One Curriculum to Rule Them All

Experimental Evidence from Utah's Jordan School District

Sam Lee Gavin Hatch Brigg Trendler*

December 2025

Abstract

Elementary curriculum choices in early grades can shape later academic outcomes. However, in our evaluation, causal evidence on alternative instructional materials remains understudied. We design a randomized crossover experiment in two second-grade classrooms in the Jordan School District in northern Utah. 33 students received either a district-mandated reading curriculum or a teacher-curated curriculum in spelling and phonics, with pre- and post-tests in each of two weeks. We estimate a bivariate fixed effects model that uses within-student changes to identify the causal effect of curated instruction on both outcomes. We then study the finite sample behavior of Wald and F-tests for the joint hypothesis of no effect using a simulation design that matches the structure of the experiment. To allow for student-specific heterogeneity in baseline achievement, we also fit a bivariate Bayesian random effects model. Frequentist joint tests do not reject the null at conventional levels in this small sample. However, posterior draws from the random effects model indicate a substantial positive treatment effect. The posterior probability that curated instruction improves at least one of the two scores is estimated at 0.98. Because test scores are censored at 100%, these estimates likely understate the magnitude of the true gains. Notwithstanding small sample limitations, we suggest that a simple curated curriculum can generate meaningful short-run improvements in reading skills for students in this setting.

1 Introduction

Elementary curriculum is an important driver of early academic success. When instructional materials fail to engage students, they may lose interest in school at a young age and have difficulty recovering. In practice, districts often adopt a single mandated curriculum, while individual teachers develop their own materials in an effort to keep students motivated. Despite the prevalence of these informal alternatives, there is little direct causal evidence that compares district-mandated curricula with teacher-curated instruction in real classrooms.

We address this gap with a randomized crossover experiment in two second-grade classrooms in the Jordan School District in northern Utah. A total of 33 students followed either a district-mandated reading curriculum or a teacher-curated curriculum in spelling and phonics over two weeks, with pre- and post-tests in each week. The experiment is designed with crossover assignment at the classroom week level and with repeated measures for each student. This yields within-student variation that supports a clear causal interpretation for short-run gains in test scores. Under standard assumptions, the design identifies the effect of curated instruction on learning outcomes for this class and time period.

The experiment is multivariate by construction, since each observation consists of a pair of outcomes for spelling and phonics scores. We use this bivariate structure to obtain tractable results in a multivariate setting. Working with two related outcomes allows us to study joint treatment effects and multivariate test statistics without the complexity of higher-dimensional multivariate methods. The bivariate model provides a natural setting to compare Wald and F-tests for a joint hypothesis and to examine how these tests behave in small samples that reflect our realistic classroom experiment.

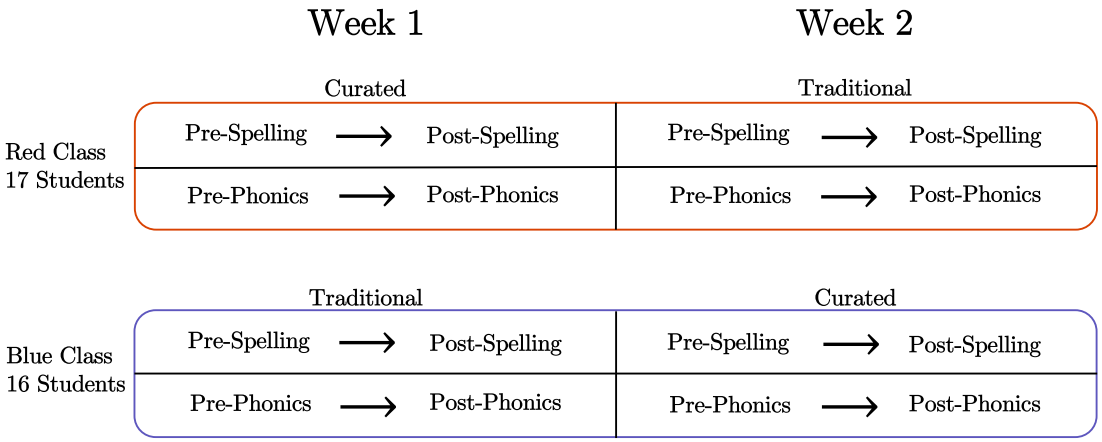
The remainder of the paper is organized as follows. Section 2 describes the experimental design, the timing of the crossover assignment, and the resulting dataset. Section 3 introduces our methods. We first develop a bivariate fixed effects model that uses within-student changes to identify the causal effect of curated instruction, then study the finite sample behavior of Wald and F-tests through a simulation calibrated to our

*Department of Statistics, Brigham Young University. We acknowledge the generous support that made this research possible. We thank Ella Hatch, who secured permission for data collection. We thank Michael Christensen for educating us in multivariate methods.

design, and finally extend the framework to a Bayesian random effects model that allows for heterogeneity between students. Section 4 reports the empirical results from both the fixed effects and random effects analyses, and interprets the joint posterior distribution of the treatment effect vector. Section 4.1 discusses how the bounded nature of test scores affects the interpretation of estimated gains. Section 5 concludes, outlines limitations of the study, and describes threats to identification and directions for future work.

2 Data and Experimental Design

The data were collected on 33 students across two second-grade classes (Blue and Red) taught by Mrs. Ella Hatch over the course of two weeks at Fox Hollow Elementary (See Figures 7 and 8 in the appendix for raw grades). She used two different curricula: traditional and teacher-curated. The traditional curriculum was written by the private company, Really Great Reading, and chosen by the district. The curated curriculum covered the same content but was taught by Mrs. Hatch in a way that she thought would be more engaging for the students. The response variable is the recorded test scores for two subjects: spelling and phonics. Each class was randomly assigned an instruction type in the first week and assigned the opposite in the following week. Each week, a pre-test and post-test for each subject were administered on Monday and Friday, respectively. Figure 1 shows this experiment design. This design results in a total of eight scores per student (two tests for two subjects over two weeks). In addition to the designed experiment, we also include a covariate for student age. Since we implement our experiment in two classes of second-graders, each student will be either seven or eight years old. While less than one year seems like an insignificant difference, numerous studies have shown that small differences in age in early childhood development can have a significant effect on learning ability [Sharp et al., 2009, Bedard and Dhuey, 2006, Navarro et al., 2016].



Notes: This figure summarizes the randomized crossover design for the two second-grade classrooms. Each row represents a classroom (Blue or Red), and each column represents a week of instruction. Boxes indicate the assigned curriculum in that classroom week (traditional or teacher-curated). Within each cell, the pre-test is given on Monday, and the post-test is given on Friday for both spelling and phonics, so each student contributes four scores per week and eight scores in total across the two-week study.

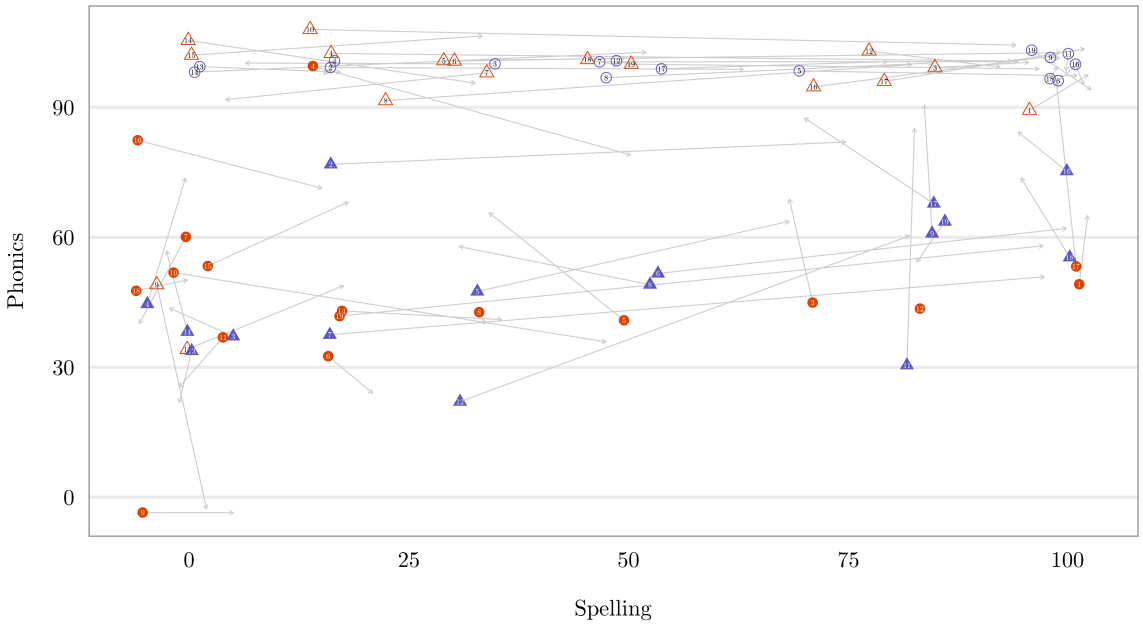
Figure 1: Experimental design

Exploratory analysis in Figure 2 shows each student's progression in score from pre-tests to post-tests broken up by class, week, and instruction type. Most arrows show improvement by pointing to the right and upwards, as we expect by virtue of any amount of instruction. In the subsequent section, we estimate which curriculum has a greater influence on score improvement. We also aggregate[1] these trends in Figure 6 (see appendix) for a more simplified view of the data.

Some students were absent on test days and do not have a recorded score. There were a total of 5 missing values. We reached out to the teacher, Mrs. Hatch, and asked if she had any thoughts on how these students

[1]Note that in our analysis, the aggregate trends may ignore within-student improvement. Hence, we account for this heterogeneity in Section 3.

would have scored had they been present. She was confident in her assessment of the missing scores due to her prior experience with those particular students, so we decided to impute[2] the grades based on "teacher's intuition."



Notes: This figure displays the observed bivariate outcomes from the experiment. Each point represents a student's spelling-phonics pre-test score, with arrows depicting within-student changes from pre-test to post-test in each week. For the purpose of presentation, we jitter each pre-test score according to a Gaussian distribution with variance 9. Triangles denote curated-instruction weeks and circles denote traditional-instruction weeks; blue markers correspond to the blue classroom and red markers to the red classroom.

Figure 2: Observed Spelling and Phonics Scores

3 Methods

Our empirical strategy proceeds in three steps. We first develop a bivariate fixed-effects model that leverages the crossover structure of the experiment to identify the short-run causal effect of curated instruction on spelling and phonics performance. We then assess the small-sample behavior of the resulting multivariate hypothesis tests through a simulation study calibrated to the structure of our data. Finally, to allow for student-level heterogeneity in baseline proficiency, we extend the framework to a bivariate random-effects specification and estimate it using a Bayesian multilevel model.

3.1 Fixed Effects Model

We begin by outlining a simple bivariate fixed-effects regression model to evaluate the treatment effect on a student's spelling and phonics scores. We model the student's outcome vector jointly as,

$$\boldsymbol{y}_{iwt} = \begin{pmatrix} \text{spelling}_{iwt} \\ \text{phonics}_{iwt} \end{pmatrix} \in \mathbb{R}^2, \tag{1}$$

for each student $i = 1, \ldots, N$, week $w \in \{1, 2\}$, and time $t \in \{\text{pre}, \text{post}\}$ indicating the time periods before and after (respectively) a student takes the quiz during a given week. Recall in Section 2 that treatment

[2]We are confident that the imputation of the five missing values over 264 observations, or 132 bivariate observations, has little effect on our resulting analysis.

occurs at the classroom level. Hence, we first propose modeling the outcome vector through the following specification.

$$\begin{aligned} \boldsymbol{y}_{iwt} &= \boldsymbol{\alpha}_i + \boldsymbol{\lambda}_w + \boldsymbol{\kappa}_c + \boldsymbol{x}'_{iwt}\boldsymbol{\gamma} \\ &\quad + \mathbb{1}\{t = \text{post}\}\boldsymbol{\beta_1} + \mathbb{1}\{\text{instruction}_{cw} = \text{curated}\}\boldsymbol{\beta}_2 \\ &\quad + \mathbb{1}\{t = \text{post}\}\mathbb{1}\{\text{instruction}_{cw} = \text{curated}\}\boldsymbol{\beta_3} \\ &\quad + \boldsymbol{u}_{iwt}, \end{aligned} \tag{2}$$

$$\text{where} \quad \boldsymbol{u}_{iwt} \sim \mathcal{N}_2\left(\boldsymbol{0}, \boldsymbol{\Sigma}_u = \begin{bmatrix} \sigma_s^2 & \rho \\ \rho & \sigma_p^2 \end{bmatrix}\right) \tag{3}$$

where we include bivariate fixed effects[3] for each student ($\boldsymbol{\alpha}_i$), week ($\boldsymbol{\lambda}_w$), and each class (either blue or red) ($\boldsymbol{\kappa}_c$). The vector $\boldsymbol{x}_{iwt}$ represents additional controls such as the age of the student. We are particularly interested in $\boldsymbol{\beta}_3$, which represents how the student performs at the end of the week on their quiz after receiving the curriculum treatment. Stacking score vectors into the corresponding $n \times 2$ matrix $\boldsymbol{Y}$, using $\boldsymbol{X}$ to represent $n \times k$ the matrix of fixed effect, time, and treatment indicators, and using the $k \times 2$ matrix $\boldsymbol{B}$ to represent the matrix of regression coefficients to be estimated $(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, the standard maximum likelihood estimators hold. Namely[4],

$$\hat{\boldsymbol{B}} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_u := \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}\right). \tag{5}$$

Thus, it follows[5] that,

$$\text{Var}\left(\text{vec}(\hat{\boldsymbol{B}})\right) := \boldsymbol{\Sigma}_u \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{6}$$

We are particularly interested in whether at least one score (either that of spelling or phonics) is improved using the curated curriculum. Formally, we state this hypothesis as,

$$\begin{aligned} &\mathcal{H}_0 : \boldsymbol{B}'\boldsymbol{c} = 0, \qquad \boldsymbol{c} = \begin{pmatrix} 0 & 0 & \cdots & 1 \end{pmatrix}' \\ &\mathcal{H}_1 : \boldsymbol{B}'\boldsymbol{c} \neq 0, \end{aligned} \tag{7}$$

where the last entry in $\boldsymbol{c}$ indicates the row position of $\boldsymbol{\beta}_3$ in $\boldsymbol{B}$. Using a consistent estimator[6] for Eq. (6), we compute the corresponding Wald statistic as,

$$W := \left(\hat{\boldsymbol{B}}'\boldsymbol{c}\right)'\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}\hat{\boldsymbol{\Sigma}}_u\right]^{-1}\left(\hat{\boldsymbol{B}}'\boldsymbol{c}\right) \xrightarrow{d} \chi^2_2. \tag{8}$$

It is not difficult to see that because $\boldsymbol{B}'\boldsymbol{c} \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{V})$ by the Normality assumption in Eq. (3), where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}\boldsymbol{\Sigma}_u$, W can be decomposed into two statistics that converge to bivariate standard Normal distributions by constructing $\boldsymbol{Z}$ such that $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}(\boldsymbol{B}'\boldsymbol{c})$. Thus, $E[\boldsymbol{Z}] = \boldsymbol{V}^{-1/2}E[\boldsymbol{B}'\boldsymbol{c}] = \boldsymbol{0}$, and $\text{Var}(\boldsymbol{Z}) = \boldsymbol{V}^{-1/2}\boldsymbol{V}\boldsymbol{V}^{-1/2} = \boldsymbol{I}_2$. Hence,

$$W = \left(\hat{\boldsymbol{B}}'\boldsymbol{c}\right)'\hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{V}}^{-1/2}\left(\hat{\boldsymbol{B}}'\boldsymbol{c}\right) = \left(\hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{B}}'\boldsymbol{c}\right)'\left(\hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{B}}'\boldsymbol{c}\right) = \hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}.$$

We show this, frankly, obvious result to motivate our use of Hotelling's correction, which we outline as follows. While $\hat{\boldsymbol{Z}} \xrightarrow{d} \boldsymbol{Z} \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I}_2)$, it may be the case that $\hat{\boldsymbol{V}}$ is not the best estimator for $\boldsymbol{V}$ to the extent that $\hat{\boldsymbol{Z}}$ differs significantly from the Normal distribution in finite samples. To this end, we remark that $\hat{\boldsymbol{V}}$

[3]In this case, since we only have two classrooms in our experiment, the fixed effects given by $\boldsymbol{\lambda}_w$ render $\boldsymbol{X}'\boldsymbol{X}$ rank deficient and are thus excluded in empirical estimation. In other words, we can always perfectly identify all the parameters in the model if we are given a student and his or her corresponding classroom. We include theses fixed effects in Eq. (3) simply for tractability.

[4]In practice, we use the unbiased estimator for $\boldsymbol{\Sigma}_u$, that is, $\frac{1}{n-k}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}\right)$.

[5]This is a standard result. However, for the interested reader, we reproduce it in Section B.

[6]Specifically, we use $\hat{\text{V}}\left(\text{vec}(\hat{\boldsymbol{B}})\right) = \hat{\boldsymbol{\Sigma}}_u \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}$, where $\hat{\boldsymbol{\Sigma}}_u$ is given in Eq. (5).

follows a Wishart distribution [Wishart, 1928]. Therefore, the statistic $W = \hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}$ follows a Hotelling T^2 distribution. Using the result from Hotelling [1931], it follows that

$$F := \frac{n-p-1}{2(n-p)} W \sim \mathrm{F}_{2,n-p-1}\,. \tag{9}$$

We evaluate the performance of both the Wald and F-statistics for testing Eq. (7) under our experimental design in the subsequent section. We do not consider eigenvalue-based MANOVA tests, since in our setting they collapse to the same underlying quadratic form; see Section A.1.

The selling feature of this specification is its causal implications within the experimental design. The inclusion of student and week fixed effects absorbs all time-invariant heterogeneity across students and all common shocks within a given week of instruction and thus enables us to recover an unbiased treatment effect. The bivariate student effects control for persistent differences in ability, motivation, or home environment that jointly influence spelling and phonics scores; likewise, the week effects $\boldsymbol{\lambda}_w$ remove any systematic differences in testing conditions or instructional content that are shared by all students in a given week. Conditional on these fixed effects, the remaining within-student variation across the pre- and post-tests isolates the short-run gain (linearly) attributable to instruction during that week[7]. Under this design, given usual model assumptions, $\boldsymbol{\beta}_3$ is identified from a clean difference-in-differences comparison of within-student gains. The estimator of $\boldsymbol{\beta}_3$ unbiasedly recovers the causal effect of the curated curriculum on student performance.

3.1.1 Simulation Study

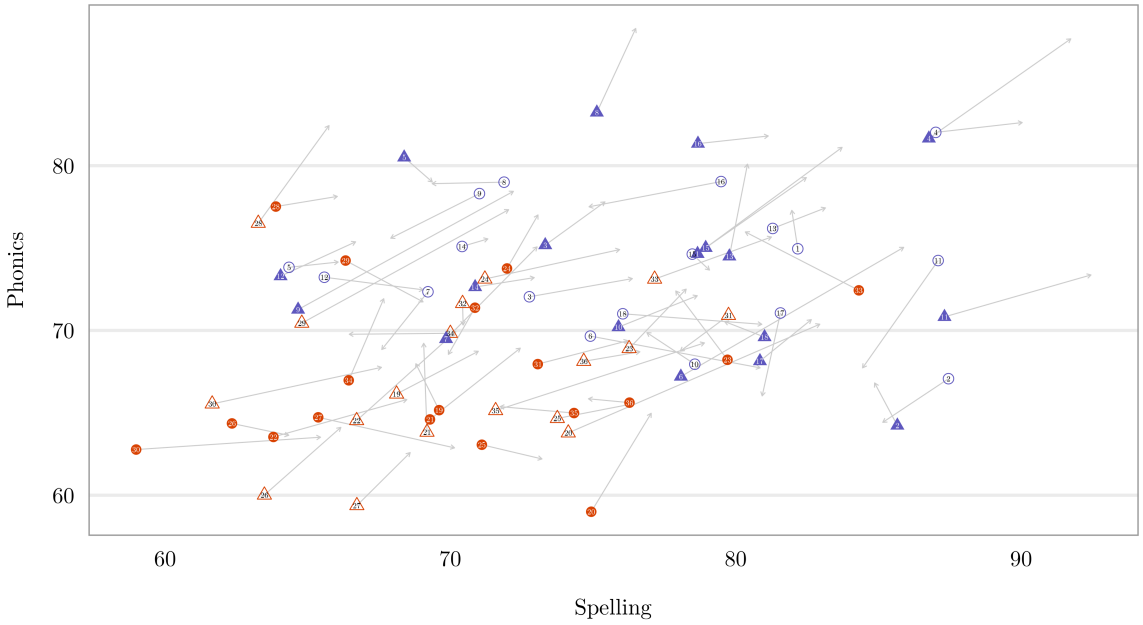
The joint hypothesis in Eq. (7) concerns a bivariate treatment effect. The curated curriculum may increase spelling performance, phonics performance, or both. Because our experiment is small by design, large-sample approximations supporting multivariate Wald tests may perform poorly in finite samples. Moreover, the power of multivariate hypothesis tests in short panels can depend sensitively on the covariance structure of the data-generating process. To justify our inferential strategy and to evaluate the operating characteristics of our test procedure in a setting commensurate with our empirical application, we conduct a simulation study calibrated to our fixed-effects model described in the previous section.

For each simulated dataset, we preserve the exact crossover structure of our experiment: two classrooms, two weeks, pre- and post-measurements in each week, and a single bivariate observation for each (i, w, t). We generate synthetic data under the multivariate fixed-effects model in Eq. (3). Specifically, each student receives independent bivariate intercepts, and scores evolve within each week according to idiosyncratic bivariate error terms with residual covariance ρ. We vary the true treatment effect vector over a 11×11 grid of values spanning both positive and negative effect sizes, yielding 121 distinct simulation configurations. We assess both the size of the test under the null hypothesis and the power using this design. For each configuration of $\boldsymbol{\beta}_3$, we generate 2,000 Monte Carlo replications for each configuration. Figure 3 displays one such realization.

Figure 4 reports the empirical rejection probabilities of both the Wald test and the F-test across the 11×11 grid of treatment effect values. We also report the Monte Carlo errors in Table 2 in the appendix. We discuss several important implications for the purposes of our experimental study. First, along the null line $\boldsymbol{\beta}_3 = (0, 0)'$, rejection rates for the Wald test concentrate around 0.06. Thus, while the Wald test may slightly over-reject in a small experimental setting such as ours, we remark that it is generally appropriately sized. The F-test, in contrast, improves on this performance and is precisely the nominal 5 percent level under the null (subject to Monte Carlo error).

Second, the power of the procedure increases smoothly as both spelling and phonics effects move in a single direction. We suggest that this is a desirable property for our experiment. Since we hypothesize that the curated curriculum will move both spelling and phonics scores above the counterfactual expectation, if

[7] We give a brief outline of this intuition here. Because the curriculum is assigned at the classroom–week level through a deterministic crossover design, the indicator $\mathbb{1}(\text{instruction}_{cw} = \text{curated})$ varies for each student across weeks but is orthogonal to their latent ability and all other time-invariant characteristics by virtue of the inclusion of fixed effects. This follows from the experimental structure: each classroom receives both instructional regimes, but in different weeks, thereby generating quasi-experimental within-student treatment variation that is unconfounded by individual heterogeneity. Thus, the interaction term $\mathbb{1}(t = \text{post})\mathbb{1}(\text{instruction}_{cw} = \text{curated})$ compares post-test improvements that come by way of the curated curriculum *relative* to the improvement that would've otherwise come by the traditional instruction, differencing out all baseline differences captured by $\boldsymbol{\alpha}_i$ and all week-specific shocks captured by $\boldsymbol{\lambda}_w$.



Notes: This figure displays a single simulated realization of spelling and phonics scores. Each point represents an (i, w, t) observation for a student, where student IDs are denoted by the numbers inscribed in each point. Arrows from each node point to the end-of-week score vector. Treatment status is indicated by the marker shape: circles denote traditional instruction, while triangles denote the curated curriculum treatment. Red shapes indicate observations from the red class, and blue shapes indicate observations from the blue class. Open shapes indicate quiz scores in week 1, and closed shapes indicate quiz scores in week 2. We expect that some instruction, regardless of type, will improve scores at the end of the week. However, on average, we hypothesize that the idiosyncratic slopes (the measure of improvement) for each student will be larger in magnitude than the non-treated groups. Compare Figure 2.

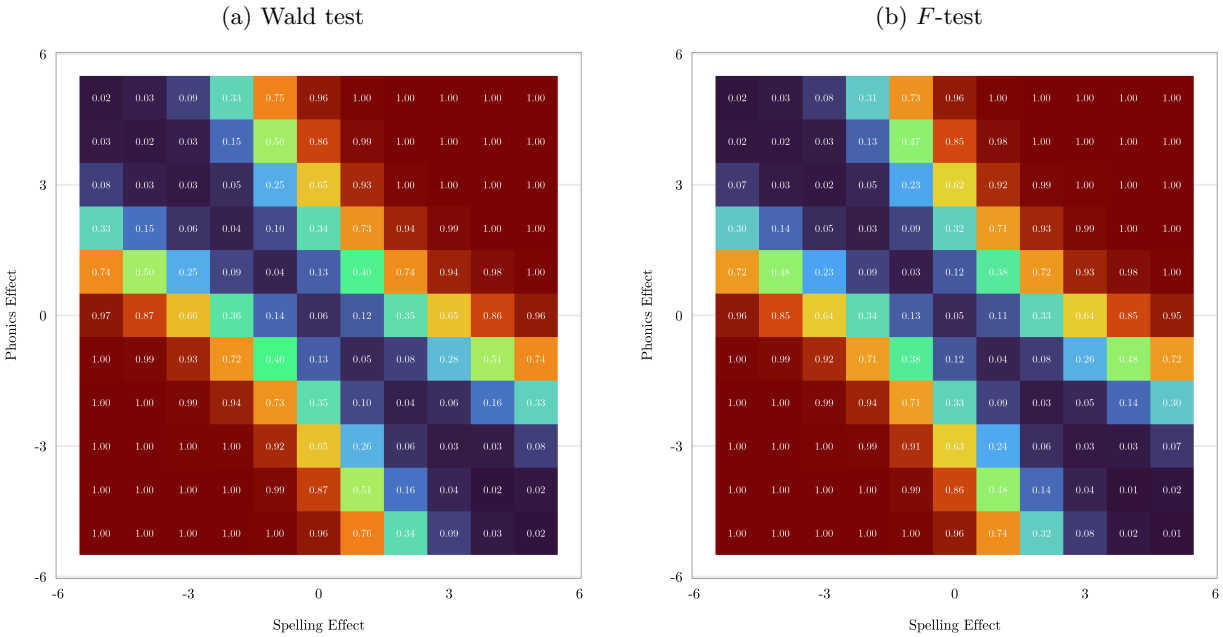
Figure 3: Simulated Bivariate Outcomes

we detect a positive treatment effect vector, we maintain that the effect we see is recovered with high power. The Wald and F-tests perform similarly in this region.

Third, when the two components of $\boldsymbol{\beta}_3$ have opposite signs, rejection probabilities fall dramatically, even when the magnitudes of the effects are large. We suspect this occurs because both the Wald statistic and the F statistic depend on the same quadratic form $\hat{\boldsymbol{\beta}}_3' \widehat{\mathbf{V}}^{-1} \hat{\boldsymbol{\beta}}_3$, which is sensitive to the direction of the true effect vector. Alternatives that move away from the origin but point into different quadrants relative to the null do not produce large test statistics. In this sense, opposite-signed effects "cancel" each other out in both procedures. We remark that this behavior is consistent with the substantive hypothesis of interest. The tests are designed to detect whether curated instruction produces simultaneous improvements in both spelling and phonics, so improvements in only one outcome, or improvements in one paired with deterioration in the other, do not provide evidence against the joint null.

3.2 Random Effects Model

A natural alternative to the fixed-effects specification in Eq. (3) is to treat student-specific heterogeneity as arising from a random sampling process. In multilevel settings, a random-effects formulation offers two advantages. First, by shrinking individual intercepts toward a common population distribution, the random-effects estimator can be substantially more efficient than the fixed-effects estimator when the student effect is uncorrelated with the regressors. We defend this assumption in our setting because the curriculum assignment varies only at the classroom–week level and thus is orthogonal to latent student ability by design. Second, modeling the dependence structure through a random-effects model allows us to quantify heterogeneity in baseline spelling and phonics proficiency across the population. To incorporate the random-effects design, we make a slight adjustment to Eq. (3) by replacing the student fixed effects $\boldsymbol{\alpha}_i$ with a bivariate random



Notes: Each panel reports Monte Carlo rejection probabilities for the joint null $\mathcal{H}_0 : \boldsymbol{\beta}_3 = (0, 0)'$ in our bivariate fixed-effects model. The horizontal and vertical axes index the true spelling and phonics components of $\boldsymbol{\beta}_3$ on an 11×11 grid ranging from -5 to 5. Cell colors and overlaid numbers give the proportion of 2,000 simulated samples in which the null is rejected at the nominal $\alpha = 5\%$ level. Panel (a) shows the performance of the χ^2 Wald test; panel (b) displays the corresponding F-test. Selected results are also presented in Table 2.

Figure 4: Empirical rejection rates for joint tests over a grid of treatment effects

intercept:

$$\begin{aligned} \boldsymbol{y}_{iwt} = \boldsymbol{\mu}_0 + \boldsymbol{\lambda}_w + \boldsymbol{\kappa}_c + \boldsymbol{x}'_{iwt}\boldsymbol{\gamma} \\ + \boldsymbol{z}'_{iwt}\boldsymbol{b}_i \\ + \mathbb{1}\{t = \text{post}\}\boldsymbol{\beta_1} + \mathbb{1}\{\text{instruction}_{cw} = \text{curated}\}\boldsymbol{\beta}_2 \\ + \mathbb{1}\{t = \text{post}\}\mathbb{1}\{\text{instruction}_{cw} = \text{curated}\}\boldsymbol{\beta_3} \\ + \boldsymbol{u}_{iwt}, \end{aligned} \tag{10}$$

$$\text{where} \quad \boldsymbol{u}_{iwt} \sim \mathcal{N}_2\left(\mathbf{0}, \boldsymbol{\Sigma}_u = \begin{bmatrix} \sigma_s^2 & \rho \\ \rho & \sigma_p^2 \end{bmatrix}\right), \tag{11}$$

and we include an overall intercept $\boldsymbol{\mu}_0$ so that the student random effects $\boldsymbol{b}_i$ may be parameterized with mean zero. We set

$$\begin{aligned} \boldsymbol{z}_{iwt} &= \boldsymbol{I}_2, \\ \boldsymbol{b}_i &\sim \mathcal{N}_2\left(\mathbf{0}, \boldsymbol{\Omega} = \begin{bmatrix} \eta_s^2 & \eta_{sp} \\ \eta_{sp} & \eta_p^2 \end{bmatrix}\right), \end{aligned}$$

so that each student has a pair of latent baseline abilities (one for spelling and one for phonics). A priori, these random intercepts may be correlated through the off-diagonal parameter η_{sp}, but in estimation we regularize this correlation toward zero via a shrinkage prior. To that end, we opt to switch to a Bayesian paradigm for ease of estimation. While it *is* possible to estimate our bivariate mixed model with a somewhat complex Newton-Rhapson optimization routine, we note that computational and inferential limitations under such an approach motivate a Bayesian treatment of the random-effects specification. We include a discussion of our decision to switch to a Bayesian framework in the appendix.

We give a descriptive outline of our Bayesian estimation strategy as follows. Conditional on the random

effects, the likelihood factorizes from Eq. (11) as,

$$\boldsymbol{y}_{iwt} \mid \boldsymbol{b}_i, \boldsymbol{B}, \boldsymbol{\Sigma}_u \sim \mathcal{N}_2(\boldsymbol{\mu}_0 + \boldsymbol{\lambda}_w + \boldsymbol{\kappa}_c + \boldsymbol{x}'_{iwt}\boldsymbol{\gamma} + \boldsymbol{z}'_{iwt}\boldsymbol{b}_i,\ \boldsymbol{\Sigma}_u)\,,$$

with $\boldsymbol{b}_i \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Omega})$ as above. For the fixed effects we place weakly informative Gaussian priors of the form

$$\beta_{jk} \sim \mathcal{N}\big(0,\ s_{jk}^2\big),$$

where the prior scales s_{jk} are chosen automatically[8] by our statistical software [Goodrich et al., 2025].

For the covariance matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}_u$ we adopt the decomposed covariance prior. Each covariance matrix can be written as $\boldsymbol{\Sigma} = \mathbf{DRD}$, where $\mathbf{D}$ is diagonal with nonnegative standard deviations on the diagonal and $\mathbf{R}$ is a 2×2 correlation matrix. Independent weakly informative priors are placed on the standard deviations.[9] Posterior inference is based on draws from the joint posterior distribution

$$\pi\big(\boldsymbol{B}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_u, \{\boldsymbol{b}_i\}_{i=1}^N \mid \{\boldsymbol{y}_{iwt}\}\big)\,,$$

obtained via Hamiltonian Monte Carlo as implemented in Stan. Again, our primary interest lies in the bivariate treatment-effect vector $\boldsymbol{\beta}_3$. Rather than relying on an asymptotic Wald test as we did for our fixed effect model, we summarize the joint posterior distribution of $\boldsymbol{\beta}_3$ directly. The posterior probability that the curated curriculum improves performance in both subjects is then estimated by

$$\widehat{\Pr}\big(\beta_{3,\text{spelling}} > 0,\ \beta_{3,\text{phonics}} > 0 \mid \text{data}\big) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\Big\{\beta_{3,\text{spelling}}^{(m)} > 0,\ \beta_{3,\text{phonics}}^{(m)} > 0\Big\}\,, \tag{12}$$

for each posterior draw, m. We are also interested in the following posterior probability.

$$1 - \widehat{\Pr}\big(\beta_{3,\text{spelling}} < 0,\ \beta_{3,\text{phonics}} < 0 \mid \text{data}\big), \tag{13}$$

which is the posterior probability that it is *not* the case that curated instruction harms both spelling and phonics simultaneously. Or in other words, this represents the posterior probability that at least one test score is improved relative to the traditional curriculum, even at the expense of the improvement of the other score metric. These posterior probabilities provide a Bayesian equivalent of a joint test of $\mathcal{H}_0 : \boldsymbol{\beta}_3 = \mathbf{0}$, with two clear advantages. First, it does not rely on large-sample χ^2 approximations and is straightforward to interpret as the posterior mass assigned to the region of $(\beta_{3,\text{spelling}}, \beta_{3,\text{phonics}})$ space corresponding to positive treatment effects in both outcomes.

4 Results

In this section, we present the empirical findings from our experiment. We begin by examining the estimated effects of our main parameter vector of interest, that is, the effect of the curated instruction relative to the effects of the traditional curriculum. The estimated coefficient vector ($\boldsymbol{\beta}_3$) and other parameters of interest are reported in Table 1. The fixed effects estimates indicate that the point estimates of the curated curriculum interaction are positive for both spelling and phonics. Under our frequentist estimation, neither component is individually significant, and we fail to reject the joint null $H_0 : \boldsymbol{\beta}_3 = \mathbf{0}$ and $\alpha = 0.05$. We suggest that this is consistent with the limited size of the experiment and the resulting variability in the estimated effects. Beyond the treatment effects, Table 1 also reveals interesting patterns in the covariates. Age is positively associated with both spelling and phonics scores in all specifications. This supports the evidence given in Section 2 that small age differences yield the strongest marginal effects in scholastic performance for younger students. Students in the red classroom score lower on average in both subjects, which indicates baseline differences across classes that the fixed-effects specification is intended to absorb. As

[8] Specifically, s_{jk} is proportional to the ratio of the marginal standard deviation of outcome j to that of predictor k, multiplied by a fixed constant. This yields a weakly informative prior that is invariant to linear rescaling of the predictors and outcomes.

[9] These are exponential-type priors that downweight unrealistically large values. Additionally, an LKJ(2) prior with concentration parameter is placed on the correlation matrix. The LKJ prior centers $\mathbf{R}$ at the identity and mildly shrinks correlations toward zero while still allowing substantial posterior correlation when supported by the data. We specify the same prior family for both $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}_u$, with a regularization parameter tuned to avoid overfitting in this small sample.

a result, it may be reasonable to hypothesize that potential peer spillover effects may differ by classroom, since students surrounded by higher-performing peers may experience stronger motivational or strategic incentives to improve. We view this as a natural direction for future work for researchers who seek to replicate or extend a similar experimental design. We also point out that the main effect of curated instruction (β_2) is near zero in both models. This suggests that differences across instruction types arise primarily through within-week improvements.

Table 1: Frequentist and Bayesian Estimates for Curriculum Effect

Panel A. Coefficient Summary

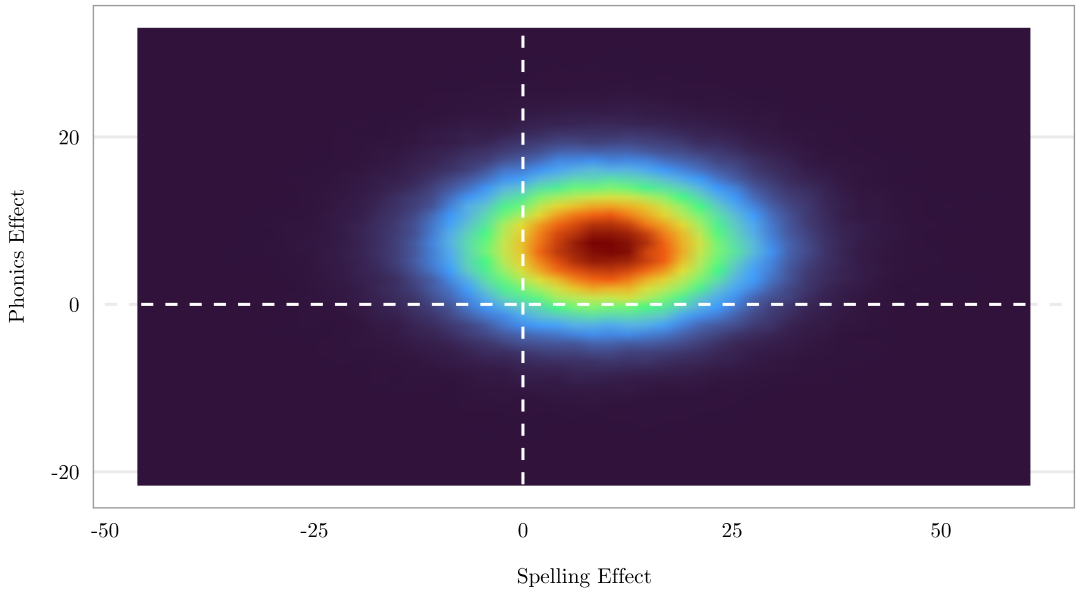
	Frequentist (FE model)		**Bayesian (RE model)**		
	Estimate	SE	Mean	Median	95% HPD
Spelling					
Age	22.92	9.19	17.37	17.38	[2.34, 32.40]
Class Red	-19.03	6.11	-17.40	-17.42	[-28.79, -5.90]
β_1	13.09	7.83	13.12	13.14	[-1.81, 28.09]
β_2	-1.59	7.84	-1.56	-1.56	[-16.71, 13.22]
β_3	9.64	11.08	9.60	9.63	[-11.36, 30.83]
Phonics					
Age	16.54	4.72	13.69	13.69	[4.37, 22.65]
Class Red	-7.74	5.40	-8.00	-7.99	[-14.12, -1.88]
β_1	1.00	4.02	1.00	0.98	[-6.91, 8.74]
β_2	-3.41	4.02	-3.40	-3.40	[-11.32, 4.26]
β_3	6.82	5.69	6.81	6.82	[-4.10, 17.92]

Panel B. Inference Summary

	Test	Statistic	p-value
Frequentist	Wald	$W = 1.44$	0.48
	F	$F = 0.72$	0.49
Bayesian	$\Pr(\beta_{3,\text{spell}} > 0,\ \beta_{3,\text{phon}} > 0 \mid \text{data})$	0.72	
	$1 - \Pr(\beta_{3,\text{spell}} < 0,\ \beta_{3,\text{phon}} < 0 \mid \text{data})$	0.98	

Notes: Parameters estimated with $n = 132$ observations with 33 total students from two classes. Frequentist estimates come from the bivariate fixed-effects model estimation in Section 3.1. The Wald statistic in Panel B is computed using the quantiles of a χ^2_2 distribution, and the F-stat is computed using the quantiles of the $F_{2,108}$ distribution. Bayesian estimates come from the random-effects model in Section 3.2 and reflect estimates on the posterior draws. HPD represents the 95% highest posterior density intervals. Joint inference evaluates the null hypothesis $H_0 : \boldsymbol{\beta}_3 = \mathbf{0}$ for the spelling and phonics treatment effects as given in Eq. (7). Posterior probabilities summarize the mass assigned to the positive-positive and non-negative quadrants of the joint posterior distribution.

However, we contend that our Bayesian random effects analysis provides a more informative description of the joint parameter space. Figure 5 displays the posterior density of $(\beta_{3,\text{spelling}}, \beta_{3,\text{phonics}})$, which places most of its mass in the region where both effects are positive. Posterior means, medians, and highest posterior density intervals in Table 1 reflect the same pattern. The two joint posterior probabilities computed from Eq. (13) and (12) summarize this evidence. The probability that both effects are positive is computed as 0.72. Hence, we suggest that the effect we estimate is practically significant even if it does not show statistical significance from a frequentist standpoint. The posterior probability that the curated curriculum instruction improves *at least one* score relative to the traditional curriculum is computed as 0.98 and grants our most substantive result of the analysis. This shows that the data place almost all of their weight on outcomes in which students benefit.



Notes: This figure displays the estimated posterior joint density of the bivariate treatment-effect vector from the Bayesian random-effects model. Warmer colors indicate regions of higher posterior probability. The dashed lines mark the axes corresponding to zero effects on spelling and phonics. The posterior mass is concentrated in the positive-positive quadrant, indicating strong posterior support that curated instruction improves both skills simultaneously.

Figure 5: Posterior Joint Density of the Bivariate Treatment Effect

4.1 Attenuation Bias

The astute reader will notice that we have ignored the censored nature of our data. So far, we have not considered that by construction, quiz scores are bounded between 0 and 100; we ignore this constraint in both our modeling and simulation studies. Several students achieve perfect scores before and after instruction, as can be seen in Figure 2. Since post-treatment progress cannot be expressed beyond the score cap, measured gains are compressed relative to true gains. Hence, this kind of censoring induces attenuation bias in the estimated effects and shifts mass in the posterior toward more conservative values. Under this structure, the estimates reported here should be interpreted as lower bounds on the magnitude of the true treatment effect.

5 Conclusion

We study a randomized crossover experiment in two second-grade classrooms in the Jordan School District and evaluate the effect of a teacher-curated reading curriculum on spelling and phonics scores. A bivariate fixed effects model uses within-student changes to identify the short-run causal effect of curated instruction. In this framework, the multivariate Wald and F-tests are approximately well-sized in finite samples and exhibit strong power when effects move in the same direction. The quadratic form of these statistics limits hypothesis testing when effects diverge in opposite directions. A Bayesian random effects model complements this analysis by providing a richer summary of the joint treatment effect. Posterior draws place most of their mass on regions where curated instruction improves at least one outcome relative to the effects of the traditional curriculum. The posterior probability that at least one of the two scores improves relative to the traditional curriculum is estimated at 0.98, which we interpret as strong evidence that the curated curriculum benefits students in this setting.

5.1 Limitations

The empirical design and data impose several constraints on external validity. The sample is small, with 33 students drawn from a single school and a single instructor. The experiment also covers only two weeks, so we do not study the longer-run persistence of gains or the possibility of dynamic treatment effects. As discussed in earlier sections, these features limit the generalizability of our conclusions and suggest that our estimates should be viewed as local to this population, teacher, and time horizon. Moreover, we conclude that any effects that are present represent a lower bound on the true treatment effect due to the attenuation bias that occurs by the nature of the data.

Despite these limitations, we conclude that our study provides a concrete example of how elementary schools can use simple experimental designs and multivariate methods to evaluate instructional choices. The evidence indicates that a curated curriculum can generate meaningful improvements in reading skills over a short horizon and that Bayesian multilevel models can extract more information from small educational experiments than standard large sample tests. Future work could extend this design to more classrooms, grade levels, and schools with different demographic and socioeconomic profiles. Larger studies that randomize curricula across multiple teachers in a double-blind fashion, track students over longer periods, and measure peer interactions would help separate teacher effort from curricula effects that researchers intend to assess and clarify how alternative curricula shape the joint development of reading skills.

5.1.1 Threats to Identification

A central threat to identification arises from potential experimenter bias. Because the instructor in our experiment expressed a strong preference against the traditional curriculum, differential effort itself may account for at least part of the observed treatment effect. In this case, our estimates would capture a compound effect of instructional content and teacher motivation, and hence, we cannot disentangle these channels within the confines of this study. A more credible design would randomize curricula at the classroom level without disclosing assignments to teachers, thereby holding teacher beliefs and effort constant while isolating curricular effects. Nevertheless, our findings provide motivation for such a design. Even if the estimated gains reflect effort, the result is substantively meaningful. We conclude that variation in teacher engagement can generate measurable differences in short-run learning outcomes for their students.

References

K. Bedard and E. Dhuey. The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics*, 121(4):1437–1472, 2006. doi: 10.1093/qje/121.4.1437. URL `https://academic.oup.com/qje/article-abstract/121/4/1437/1855234`.

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. *rstanarm: Bayesian Applied Regression Modeling via Stan*, 2025. URL `https://mc-stan.org/rstanarm/`. R package version 2.32.2.

Harold Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.

R. M. Navarro, M. C. Aguilar, and D. Alonso. The relative age effect and its influence on academic performance in primary school. *PloS One*, 10(10):e0141895, 2016. doi: 10.1371/journal.pone.0141895. URL `journals.plos.org`.

Caroline Sharp, Nalia George, Claire Sargent, Sharon O'Donnell, and Maureen Heron. The influence of relative age on learner attainment and development: A review of international evidence. Technical report, Department for Education, February 2009. URL `https://assets.publishing.service.gov.uk/media/5a82a07de5274a2e87dc2281/0209\_CarolineSharp_et_al_RelativeAgeReviewRevised.pdf`. A UK government review confirming effects in primary school.

John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):32–52, 1928.

A Discussion

A.1 Eigenvalue Tests

In principle, one could replace the Wald statistic with classical MANOVA tests that are functions of the eigenvalues of the matrix $\mathbf{E}^{-1}\mathbf{H}$, such as Wilks's Λ, Pillai's trace, the Hotelling–Lawley trace, or Roy's largest root. However, in our setting, the linear hypothesis $H_0 : \boldsymbol{\beta}_3 = \mathbf{0}$ has rank one, so the associated hypothesis sum of squares and products matrix $\mathbf{H}$ has rank at most one and $\mathbf{E}^{-1}\mathbf{H}$ has at most a single nonzero eigenvalue, say λ_1. All four MANOVA test statistics become monotone functions of this same scalar quantity. In particular, the Hotelling–Lawley trace reduces to $T = \lambda_1$, and the Wald statistic for $H_0 : \boldsymbol{\beta}_3 = \mathbf{0}$ is proportional to T and hence to the usual Hotelling T^2 statistic. Eigenvalue-based tests, therefore, compress the sample information into the same quadratic form that underlies our Wald and F tests, so they cannot remedy the low power against mixed-sign alternatives documented in our simulations.

A.2 A Case for a Bayesian Approach

To illustrate why estimating the multivariate random effects model in Eq. (11) may be more tractable in a Bayesian paradigm, we first consider a motivating result from statistical theory. Observe that, given our design, by conditioning on the week and classroom effects and the observed covariates, the marginal variance of an individual observation can be decomposed as,

$$\begin{aligned}\mathrm{Var}(\boldsymbol{y}_{iwt} \mid \boldsymbol{\lambda}_w, \boldsymbol{\kappa}_c, \boldsymbol{x}_{iwt}, \boldsymbol{z}_{iwt}) &= \mathrm{Var}(\boldsymbol{z}'_{iwt}\boldsymbol{b}_i) + \mathrm{Var}(\boldsymbol{u}_{iwt}) \\ &= \boldsymbol{\Omega} + \boldsymbol{\Sigma}_u. \end{aligned} \tag{14}$$

Stacking the four (w, t) observations for student i into the vector $\boldsymbol{y}_i \in \mathbb{R}^8$, and letting $\boldsymbol{J}$ and $\boldsymbol{I}$ denote 4×4 matrices of ones and identity matrices respectively, the implied covariance structure for the multivariate repeated-measures vector is

$$\mathrm{Var}(\boldsymbol{y}_i) = (\boldsymbol{J} \otimes \boldsymbol{\Omega}) + (\boldsymbol{I} \otimes \boldsymbol{\Sigma}_u), \tag{15}$$

where $\otimes$ denotes the Kronecker product. We continue to assume that students are independent across i, so the full covariance matrix of all stacked outcomes is block diagonal with blocks given by $\mathrm{Var}(\boldsymbol{y}_i)$. In principle, under the Gaussian specification above, the generalized least squares estimator of the coefficient matrix $\boldsymbol{B}$ that collects all fixed effects across both outcomes is

$$\hat{\boldsymbol{B}}(\boldsymbol{\theta}) = \left(\boldsymbol{X}' \,\mathrm{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}' \,\mathrm{V}(\boldsymbol{\theta})^{-1} \boldsymbol{Y},$$

where $\boldsymbol{\theta}$ stacks the variance parameters in $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Omega}$. The corresponding asymptotic covariance of $\mathrm{vec}(\hat{\boldsymbol{B}})$ has the form

$$\mathrm{Var}\Big(\mathrm{vec}(\hat{\boldsymbol{B}}(\boldsymbol{\theta}))\Big) = \boldsymbol{\Sigma}_u \otimes \left(\boldsymbol{X}' \,\mathrm{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X}\right)^{-1},$$

so in theory one could obtain a Wald-type test for the contrast $\boldsymbol{B}'\boldsymbol{c}$ by plugging in a REML estimate $\hat{\boldsymbol{\theta}}$ of the variance parameters and a consistent estimate of $\boldsymbol{\Sigma}_u$. This follows as,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \; \ell(\boldsymbol{\theta}) \equiv \arg\max_{\boldsymbol{\theta}} \left[-\tfrac{1}{2}\log|\mathrm{V}(\boldsymbol{\theta})| - \tfrac{1}{2}\log\left|\boldsymbol{X}' \,\mathrm{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X}\right| - \tfrac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}(\boldsymbol{\theta})\right)' \mathrm{V}(\boldsymbol{\theta})^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}(\boldsymbol{\theta})\right) \right].$$

In practice, however, this approach is cumbersome for two reasons. First, standard frequentist mixed-effects software typically fits separate univariate mixed models for each outcome and does not expose the joint multivariate covariance structure needed to evaluate the expressions above. Implementing a full multivariate REML estimator for $\boldsymbol{\theta}$ and the corresponding Wald covariance would require jointly optimizing over a high-dimensional, non-convex parameter space. Moreover, the REML estimates of the fixed effects and the variance components are mutually dependent, so obtaining the Wald covariance entails solving a system with no closed-form expression. Second, for our empirical purposes, the available sample size is small (33 students with four repeated measurements per student), and the large-sample χ^2_2 approximation underlying a multivariate Wald test is unlikely to be reliable, as demonstrated in Section 3.1.1.

B Additional Derivations

To obtain the sampling variance of $\mathrm{vec}(\hat{\boldsymbol{B}})$ as given in Eq. (6), we write the multivariate regression model as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}$ with $\mathrm{vec}(\boldsymbol{U}) \sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{\Sigma}_u \otimes \boldsymbol{I}_n)$. The maximum likelihood estimator satisfies

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{B} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U},$$

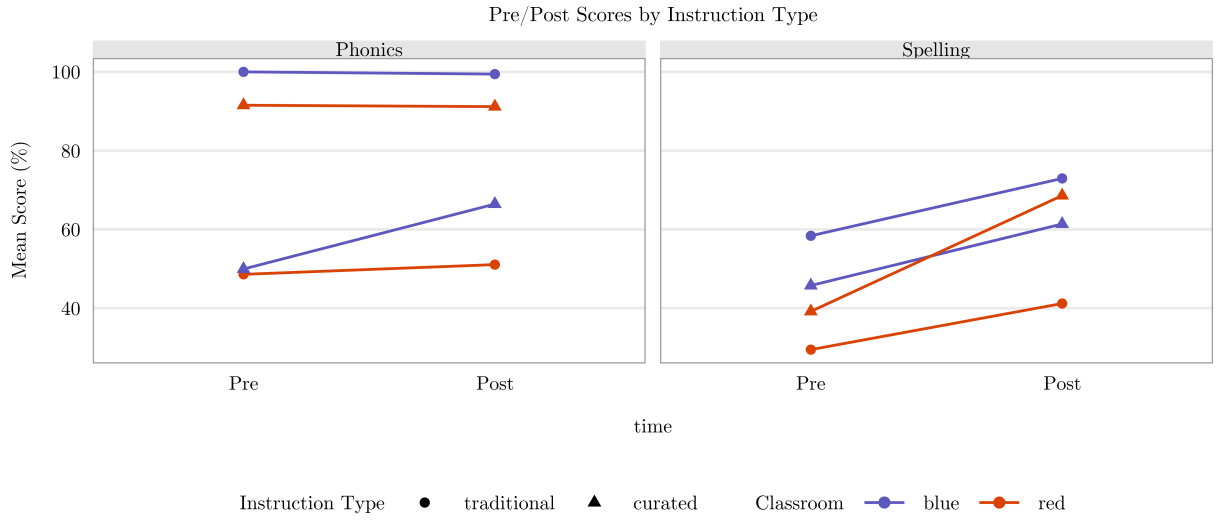
so vectorization yields

$$\mathrm{vec}(\hat{\boldsymbol{B}}) = \mathrm{vec}(\boldsymbol{B}) + (\boldsymbol{I}_2 \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\ \mathrm{vec}(\boldsymbol{U}).$$

Hence, it follows directly,

$$\begin{aligned}
\mathrm{Var}\Big(\mathrm{vec}(\hat{\boldsymbol{B}})\Big) &= (\boldsymbol{I}_2 \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\ \mathrm{Var}(\mathrm{vec}(\boldsymbol{U}))\ (\boldsymbol{I}_2 \otimes \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
&= (\boldsymbol{I}_2 \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\,(\boldsymbol{\Sigma}_u \otimes \boldsymbol{I}_n)\,(\boldsymbol{I}_2 \otimes \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
&\quad (\text{using } (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C}) \otimes (\boldsymbol{B}\boldsymbol{D})) \\
&= \big(\boldsymbol{I}_2\boldsymbol{\Sigma}_u\boldsymbol{I}_2\big) \otimes \big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\big) \\
&= \boldsymbol{\Sigma}_u \otimes \big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\big) \\
&= \boldsymbol{\Sigma}_u \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

C Additional Tables & Figures



Notes: This figure aggregates the student-level data from Figure 2 to show mean pre- and post-test scores for spelling and phonics by instruction type (traditional vs. curated) and classroom (blue vs. red). Circles represent traditional instruction and triangles represent the curated curriculum. Lines connect pre/post means within each classroom–instruction cell. The figure summarizes the same underlying observations as Figure 2 but collapses them to group means for a comparison of average gains across instructional regimes.

Figure 6: Mean pre/post scores by instruction type and classroom

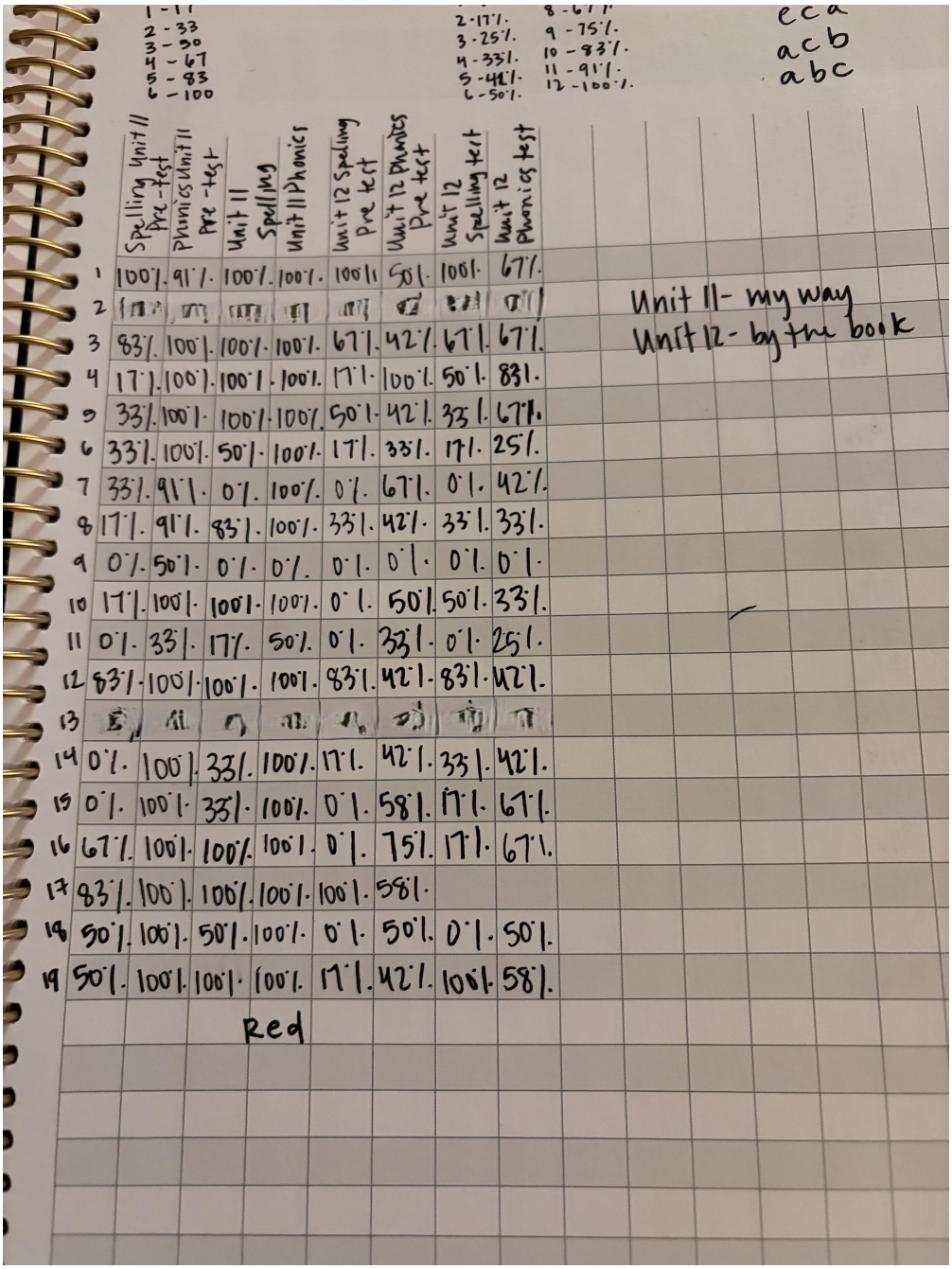


Figure 7: Gradebook for Red Class

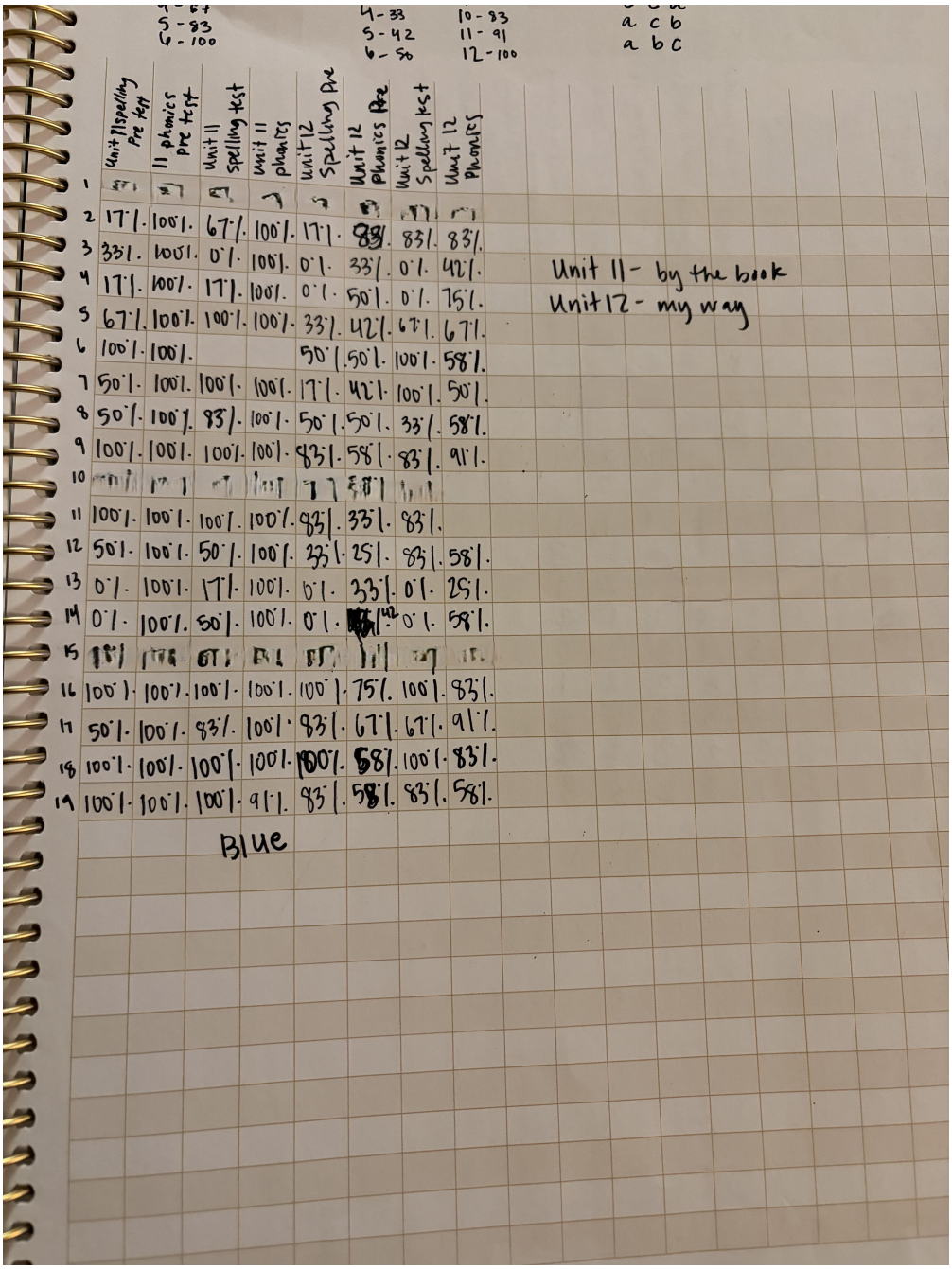


Figure 8: Gradebook for Blue Class

Table 2: Selected empirical rejection rates ($\alpha = 0.05$) for joint tests of $\mathcal{H}_0 : \boldsymbol{\beta}_3 = (0, 0)'$

True effects		**Wald test**		F **test**	
Spelling	Phonics	Rej. prob.	MCSE	Rej. prob.	MCSE
Panel A: Positive–positive effects					
0	0	0.057	0.005	0.050	0.005
1	1	0.396	0.011	0.375	0.011
2	2	0.936	0.005	0.930	0.006
3	3	0.999	0.001	0.999	0.001
Panel B: Opposite-sign effects					
4	-4	0.017	0.003	0.015	0.003
2	-2	0.035	0.004	0.029	0.004
-2	2	0.038	0.004	0.034	0.004
-4	4	0.024	0.003	0.022	0.003

Notes: Each entry reports the empirical rejection frequency across 2,000 Monte Carlo replications for tests of $\mathcal{H}_0 : \boldsymbol{\beta}_3 = (0, 0)'$ in the bivariate fixed-effects model of Section 3.1. The data-generating process uses two classrooms (blue and red) with $n = 18$ students per class; baseline ability is generated with $\alpha_i \sim \mathcal{N}(70, 5^2)$, and subject-specific test shifts of 2 points in spelling and -2 points in phonics. Week effects are $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \mathbf{0}$, class effects are $\boldsymbol{\kappa}_{\text{blue}} = (5, 5)'$ and $\boldsymbol{\kappa}_{\text{red}} = \mathbf{0}$, and main effects satisfy $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}$. Idiosyncratic errors are bivariate normal with standard deviations $\sigma_s = \sigma_p = 2$ and correlation $\rho = 0.9$. The true treatment-effect vector in each row is $\boldsymbol{\beta}_3 = (\beta_{3,\text{spelling}}, \beta_{3,\text{phonics}})'$. The bold entry marks the rejection probability closest to the nominal size under the null.

D R Code

All code is also documented and publicly available here: `https://github.com/SamLeeBYU/south-jordan-curriculum-experiment`.

D.1 Data Cleaning

```r
## data_cleaning.R

library(tidyverse)

red <- read.csv("data/raw_data/red_scores.csv")
blue <- read.csv("data/raw_data/blue_scores.csv")

# Red class curated first
# Blue class curated second
# Red student 17 -> 100 both post assessments
# Blue
#   6 -> 100 for both assessments
#   10, 11 -> 83 for both

### 1. Impute data based on Ella's feedback
red$Unit.12.Spelling.Post[17] = "100.00%"
red$Unit.12.Phonics.Post[17] = "100.00%"

blue$Unit.11.Spelling.Post[6] = 100
blue$Unit.11.Phonics.Post[6] = 100

blue$Unit.12.Phonics.Post[10] = 83
blue$Unit.12.Phonics.Post[11] = 83

### 2. Drop rows where Student.Number is NA ----
red <- red %>% filter(!is.na(Student.Number))
blue <- blue %>% filter(!is.na(Student.Number))

### 3. Add classroom labels ----
red <- red %>% mutate(classroom = "red")
blue <- blue %>% mutate(classroom = "blue")

### 4. Make score columns numeric ----
pct_to_num <- function(x) {
  x %>% str_replace("%", "") %>% as.numeric()
}

red <- red %>%
  mutate(across(starts_with("Unit"), pct_to_num))

blue <- blue %>%
  mutate(across(starts_with("Unit"), as.numeric))


### 5. Combine into one dataset ----
all <- bind_rows(red, blue)

### 6. Pivot to long format ----
long <- all %>%
  pivot_longer(
    cols = matches("^Unit\\.\\d+"),
```

```
    names_to = c("unit", "test", "time"),
    names_pattern = "^(Unit\\.\\d+)\\.(Spelling|Phonics)\\.(Pre|Post)$",
    values_to = "score"
  )

### 7. Recode variables ----
long <- long %>%
  mutate(
    id = Student.Number,
    week = ifelse(unit == "Unit.11", 1L, 2L),
    time = factor(time, levels = c("Pre", "Post")),
    test = tolower(test), # "Spelling" -> "spelling", "Phonics" -> "phonics"
    classroom = factor(classroom)
  )

### 8. Assign instruction type based on classroom times week ----
instruction_fun <- function(classroom, week) {
  if (classroom == "blue" && week == 1) {
    return("traditional")
  }
  if (classroom == "blue" && week == 2) {
    return("curated")
  }
  if (classroom == "red" && week == 1) {
    return("curated")
  }
  if (classroom == "red" && week == 2) return("traditional")
}

long <- long %>%
  rowwise() %>%
  mutate(instruction = instruction_fun(classroom, week)) %>%
  ungroup() %>%
  mutate(
    instruction = factor(instruction, levels = c("traditional", "curated"))
  )

### 9. Final cleanup ----
final_data <- long %>%
  dplyr::select(id, Age, classroom, week, time, instruction, test, score) %>%
  arrange(id, week, test, time)

glimpse(final_data)
saveRDS(final_data, "data/final_project_scores.rds")

## data.R
library(dplyr)

ellas.class <- readRDS("data/final_project_scores.rds") %>%
  pivot_wider(names_from = "test", values_from = "score")
```

D.2 DGP

```
##dgp.R

library(dplyr)
library(tidyr)
```

```r

#This DGP was originally written by Gavin Hatch,
# recoded by Sam Lee to match regression model in the paper

library(dplyr)
library(MASS)

generate_classroom_data <- function(
  n_per_class = 18,
  classes = c("blue", "red"),

  # student ability
  ability_mean = 70,
  ability_sd = 6,

  # test-specific shift
  test_effect = c(spelling = 2, phonics = -2),

  # week and class fixed effects
  lambda_week = c(`1` = 0, `2` = 1),
  kappa_class = c(blue = 0, red = 0),

  # regression parameters
  beta1 = c(spelling = 4, phonics = 4), # effect of post indicator
  beta2 = c(spelling = 0, phonics = 0), # main effect of curated
  beta3 = c(spelling = 3, phonics = 3), # post x curated gain

  # bivariate error covariance
  sigma_s = 4, # spelling SD
  sigma_p = 4, # phonics SD
  rho = 0, # covariance term
  clip_lo = 0,
  clip_hi = 100,
  seed = NULL
) {
  if (!is.null(seed)) {
    set.seed(seed)
  }

  # error covariance matrix
  Sigma_u <- matrix(
    c(sigma_s^2, rho, rho, sigma_p^2),
    nrow = 2,
    byrow = TRUE
  )
  L.u <- chol(Sigma_u)

  # cross-over assignment
  instruction_fun <- function(classroom, week) {
    if (week == 1 && classroom == "blue") {
      return("traditional")
    }
    if (week == 1 && classroom == "red") {
      return("curated")
    }
    if (week == 2 && classroom == "blue") {
      return("curated")
    }
    if (week == 2 && classroom == "red") {
```

```
      return("traditional")
    }
  }

  # students and design
  students <- data.frame(
    id = 1:(length(classes) * n_per_class),
    classroom = rep(classes, each = n_per_class)
  )

  design <- expand.grid(
    id = students$id,
    week = 1:2,
    time = c("pre", "post"),
    KEEP.OUT.ATTRS = FALSE,
    stringsAsFactors = FALSE
  ) %>%
    left_join(students, by = "id") %>%
    mutate(
      instruction = mapply(instruction_fun, classroom, week),
      week = factor(week),
      time = factor(time, levels = c("pre", "post")),
      instruction = factor(instruction, levels = c("traditional", "curated")),
      classroom = factor(classroom)
    )

  # student abilities
  ability_student_s <- rnorm(
    nrow(students),
    mean = ability_mean + test_effect["spelling"],
    sd = ability_sd
  )
  ability_student_p <- rnorm(
    nrow(students),
    mean = ability_mean + test_effect["phonics"],
    sd = ability_sd
  )
  names(ability_student_s) <- students$id
  names(ability_student_p) <- students$id

  clip01 <- function(x, lo = clip_lo, hi = clip_hi) pmin(pmax(x, lo), hi)

  alpha.mat <- matrix(
    c(ability_student_s[design$id], ability_student_p[design$id]),
    ncol = 2
  )
  lambda <- matrix(
    c(lambda_week[design$week], lambda_week[design$week]),
    ncol = 2
  )
  kappa <- matrix(
    c(kappa_class[design$classroom], kappa_class[design$classroom]),
    ncol = 2
  )
  post = 1 * (design$time == "post")
  cur = 1 * (design$instruction == "curated")

  mu = alpha.mat +
    lambda +
```

```
    kappa +
    post * beta1 +
    cur * beta2 +
    post * cur * beta3

  Z = matrix(rnorm(nrow(design) * 2), ncol = 2)

  Y <- Z %*% L.u + mu

  long_data <- design %>%
    mutate(
      spelling = clip01(Y[, 1]),
      phonics = clip01(Y[, 2])
    )

  long_data
}
```

D.3 Model Code

```
#Fixed effects model
fit.lm <- function(dat) {
  X <- model.matrix(
    ~ 0 +
      as.factor(id) +
      as.factor(week) +
      as.factor(time) * as.factor(instruction),
    data = dat
  )
  m <- lm(
    cbind(spelling, phonics) ~ 0 + X,
    data = dat
  )
  Uhat <- residuals(m)
  Sigma.u.hat <- crossprod(Uhat) / df.residual(m)
  Var.B <- Sigma.u.hat * solve(crossprod(X))[ncol(X), ncol(X)]
  list(
    est = coef(m)[ncol(X), ],
    var = Var.B,
    n = nrow(X),
    p = ncol(X)
  )
}

wald.test <- function(estimates, alpha = 0.05) {
  v.i <- solve(estimates$var)
  w <- t(estimates$est) %*% v.i %*% estimates$est
  (1 - pchisq(w, df = 2)) < alpha
}

f.test <- function(estimates, alpha = 0.05) {
  v.i <- solve(estimates$var)
  w <- t(estimates$est) %*% v.i %*% estimates$est

  n = estimates$n
  p = estimates$p
  nu <- (n - p - 1)
```

```
  f <- nu / (2 * (n - p)) * w

  (1 - pf(f, df1 = 2, df2 = nu)) < alpha
}
```

D.4 Bayesian Model Code

```
library(rstanarm)

source("dgp.R")

ellas.class <- generate_classroom_data(
  rho = 0.9,
  sigma_s = 2,
  sigma_p = 2,
  ability_sd = 5,

  #This is effectively deviations in mu_0
  test_effect = c(spelling = 2, phonics = -2),

  lambda_week = c(`1` = 0, `2` = 0),
  kappa_class = c(blue = 5, red = 0),

  beta1 = c(spelling = 0, phonics = 0), # effect of post indicator
  beta2 = c(spelling = 0, phonics = 0), # main effect of curated
  beta3 = c(spelling = 3, phonics = 3), # post x curated gain

  seed = 666
)

source("data.R")

fit_mv <- stan_mvmer(
  formula = list(
    spelling ~ 0 +
      Age +
      factor(week) +
      factor(classroom) +
      factor(time) * factor(instruction) +
      (1 | id),

    phonics ~ 0 +
      Age +
      factor(week) +
      factor(classroom) +
      factor(time) * factor(instruction) +
      (1 | id)
  ),
  data = ellas.class,
  family = gaussian(),
  prior = normal(0, 5, autoscale = TRUE),
  prior_intercept = normal(0, 5, autoscale = TRUE),
  prior_covariance = decov(regularization = 2),
  chains = 4,
  iter = 100000,
  cores = 4,
  refresh = T
```

```
)

allsamps <- as.matrix(fit_mv)
saveRDS(allsamps, "mcmc-samples.rds")
```

D.5 Simulation Code

```
source("dgp.R")
source("models.R")

#Simulation size and power
plan <- expand.grid(
  #different beta 3
  spelling = -5:5,
  phonics = -5:5
)
simulate <- function(plan, B = 1000, seed = NULL) {
  if (!is.null(seed)) {
    set.seed(seed)
  }
  plan$phat.w <- plan$phat.f <- 0
  for (i in 1:nrow(plan)) {
    rejections.w <- matrix(0, nrow = B, ncol = 1)
    rejections.f <- matrix(0, nrow = B, ncol = 1)
    plan.i <- plan[i, ]
    for (b in 1:B) {
      ellas.class <- generate_classroom_data(
        rho = 0.9,
        sigma_s = 2,
        sigma_p = 2,
        ability_sd = 5,

        #This is effectively deviations in mu_0
        test_effect = c(spelling = 2, phonics = -2),

        lambda_week = c(`1` = 0, `2` = 0),
        kappa_class = c(blue = 5, red = 0),

        beta1 = c(spelling = 0, phonics = 0), # effect of post indicator
        beta2 = c(spelling = 0, phonics = 0), # main effect of curated
        beta3 = c(spelling = plan.i$spelling, phonics = plan.i$phonics), # post x
            curated gain

        seed = NULL
      )

      fe <- fit.lm(ellas.class)
      rejections.w[b, 1] <- wald.test(fe)
      rejections.f[b, 1] <- f.test(fe)
    }
    plan$phat.w[i] = mean(rejections.w[, 1])
    plan$phat.f[i] = mean(rejections.f[, 1])
    print(round(100 * i / nrow(plan)) / 100)
  }
  plan$MCSE.w <- sqrt(plan$phat.w * (1 - plan$phat.w) / B)
  plan$MCSE.f <- sqrt(plan$phat.f * (1 - plan$phat.f) / B)
  plan
```

```r
}

rejection.rates <- simulate(
  plan,
  B = 2000
)
saveRDS(rejection.rates, "rejection-rates-sim.RDS")
```

D.6 Code for Analysis

```r
#Frequentist analysis
source("data.R")
source("models.R")

X <- model.matrix(
  ~ 0 +
    as.factor(id) +
    as.factor(classroom) +
    Age +
    as.factor(time) * as.factor(instruction),
  data = ellas.class
)
m <- lm(
  cbind(spelling, phonics) ~ 0 + X,
  data = ellas.class
)
summary(m)

#Spelling | Coef | SE
#Age | 22.917 | 9.186
#Class Red | -19.025 | 6.108
#\beta_1 | 13.091 | 7.834
#\beta_2 | -1.586 | 7.835
#\beta_3 | 9.636 | 11.078

#Phonics | Coef | SE
#Age | 16.542 | 4.715
#Class Red | -7.738 | 5.403
#\beta_1 | 1.000 | 4.021
#\beta_2 | -3.410 | 4.022
#\beta_3 | 6.818 | 5.687

estimates <- fit.lm(ellas.class)
#F-test
v.i <- solve(estimates$var)
w <- t(estimates$est) %*% v.i %*% estimates$est

#Chisq test
w
# 1.44
1 - pchisq(w, 2)
#0.48

n = estimates$n
p = estimates$p
nu <- (n - p - 1)
f <- nu / (2 * (n - p)) * w
```

```r
f
#0.7169982
(1 - pf(f, df1 = 2, df2 = nu))
#0.4905246

#Bayesian analysis
allsamps <- readRDS("mcmc-samples.rds")
relevant.coefs <- allsamps[, c(1:14)]

#Posterior Estimates
cbind(
  colMeans(relevant.coefs),
  apply(relevant.coefs, 2, median),
  apply(relevant.coefs, 2, \(x) HDInterval::hdi(x)) |> t()
)

#Spelling | Mean | Median | 95% HPD
#Age | 17.37 | 17.38 | 2.34, 32.398
#Class Red | -17.40 | -17.42 | -28.79, -5.898
#\beta_1 | 13.124 | 13.139 | -1.8146, 28.0868
#\beta_2 | -1.5608 | -1.5621 | -16.706, 13.2158
#\beta_3 | 9.599 | 9.625 | -11.36, 30.83

#Phonics | Mean | Median | 95% HPD
#Age | 13.687 | 13.69 | 4.373, 22.65
#Class Red | -8.00 | -7.99 | -14.11697, -1.88
#\beta_1 | 0.996 | 0.982 | -6.911, 8.735574
#\beta_2 | -3.39778 | -3.3988075 | -11.3178, 4.25631
#\beta_3 | 6.810 | 6.8184 | -4.1028, 17.918

#Bayesian hypothesis testing
beta3 <- relevant.coefs[, c(7, 14)]
#Pr(beta_3,spelling > 0, beta_3,phonics > 0)
mean(beta3[, 1] > 0 & beta3[, 2] > 0)
# = 0.72305
#1-Pr(beta_3,spelling < 0, beta_3,phonics < 0)
1 - mean(beta3[, 1] < 0 & beta3[, 2] < 0)
# = 0.9785
```

D.7 Reproduce Figures

```r
##themes.R
library(ggplot2)
library(patchwork)
library(sysfonts)

font_add("cm", regular = "fonts/cmunrm.ttf")
showtext::showtext_auto()

theme_paper <- function(base_size = 14, text_size = 14, base_family = "cm") {
  theme_minimal(base_size = base_size, base_family = base_family) %+replace%
    theme(
      # Text
      plot.title = element_text(
        face = "bold",
        size = text_size + 2,
        hjust = 0.5,
```

```r
      margin = margin(b = 6)
    ),
    plot.subtitle = element_text(
      size = text_size,
      hjust = 0.5,
      margin = margin(b = 8)
    ),
    plot.caption = element_text(
      size = text_size - 2,
      hjust = 1,
      margin = margin(t = 6)
    ),
    axis.title.x = element_text(size = text_size, margin = margin(t = 8)),
    axis.title.y = element_text(
      size = text_size,
      angle = 90,
      margin = margin(r = 8)
    ),
    axis.text = element_text(size = text_size - 1),

    # Panel & grid
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank(),
    panel.grid.minor.y = element_blank(),
    #panel.grid.major.y = element_line(linewidth = 0.3, color = "grey85"),
    panel.border = element_rect(fill = NA, color = "grey60", linewidth = 0.4),

    # Legend
    legend.position = "bottom",
    legend.title = element_text(size = base_size - 1, face = "bold"),
    legend.text = element_text(size = base_size - 2),
    legend.key.width = unit(1.2, "lines"),

    # Strips (for facets)
    strip.background = element_rect(fill = "grey90", color = NA),
    strip.text = element_text(face = "bold", size = base_size - 1),

    # Margins
    plot.margin = margin(t = 8, r = 8, b = 8, l = 8)
  )
}

paper_palette <- c(
  "#5e57be",
  "#d94202",
  "#1b9e77",
  "#e7298a",
  "#66a61e",
  "#e6ab02"
)

scale_color_paper <- function(...) {
  scale_color_manual(values = paper_palette, ...)
}

scale_fill_paper <- function(...) {
  scale_fill_manual(values = paper_palette, ...)
}

```

```r
scale_fill_gradient2_paper <- function(
  low = "#7570b3",
  mid = "white",
  high = "#d95f02",
  midpoint = 0,
  ...
) {
  scale_fill_gradient2(
    low = low,
    mid = mid,
    high = high,
    midpoint = midpoint,
    ...
  )
}

##figures.R
source("themes.R")
source("dgp.R")

ellas.class <- generate_classroom_data(
  rho = 0.9,
  sigma_s = 2,
  sigma_p = 2,
  ability_sd = 5,

  #This is effectively deviations in mu_0
  test_effect = c(spelling = 2, phonics = -2),

  lambda_week = c(`1` = 0, `2` = 0),
  kappa_class = c(blue = 5, red = 0),

  beta1 = c(spelling = 0, phonics = 0), # effect of post indicator
  beta2 = c(spelling = 0, phonics = 0), # main effect of curated
  beta3 = c(spelling = 3, phonics = 3), # post x curated gain

  seed = 666
)
delta <- 0.04

arrows.df <- ellas.class %>%
  mutate(time = tolower(time)) %>%
  dplyr::select(id, week, instruction, time, spelling, phonics) %>%
  tidyr::pivot_wider(
    names_from = time,
    values_from = c(spelling, phonics)
  ) %>%
  mutate(
    x_end = spelling_pre + (1 - delta) * (spelling_post - spelling_pre),
    y_end = phonics_pre + (1 - delta) * (phonics_post - phonics_pre)
  )

base.dat <- ellas.class %>%
  filter(time == "pre") %>%
  mutate(
    shape_class = ifelse(
      instruction == "traditional",
      ifelse(week == 1, "open_circle", "closed_circle"),
      ifelse(week == 1, "open_triangle", "closed_triangle")
```

```
    )
  )

ellas.class.dgp <- ggplot(
  base.dat,
  aes(
    x = spelling,
    y = phonics
  )
) +
  # arrows
  geom_segment(
    data = arrows.df,
    aes(
      x = spelling_pre,
      y = phonics_pre,
      xend = x_end,
      yend = y_end
    ),
    arrow = arrow(length = unit(0.05, "cm")),
    color = "#CCC",
    linewidth = 0.15,
    inherit.aes = FALSE
  ) +
  geom_point(
    aes(
      color = classroom,
      shape = shape_class
    ),
    stroke = 0.2,
    show.legend = FALSE,
  ) +
  # labels: white for week 2
  geom_text(
    data = dplyr::filter(base.dat, week == 2),
    aes(label = id),
    color = "white",
    size = 8,
    vjust = 0.5,
    hjust = 0.5,
    show.legend = FALSE
  ) +
  # labels: black for week 1
  geom_text(
    data = dplyr::filter(base.dat, week == 1),
    aes(label = id),
    color = "black",
    size = 8,
    vjust = 0.5,
    hjust = 0.5,
    show.legend = FALSE
  ) +
  scale_color_paper() +
  scale_shape_manual(
    values = c(
      open_circle = 1, # open circle
      open_triangle = 2, # open triangle
      closed_circle = 16, # filled circle
      closed_triangle = 17 # filled triangle
```

```
    )
  ) +
  labs(
    x = "Spelling",
    y = "Phonics",
    shape = "Instruction",
    color = "Class"
  ) +
  theme_paper(base_size = 10, text_size = 64) +
  theme(
    legend.title = element_text(size = 64),
    legend.text = element_text(size = 64)
  )

ggsave(
  "Figures/dgp.png",
  ellas.class.dgp,
  width = 5,
  height = 2.8,
  dpi = 900
)

###### Real Data ######

source("data.R")

jitter.sd <- 3
ellas.class <- ellas.class %>%
  group_by(time == "Pre") %>%
  mutate(
    spelling = spelling + rnorm(n(), sd = jitter.sd),
    phonics = phonics + rnorm(n(), sd = jitter.sd)
  ) %>%
  ungroup()

delta <- 0.04
arrows.df <- ellas.class %>%
  mutate(
    time = tolower(time)
  ) %>%
  dplyr::select(id, week, instruction, time, spelling, phonics) %>%
  tidyr::pivot_wider(
    names_from = time,
    values_from = c(spelling, phonics)
  ) %>%
  mutate(
    x_end = spelling_pre + (1 - delta) * (spelling_post - spelling_pre),
    y_end = phonics_pre + (1 - delta) * (phonics_post - phonics_pre)
  )
base.dat <- ellas.class %>%
  filter(time == "Pre") %>%
  mutate(
    shape_class = ifelse(
      instruction == "traditional",
      ifelse(week == 1, "open_circle", "closed_circle"),
      ifelse(week == 1, "open_triangle", "closed_triangle")
    )
  )

```

```r
ellas.class.plt <- ggplot(
  base.dat,
  aes(
    x = spelling,
    y = phonics
  )
) +
  # arrows
  geom_segment(
    data = arrows.df,
    aes(
      x = spelling_pre,
      y = phonics_pre,
      xend = x_end,
      yend = y_end
    ),
    arrow = arrow(length = unit(0.05, "cm")),
    color = "#CCC",
    linewidth = 0.15,
    inherit.aes = FALSE
  ) +
  geom_point(
    aes(
      color = classroom,
      shape = shape_class
    ),
    stroke = 0.2,
    show.legend = FALSE,
  ) +
  # labels: white for week 2
  geom_text(
    data = dplyr::filter(base.dat, week == 2),
    aes(label = id),
    color = "white",
    size = 8,
    vjust = 0.5,
    hjust = 0.5,
    show.legend = FALSE
  ) +
  # labels: black for week 1
  geom_text(
    data = dplyr::filter(base.dat, week == 1),
    aes(label = id),
    color = "black",
    size = 8,
    vjust = 0.5,
    hjust = 0.5,
    show.legend = FALSE
  ) +
  scale_color_paper() +
  scale_shape_manual(
    values = c(
      open_circle = 1, # open circle
      open_triangle = 2, # open triangle
      closed_circle = 16, # filled circle
      closed_triangle = 17 # filled triangle
    )
  ) +
  labs(
```

```
    x = "Spelling",
    y = "Phonics",
    shape = "Instruction",
    color = "Class"
  ) +
  theme_paper(base_size = 10, text_size = 64) +
  theme(
    legend.title = element_text(size = 64),
    legend.text = element_text(size = 64)
  )

ggsave(
  "Figures/observed.png",
  ellas.class.plt,
  width = 5,
  height = 2.8,
  dpi = 900
)

##### Simulation Study Results #####

rr <- readRDS("rejection-rates-sim.RDS")
lims <- range(c(rr$phat.w, rr$phat.f), na.rm = TRUE)

rr.plt.W <- ggplot(rr, aes(spelling, phonics, fill = phat.w)) +
  geom_raster(interpolate = FALSE) +
  geom_text(aes(label = sprintf("%.2f", phat.w)), color = "white", size = 16) +
  scale_fill_viridis_c(option = "turbo", limits = lims) +
  coord_equal() +
  labs(fill = "Rejection Rates", x = "Spelling Effect", y = "Phonics Effect") +
  theme_paper(base_size = 10, text_size = 64) +
  theme(legend.position = "none")

rr.plt.F <- ggplot(rr, aes(spelling, phonics, fill = phat.f)) +
  geom_raster(interpolate = FALSE) +
  geom_text(aes(label = sprintf("%.2f", phat.f)), color = "white", size = 16) +
  scale_fill_viridis_c(option = "turbo", limits = lims) +
  coord_equal() +
  labs(fill = "Rejection Rates", x = "Spelling Effect", y = "Phonics Effect") +
  theme_paper(base_size = 10, text_size = 64) +
  theme(legend.position = "none")

ggsave(
  "Figures/rrw.png",
  rr.plt.W,
  width = 4,
  height = 4,
  dpi = 900
)

ggsave(
  "Figures/rrf.png",
  rr.plt.F,
  width = 4,
  height = 4,
  dpi = 900
)

##### BAYESIAN ESTIMATION #####
```

```r

allsamps <- readRDS("mcmc-samples.rds")

curriculum.effects <- allsamps[, c(7, 14)]
x <- allsamps[, 7] #Spelling
y <- allsamps[, 14] #Phonics
colMeans(curriculum.effects)
# mean(curriculum.effects[,1] > 0 | curriculum.effects[,2] > 0)

dens <- kde2d(x, y, n = 50)
df <- with(
  dens,
  expand.grid(x = x, y = y) |>
    mutate(z = as.vector(z))
)

bayes.plt <- ggplot(df, aes(x, y, fill = z)) +
  geom_raster(interpolate = T) +
  scale_fill_viridis_c(option = "turbo") +
  coord_equal() +
  geom_vline(xintercept = 0, color = "white", linetype = 2) +
  geom_hline(yintercept = 0, color = "white", linetype = 2) +
  labs(fill = "Density", x = "Spelling Effect", y = "Phonics Effect") +
  theme_paper(base_size = 10, text_size = 64) +
  theme(
    legend.position = "none"
  )

ggsave(
  "Figures/bayes.png",
  bayes.plt,
  width = 5,
  height = 2.8,
  dpi = 900
)

final_data <- readRDS("data/final_project_scores.rds")
post.scores.by.instruction <-
  final_data %>%
  group_by(instruction, test, time, classroom) %>%
  summarize(mean_score = mean(score), .groups = "drop") %>%
  ggplot(aes(
    time,
    mean_score,
    group = interaction(instruction, classroom),
    color = classroom
  )) +
  geom_line() +
  geom_point(aes(shape = instruction)) +
  scale_color_paper() +
  facet_wrap(
    ~test,
    labeller = labeller(
      test = c(
        "phonics" = "Phonics",
        "spelling" = "Spelling"
      )
    )
  ) +
```

```
  labs(
    title = "Pre/Post Scores by Instruction Type",
    y = "Mean Score (%)",
    color = "Classroom",
    shape = "Instruction Type"
  ) +
  theme_paper(base_size = 10, text_size = 64) +
  theme(
    legend.title = element_text(size = 64),
    legend.text = element_text(size = 64),
    strip.text.x = element_text(size = 64),
    strip.text.y = element_text(size = 64)
  )
ggsave(
  "Figures/post_scores_by_instruction.png",
  post.scores.by.instruction,
  width = 6,
  height = 2.8,
  dpi = 900
)
```