Studies in Educational Evaluation 89 (2026) 101594

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

A comparative analysis of difference-in-differences estimators in observational educational evaluation research

Princy Sebastian [1], Yan Zhang [*,1], Darrell M. Hull

Department of Educational Psychology, College of Education, University of North Texas, United States

ARTICLE INFO

Keywords:
Causal inference
Difference-in-Differences (DiD)
Linear Mixed Effects
Quasi-experimental design
Staggered design

ABSTRACT

The present study illustrates implementation of Difference-in-Differences (DiD) estimators for a multi-cohort, pretest-posttest evaluation of a guided reading intervention. We assessed intervention effects using student outcome data from a series of four replicated cohorts, comparing four analytical frameworks: (1) OLS regression as a baseline, (2) OLS with cluster robust standard errors (CRSE), (3) linear mixed effects (LME) with random intercepts model, and (4) a staggered DiD estimator. While results consistently suggest treated students experienced gains in reading and vocabulary, the staggered DiD approach provides more robust validation by accounting for heterogeneous treatment timing and avoiding "forbidden comparisons." The present study demonstrates how estimator choice impacts causal inference within complex, unbalanced designs common in applied educational research.

1. Introduction

The Difference-in-Differences (DiD) estimator method can be an efficient analytical procedure for educational researchers and evaluators in answering causal inference questions (Corral & Yang, 2024). These estimators typically provide unbiased treatment effect estimates when, in the absence of treatment, average outcomes for treatment and the comparison groups follow parallel trends over time (O'Neill et al., 2016) prior to treatment. Difference-in-Differences is popular in econometrics and in health outcomes studies yet rarely implemented in educational research or evaluation (Angrist & Pischke, 2009).

Educational evaluation faces two primary structural challenges that complicate traditional DiD applications. First, educational data is inherently nested (e.g., students clustered within classrooms or cohorts), and failing to account for this dependency leads to deflated standard errors and an inflated risk of Type I errors. Second, interventions in school districts are rarely implemented simultaneously across all campuses; rather, they often follow a staggered adoption or "wave" rollout. Traditional 2×2 DiD and Two-Way Fixed Effects (TWFE) models are increasingly recognized as inadequate for staggered designs, as they may lead to "forbidden comparisons" where early-treated units serve as controls for later-treated units.

The purpose of the present study is to provide a methodological demonstration of how different DiD estimators perform when applied to staggered, longitudinal educational data. Specifically, we utilize a multi-cohort dataset involving a small-group reading intervention for English Language Learners (ELLs) to compare the performance of four analytical frameworks: Ordinary Least Squares (OLS), Cluster-Robust Standard Errors (CRSE), Linear Mixed Effects (LME), and the Callaway and Sant'Anna (2021) staggered DiD estimator.

2. DiD in educational research

For interventional researchers, randomized controlled trial (RCT) designs are the first choice, as the gold standard, but where RCTs are not possible or feasible, quasi-experimental designs may be the next best choice (Kenny, 1975). Consequently, DiD is suitable data analytic method for quasi-experimental data (Angrist & Pischke, 2009; Horii et al., 2021; Corral & Yang, 2024). DiD uses longitudinal data from treatment and comparison groups to obtain an appropriate counterfactual to estimate a causal effect, i.e., in absence of treatment, where unobserved differences between treatment and comparison groups should be the same overtime (Card and Krueger, 2000; Lechner, 2011; Snow, 1856). In an educational context, this usually involves a treatment group (T) and a comparison group (C) measured across multiple time points. The grand mean difference of the two group differences

* Correspondence to: Department of Educational Psychology, College of Education, University of North Texas, 1155 Union Circle, Denton, TX 76201, United States.
E-mail address: Yan.Zhang@unt.edu (Y. Zhang).

[1] These authors contributed equally to this work and share first authorship.

https://doi.org/10.1016/j.stueduc.2026.101594
Received 15 October 2024; Received in revised form 10 February 2026; Accepted 2 March 2026
Available online 5 March 2026

across time is the difference-in-differences estimate or Average Treatment Effect (Furquim et al., 2019; Wooldridge, 2015). In Eq. (1), β_3 represents the Average Treatment Effect (ATE).

$$Y_{it} = \beta_0 + \beta_1 \text{ Group}_i + \beta_2 \text{ Time}_t + \beta_3 \text{ (Group*Time)}_{it} + \epsilon_{it} \quad (1)$$

Notes. Group and Time are dummy variables (Group: 0 = Comparison, 1 = Treatment; Time: 0 = Pretreatment, 1 = Posttreatment); Y: outcome variable; β1: Initial difference, marginal effect of treatment (Group) at time = 0; β2: Baseline change over time, marginal effect of time at treatment (Group) = 0; β3: Treatment effect

To estimate an ATE, it is necessary to assume the average outcome that would have occurred in the absence of treatment for the treated units. However, this identifying assumption cannot be verified, as the counterfactual outcome cannot be directly observed. Here, one should consider two distinct identifying assumptions. First, it can be assumed that the change in Y_0 between periods t and t' is independent of whether the unit is assigned to the treated group, after conditioning on observables (Angrist & Pischke 2009). This parallel trends assumption requires that in the absence of treatment, the average outcomes for the treated and comparison groups would have followed the same trajectory over time (Rietveld et al., 2019). Critical questions to consider in this phase are: (a) Are the treatment and comparison group comparable at the outset? and (b) Are both groups trending the same way prior to the start of treatment or intervention? While graphical analysis provides a preliminary check of parallel trends, modern causal inference standards increasingly require formal statistical validation through event study frameworks to rule out pre-treatment differences (Callaway and Sant'Anna, 2021; Corral and Yang, 2024). If the two groups are different on observed characteristics initially, then there will not be a graphical parallel trend, and one cannot assume the difference is due to the treatment. In this scenario, regression estimates will be inconsistent, and causal inference will be invalid due to the endogeneity problem (Sieweke & Santoni, 2020) as shown below with non-parallel trend equations.

$$Y_{it} = \beta_0 + \beta_1 \text{ Group}_i + (\beta_2 {}_+ \beta_4 \text{ Group) Time}_t + \beta_3 \text{ (Group*Time)}_{it} + \epsilon_{it} \quad (2)$$

rearrange the terms; change in treatment is unobserved so gets grouped with error term

$$Y_{it} = \beta_0 + \beta_1 \text{ Group}_i + \beta_2 \text{ Time}_t + \beta_3 \text{ (Group*Time)}_{it} + \beta_4 \text{ Group*Time}_t + \epsilon_{it} \quad (3)$$

The second assumption is that the stable unit treatment value assumption (SUTVA) must hold, implying that the potential outcomes for any unit are unaffected by the treatment assignment of other units. This ensures there are no spillover effects between cohorts and that the treatment does not induce "anticipatory effects" where students or teachers change behavior in periods prior to the actual intervention. Deviation from these assumptions, particularly non-parallel trends, leads to endogeneity problems that render OLS estimates inconsistent (Sieweke & Santoni, 2020).

2.1. Addressing clustering and dependencies

Traditional DiD applications in education often overlook the correlation of errors within clusters (e.g., students nested within classrooms), leading to deflated standard errors and an inflated Risk of Type I errors (Bertrand et al., 2004). What Works Clearinghouse (WWC, 2022) standards suggest that DiD is acceptable for baseline adjustment when outcome measures have the same measurement scale at both pretest and posttest, and the correlation is greater than.60 . Accounting for dependent observations as in repeated measurements or in cluster/nested data is necessary to avoid misestimated standard errors leading to questionable inferential statistics (Huang & Li, 2021). Remedies include using cluster robust standard errors, which are more tolerant of certain assumption violations of the statistical model (Cameron & Miller, 2015). Using Cluster Robust Standard Errors (CRSE) would be an analytic method to minimize standard error misestimation with DiD if an important clustered scenario exists. When regression residuals are heteroskedastic, robust standard errors improve standard errors (Angrist & Pischke 2009). However, standard error will be calculated with less bias, which may have implications on inferential statistics (McNeish, et al., 2017). With added robust standard error calculations, one will still obtain parameter estimates similar to OLS estimates because quantities used in these calculations such as sum of squares, and expected mean squares are unaffected by the statistical correction to the standard error estimates (Hayes & Cai, 2007). However, even RSEs can underestimate standard errors for group-level variables when the number of clusters is fewer than 50 (Huang and Li, 2021; Snijders and Bosker, 1993).

Another remedy method is to use a mixed effects model, which does not have an independence assumption (Heck et al., 2013). By associating common random effects to observations sharing the same level of a classification factor, mixed-effects models flexibly represent the covariance structure induced by the grouping of the data (Pinheiro & Bates, 2000). In addition, with random effects it is assumed that unit effects are uncorrelated with covariates (Meteyard & Davies, 2020). If between subjects' variance is substantial (and statistically significant) then the random-effects model is more appropriate.

Following linear mixed effects modeling is proposed to assess the outcome scores (time/pre-post) nested within students. In each model, Time (i) is nested within each group (j). Time is represented at Level 1 and Group is at Level 2. A linear mixed effects model is presented below where Y_{ij} is the outcome variable, i referring to Test (level 1), j referring to the Student (level 2).

$$Y_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Group}_j + \beta_3 (\text{Time* Group}_{ij}) + e_{ij} + u_j \quad (4)$$

The random terms (e_{ij} and u_j) are the error terms. Level 1 error term (e_{ij}) represents time test for each student while level 2 error term (u_j) represents the deviation for each student from the grand group mean.

If there is no significant correlated variability due to grouping, then mixed effects modeling may not be needed (Heck et al., 2013). The ratio of between group variance to total variance provides the percentage of outcome variance explained by between group or clustering. This ratio is known as the Intraclass Correlation Coefficients (ICCs). As a rule of thumb, an ICC of 10% or above warrants a multi-level model (Heck et al., 2013).

2.2. Staggered adoption design

In real life educational evaluations, interventions follow a staggered adoption pattern rather than a simultaneous rollout across all participants (Roth et al., 2023). While traditional Two-Way Fixed Effects (TWFE) models, which include unit and time fixed effects, have long been the standard for analyzing such multi-period data, recent econometric literature has identified significant risks of bias in these estimators when treatment effects are heterogeneous across cohorts or time (Baker et al., 2022; Goodman-Bacon, 2021; Wing et al., 2024; de Chaisemartin and D'Haultfoeuille, 2020).

A primary consequence of applying traditional DiD to staggered designs is the "negative weighting" problem (Goodman-Bacon, 2021; Sun and Abraham, 2021). As demonstrated by de Chaisemartin, D'Haultfoeuille, 2020) and Goodman-Bacon (2021), the TWFE estimator functions as a weighted average of all possible 2 × 2 DiD comparisons within the panel (Baker et al., 2022). In a staggered setup, this often results in "forbidden comparisons", where units treated in earlier periods are utilized as controls for units treated in later periods (Callaway and Sant'Anna, 2021; Wing et al., 2024; de Chaisemartin and D'Haultfoeuille, 2020). If treatment effects evolve over time, the changes in outcomes for the early-treated group are subtracted from the changes in the later-treated group. This can produce "negative weights," where the resulting Average Treatment Effect (ATE) may be signed differently than every individual group-time effect, leading to misleading or entirely localized conclusions that do not reflect the true impact of the intervention (Baker et al., 2022; Goodman-Bacon, 2021; de Chaisemartin and D'Haultfoeuille, 2020).

To address these limitations, modern causal inference frameworks, such as the staggered DiD estimator proposed by Callaway and Sant'Anna (2021), have been developed to isolate "clean" comparisons (Borusyak et al., 2024; Sun and Abraham, 2021; Wing et al., 2024). This approach estimates the group-time average treatment effect on the treated (ATT (g, t)) by comparing a treated cohort solely to a strictly defined control group of "never-treated" or "not-yet-treated" units (Callaway & Sant'Anna, 2021). By avoiding the contamination of the counterfactual with previously treated units, this method ensures that the estimated effects are robust to temporal heterogeneity (Gardner, 2022; Wing et al., 2024). In the context of this study, the use of the Callaway and Sant'Anna estimator allows for a more granular validation of the parallel trends assumption through an event-study framework, providing a dynamic perspective on student growth that static OLS or LME models may fail to capture (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).

The primary objective of this study is to provide a rigorous causal evaluation of the literacy intervention on student linguistic outcomes within a multi-cohort, staggered-adoption framework. While previous evaluations often rely on traditional pooled models, the potential for treatment timing heterogeneity necessitates a more granular analytical approach to isolate "clean" treatment effects and evaluate their sustainability over time. By comparing robust staggered DiD estimators against traditional benchmarks, this research aims to clarify not only the immediate impact of the intervention, but also the consistency of these gains across different implementation waves. To achieve this, the following research questions guide our analysis:

2.3. *Research questions*

1. What are the parameter estimates, standard errors, and effect size estimates of MAP reading and vocabulary RIT scores of the treatment group compared to the business-as-usual instruction group with the DiD estimator using OLS regression, OLS regression with cluster robust standard errors, a linear mixed effects model with random intercepts and, staggered DiD estimator?
2. To what extent are differences in observed effects associated with the four different analytic methods, and what are the implications for practice in applied research and evaluation analyses?

3. Methodology

3.1. *Participants*

Third grade teachers and their students were recruited from twenty-four schools across a large school district in the Southern United States. The study lasted four years. Each year a new cohort of teachers and students participated in the study (Cohort 1: 2018–19, Cohort 2: 2019–20, Cohort 3: 2020–21, Cohort 4: 2021–22).

3.2. *Student demographics*

The combined student sample size from all cohorts is 1107. The treatment group included 71 students, while the comparison group included 1036 students. Specifically, Cohort 1 consisted of 20 treated and 201 comparison students; Cohort 2 included 12 treated and 136 comparison students; Cohort 3 comprised 31 treated and 134 comparison students; and Cohort 4 included 8 treated and 565 comparison students.

These students were classified by the district as Limited English Proficient (LEP) students, where 78% of LEP students spoke Spanish. The other 22% spoke Vietnamese, Arabic, Somali, Korean, Tamil and Urdu. Students were 52% female, and 91% of students were eight years of age in September as school started, and the remaining 8% were nine.

3.2.1. *Teacher demographics*

There were 20 treatment teachers who received professional development over the four cohorts and 11 participated in provision of the intervention to students. Per the district coordinator, the treatment and comparison group teachers had similar levels of experience and certifications.

3.2.2. *Sampling procedures*

Teacher participation occurred through self-selection. The school district identified the treatment and comparison group teachers and their respective students from a pool of 500 third grade students each year. The study included a small treatment group of approximately 5 self-selected teachers and their 25 students and a larger comparison group of 150–200 students in third grade that are classified as being Limited English Language proficient (LEP). In general, third grade teachers in the district have class sizes of 20–28 students, and approximately five students in any teacher's classroom are "tier 2" students. Under the Response to Intervention (RTI) framework, these are students that are not making adequate progress in the regular classroom and require intensive instruction, specifically in English language acquisition, and reading. These students who were identified as LEP were study participants for the dependent variable outcomes. Under the approved IRB protocol, the district supplied de-identified data on student participants to carry out an analysis.

3.2.3. *Measures and covariates*

Measures of Academic Progress (MAP) scores are the key student outcome data. MAP is a norm-referenced tool that measures student growth over time. MAP (created by Northwest Evaluation Association or NWEA) is a computer-adaptive test, meaning a computer adjusts the test to create a unique test for each student. The difficulty of each question is determined by how well the student answered the previous question. MAP test scores allow teachers to customize instructions to fit each student's literacy needs.

The present study used only the reading and vocabulary sub scales of the MAP assessment since these scales were most closely aligned with the intervention. Specifically, reading and vocabulary MAP RIT scores are analyzed. RIT stands for Rasch UnIT and is a measurement scale, with an equal-interval scale developed to simplify the interpretation of test scores. A RIT score also has the same meaning regardless of the grade or age of the student (NWEA website). Baseline RIT scores were obtained at the end of second grade and then in the fall, winter and in the spring of third grade for each cohort. The student intervention started in fall and lasted for 18 weeks during the third grade. Pretreatment scores were defined as the assessments conducted at the end of the second-grade year and the beginning of the third-grade year. Post-treatment scores comprised the assessments conducted at the conclusion of the winter and spring semesters of the third-grade year.

3.2.4. *Conditions and design*

The study design is a replicated cohort structure with staggered longitudinal unbalanced design containing a large untreated control group and an treatment group with small group intervention (Cook et al., 2002). The primary independent variable is the treatment status (intervention vs. control), and the dependent variables are the MAP Reading and Vocabulary RIT scores.

In this design, students are assessed at four longitudinal time points: a baseline assessment at the end of second grade, followed by fall, winter, and spring assessments during the third-grade intervention year. This structure creates a panel dataset where individual observations are nested within students and cohorts. Although data collection was limited to the end of second grade through the end of third grade—with no subsequent observations after Grade 3—the treatment is considered irreversible. That is, once a student cohort entered the intervention in Grade 3, they remained in the treated state for all subsequent observed periods. This alignment satisfies the key assumption of staggered adoption designs regarding treatment irreversibility (Callaway & Sant'Anna, 2021). By employing the Callaway and Sant'Anna (2021) method, this study not only accounts for heterogeneous cohort and time effects by avoiding "forbidden comparisons" but also reveals the dynamic trajectories of student achievement from multiple temporal perspectives.

Selected English Language Learners (ELLs) were pulled out of larger classroom settings and taught in small groups by trained treatment teachers. The hypothesis is that this targeted intervention would lead to superior growth trajectories in language acquisition, as reflected in higher RIT score gains compared to the comparison group.

3.3. Intervention: The teacher's professional development and coaching

The study described in this article focuses on the fourth of recommendations from a practice guide published by The United States Department of Education, through the Institute of Education Sciences: Provide small-group instructional intervention to students struggling in areas of literacy and English language development (WWC, 2022).

The training was composed of teacher professional development and in-classroom coaching. Professional development was provided for teachers in the form of a graduate-level course. Treatment teachers enrolled in a five-week graduate level face-to-face course, which focused on language, literacy development, alphabetic knowledge, print awareness, receptive language, expressive language, and small group instructional intervention with English learners using multimodal strategies.

During the course, teachers developed a one-week unit consisting of five lesson plans in which they targeted both the reading (TEKS) and linguistic (ELPS) needs of their incoming ELs. At the start of the academic year, teachers met with the project facilitator to review student data and select the members of their small group. Together they planned on meeting with their small group and created a coaching schedule for the 18 weeks of intervention. The total proposed treatment dosage is 30 min a day, 5 times a week, for 20 weeks equaling a maximum dose of 50 h of small group instruction for each school year. So, the researcher-expected fidelity of implementation dosage adherence for each treatment teacher is 50 h of small group instruction. Fidelity of implementation (FOI) data were gathered during the study. Each teacher submitted an online survey every 3 weeks on how much, how often, and how well they implemented the prescribed intervention as proof of fidelity of implementation. Also, an objective score of FOI was given by the coach in the classroom each week.

During the 18 weeks of intervention, the project facilitator observed teachers once per week during a 30-minute lesson sequence. The facilitator followed up on the observation with a debrief meeting to discuss the observed lesson following the Leverage Leadership 2.0 coaching model. The debriefing meeting occurred either later the same day or the following day during teacher planning periods or after school.

3.4. Leverage leadership 2.0 coaching model: See it, name it, do it

- See it: airtight activities that lead participants to the right conclusion mostly on their own
- Name it: lead participants to name the keys to the action; add formal language
- Do it: put the principles into practice during coaching and in the classroom

At the middle of the year, the teachers and project facilitator met to debrief as a whole group. They shared the glows and growth of their small group implementation thus far. Teachers were given the opportunity to hear ideas from their peers and how to implement them with their own small groups. It was a time to collaborate with each other once more. A more complete description of the treatment is provided in *blinded* (2025).

3.5. Data analysis

The school district provided all available student data each summer in a strict confidential manner with blinded student and teacher ID on multiple Excel files. Data was obtained, cleaned, and organized with respective group belonging. The outcome data analysis compared the differences in outcome scores for treatment and comparison group with respective pre and post treatment scores. Student data were matched on their Limited English Proficient (LEP) status prior to analysis to ensure that the comparison group is most closely matched to the selected treatment group (Rubin, 2004).

To examine how different analytic approaches perform under a staggered treatment adoption design, we conducted a comparative analysis using four methods: (1) a traditional difference-in-differences (DiD) estimator implemented via linear regression, (2) linear regression with cluster-robust standard errors, (3) a linear mixed-effects model with random intercepts, and (4) a staggered DiD estimator designed for multi-period panel data. The first three approaches relied solely on pre-treatment and post-treatment scores, thus incorporating only two time points and were implemented using SPSS. Recognizing the limitations of traditional Two-Way Fixed Effects (TWFE) in staggered designs (Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfoeuille, 2020), we further implemented the Callaway and Sant'Anna (2021) staggered DiD estimator using the "did" package (Callaway & Sant'Anna, 2021) in R (version 4.5.1). Unlike traditional models, this estimator utilizes the full longitudinal panel (four time-points) to capture the dynamic group-time average treatment effects and accounts for variation in treatment timing.

Due to data collection constraints during COVID-19, cohort 2 includes only three observed time points, as achievement data from the final semester of Grade 3 were unavailable. Additionally, achievement data for cohort 3 in Grade 2 were collected one semester earlier, resulting in the absence of observations for Spring 2020. Analyses were restricted to complete cases to ensure consistency across time points in the panel structure. While this resulted in a slight reduction of the control group sample size to satisfy the requirements of the staggered DiD estimator, the final analytic sample remains statistically sufficient. Detailed sample sizes by cohort and group are presented in Table 1.

Parallel trends assumptions were checked though both visual inspection and formal statistical testing. First, outcome trajectories (Reading and Vocabulary RIT scores) were plotted across four time points—from the end of second grade through the spring of third grade—to visually assess pre-treatment trends (see Figs. 1 and 2). Following visual inspection, formal pre-trend tests were conducted. For the traditional DiD models, interaction terms (Treat × Time) for the two pre-treatment periods (Spring 2nd Grade and Fall 3rd Grade) were analyzed. Each year the treatment began in August as the school year started. The first MAP testing for third grade took place in September.

Beyond traditional baseline models, the dependent variables were analyzed using a comparative framework: (1) OLS regression, (2) OLS with cluster-robust standard errors, and (3) linear mixed-effects (LME) models with random intercepts, all implemented in SPSS. To address the

Table 1
Sample size for staggered DiD estimator.

Year	Control	Treatment
Cohort 1	198	20
Cohort 2	133	13
Cohort 3	126	31
Cohort 4	497	8

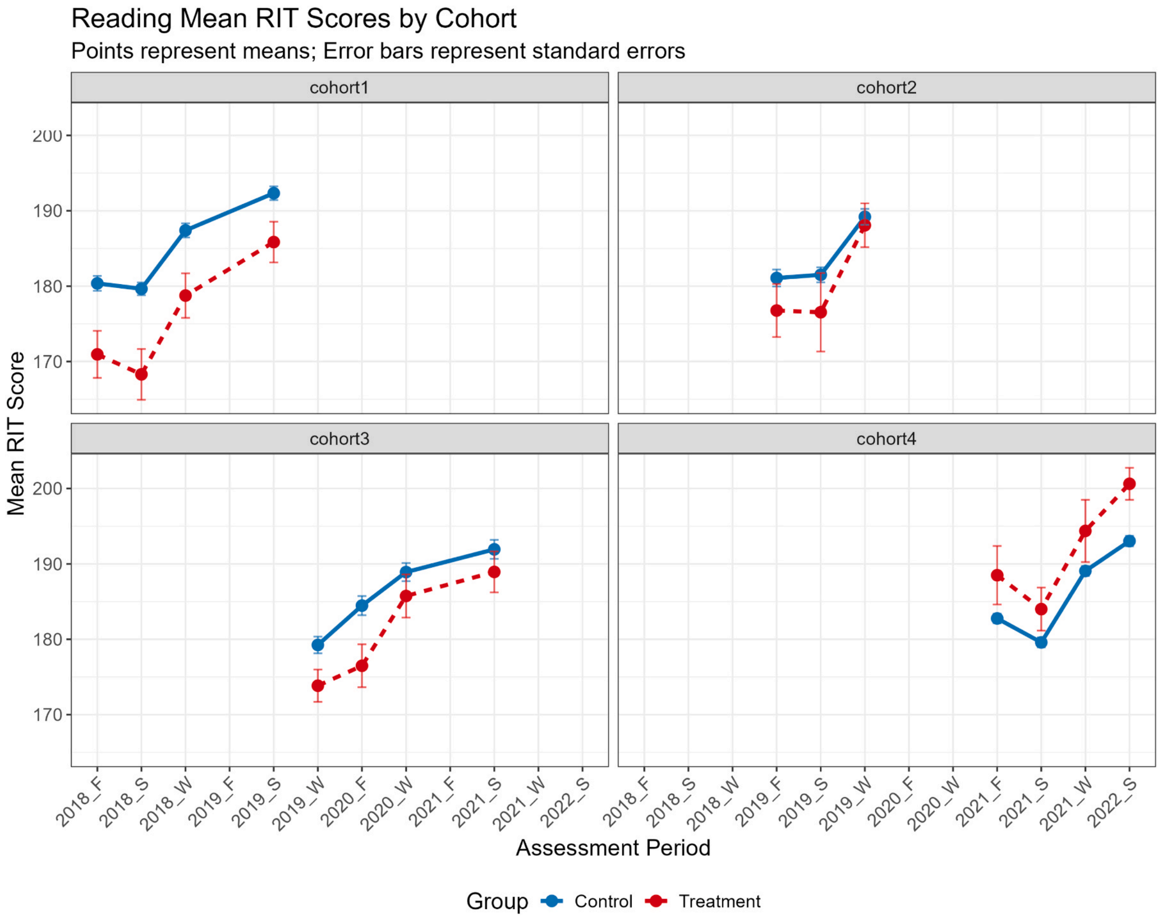


Fig. 1. Parallel trend checks with graphs for reading RIT scores overtime.

complexities of the staggered treatment adoption and potential heterogeneous effects, we further applied the Callaway and Sant'Anna (2021) staggered DiD estimator. This advanced approach enables a more robust verification of conditional parallel trends and provides a dynamic view of treatment effects across the longitudinal panel.

4. Results

Descriptive statistics for combined and each cohort are provided in Supplementary Tables S2–7 (see supplementary materials). MAP RIT score gain for the treatment group above the control group gain ranged from 3 to 6 RIT points across cohorts (average DiD for Reading MAP RIT for Cohort 1 = 6, Cohort 2 = 5, Cohort 3 = 2, Cohort 4 = 6; and average DiD for vocabulary MAP RIT score for Cohort 1 = 4, Cohort 2 = 5, Cohort 3 = 5, Cohort 4 = 3).

Parallel trends were examined for all cohorts using a group-by-time interaction model at two pre-treatment time points. The interaction terms (*Treat* × *Time*) were not statistically significant for any cohort in Reading or Vocabulary (see Tables S8-9), indicating no differential pre-intervention trends.

For the staggered DiD estimator, the unconditional parallel trends assumption was initially not met for Reading ($p = .023$) but was met for Vocabulary ($p = .083$). Consequently, ethnicity, special education status, gender, and age were included as covariates. Furthermore, Cohort 4 was excluded from the staggered DiD analysis due to noise and small treatment size during the pandemic. The final conditional parallel trends tests for the combined Cohorts 1–3 were successful (Reading: $p = .123$; Vocabulary: $p = .377$).

4.1. Reading MAP RIT scores

The DiD analysis of Reading MAP RIT Scores across three traditional analytic methods is summarized in Supplementary Table S10. Meanwhile, the results from the staggered DiD estimator for reading score are presented in Table 2.

The Linear Mixed Effects (LME) model was the only traditional approach to detect a statistically significant Average Treatment Effect (ATE) for Cohort 1 (ATE = 4.92, $SE = 2.07$, $p = .01$, $d = .16$). For Cohorts 2, 3 and 4 no traditional DiD methods yielded significant results ($p > .05$), with estimated ATEs of 4.37, 3.17 and 5.85, respectively.

Although aggregating multiple cohorts via traditional DiD is generally discouraged due to the risk of "forbidden comparisons" (Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021), where early-treated units are improperly used as controls for later-treated units, we present these combined results to provide a baseline for comparison with the more robust staggered DiD estimates. Under the traditional aggregation of Cohorts 1–3, the ATE was 4.41, reaching statistical significance under both the OLS with Cluster-Robust Standard Errors (CRSE) and LME frameworks. The LME model exhibited the highest precision ($SE = 1.35$, $p = .001$, $d = .15$), followed by the CRSE model ($SE = 1.52$, $p = .004$, $d = .13$).

In contrast, the staggered DiD estimator, which explicitly accounts for heterogeneous treatment timing (Table 2), revealed a more robust and significant Simple Average ATT of 5.812 ($SE = 1.276$, 95% CI [3.31, 8.31], $d = .485$). Analysis of group-specific effects confirmed significant positive ATTs across all cohorts, with Cohort 2 exhibiting the largest impact (ATT = 8.575, $p < .05$). Furthermore, the event study analysis formally validated the parallel trends assumption, showing pre-treatment stability (Pre-period 1: ATT = 1.716, $p > .05$) followed by a

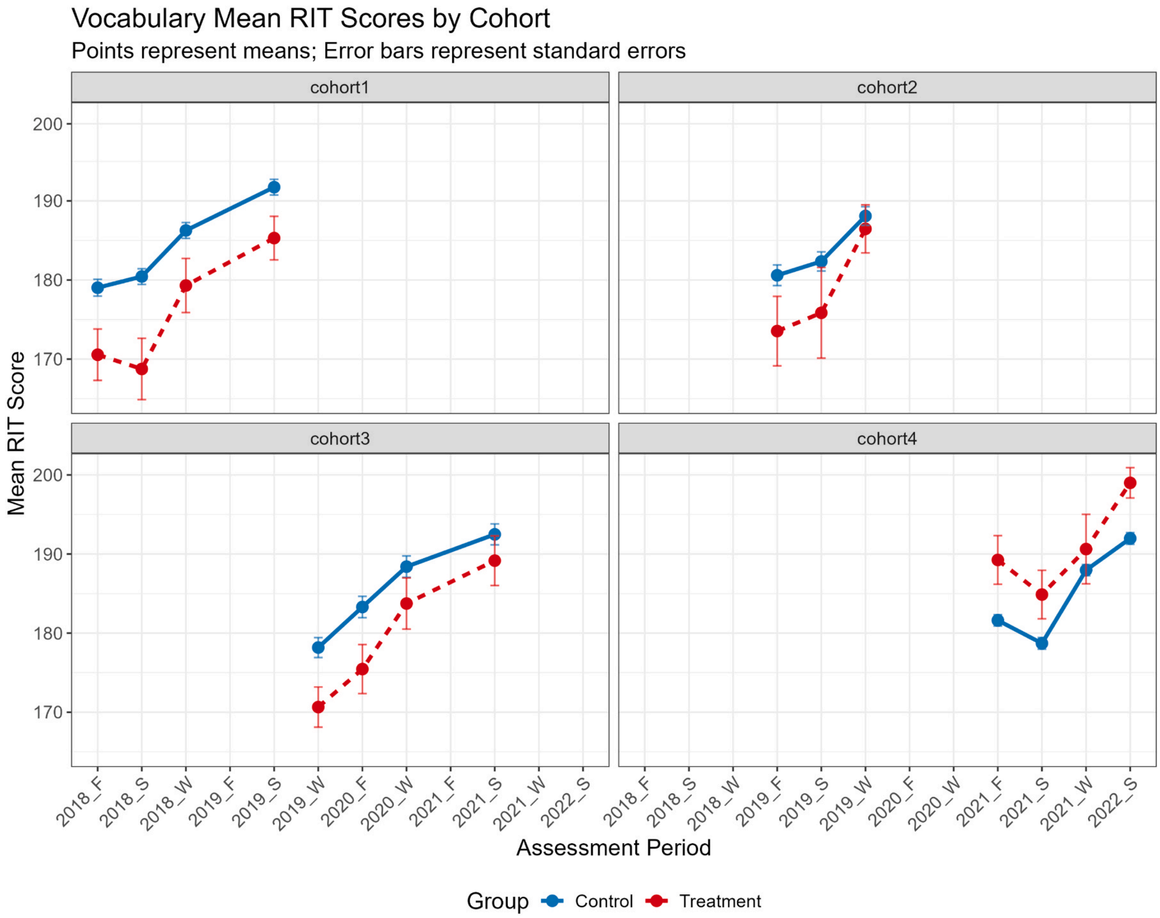


Fig. 2. Parallel trend checks with graphs for vocabulary RIT scores overtime.

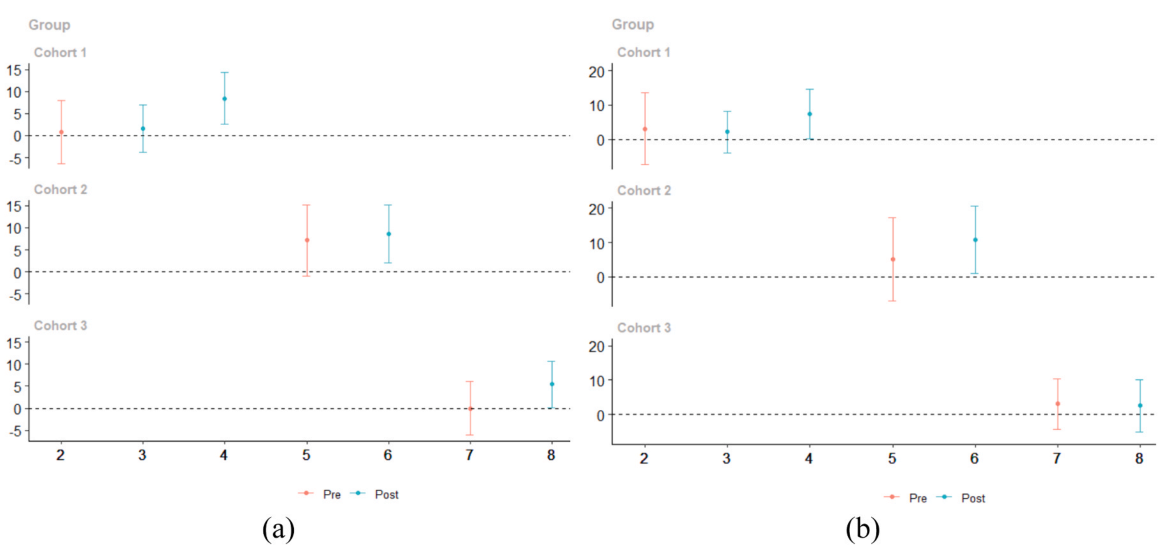


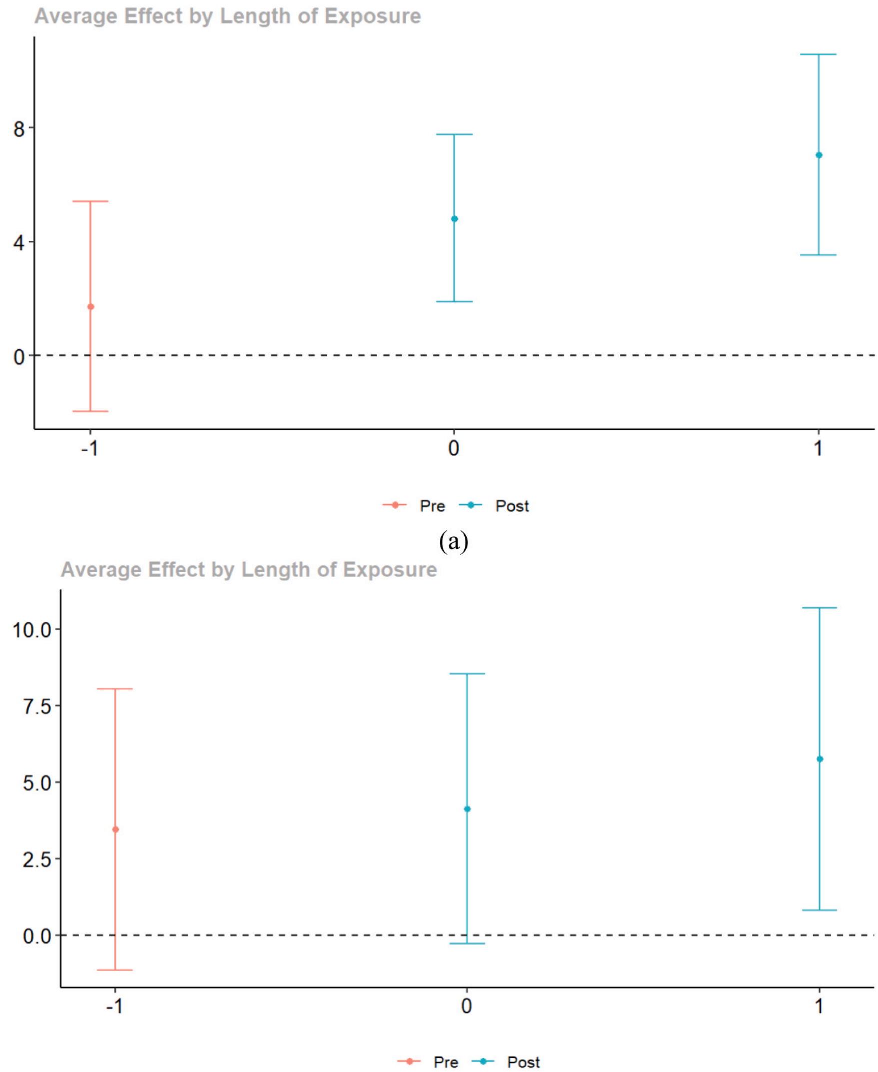
Fig. 3. Dynamic group-time average treatment effects for reading (a) and vocabulary (b).

cumulative impact that reached an ATT of 7.054 during the post-treatment period.

4.2. Vocabulary MAP RIT scores

The DiD analysis of Vocabulary MAP RIT Scores across three traditional analytic methods is summarized in Supplementary Table S11. Meanwhile, the results from the staggered DiD estimator for reading score are presented in Table 3.

For Vocabulary MAP RIT scores, traditional DiD analyses of individual cohorts yielded no statistically significant results across all models ($p > .05$), with ATEs ranging from 3.49 to 5.24. However, similar to the Reading analysis, aggregating Cohorts 1–3 significantly increased statistical power. The combined ATE of 6.26 was statistically



(a)

(b)

Fig. 4. Event study results on reading (a) and vocabulary (b) ATT.

Table 2
Staggered DiD estimates for MAP reading RIT scores.

Analysis Component	Estimate/ATT	SE	95% CI	Cohen's d
Simple Average ATT	5.812	1.276	[3.311, 8.312]*	0.485
Group Effect				
Cohort 1	4.968	1.983	[0.662, 9.275]*	0.415
Cohort 2	8.575	2.509	[3.125, 14.025]*	0.716
Cohort 3	5.775	1.784	[1.901, 9.649]*	0.482
Event Study				
Pre-period 1	1.716	1.623	[-1.980, 5.411]	0.143
Post-period 0	4.825	1.290	[1.887, 7.764]*	0.403
Post-period 1	7.054	1.551	[3.523, 10.586]*	0.589

Note. * Confidence band does not cover 0

Table 3
Staggered DiD estimates for MAP vocabulary RIT scores.

Analysis Component	Estimate/ATT	SE	95% CI	Cohen's d
Simple Average ATT	4.846	1.482	[1.941, 7.751]*	0.345
Group Effect				
Cohort 1	4.694	2.446	[-0.760, 10.148]	0.334
Cohort 2	10.873	4.341	[1.194, 20.551]*	0.775
Cohort 3	3.641	2.418	[-1.750, 9.032]	0.259
Event Study				
Pre-period 1	3.450	1.929	[-1.137, 8.037]	0.246
Post-period 0	4.125	1.851	[-0.276, 8.526]	0.294
Post-period 1	5.753	2.078	[0.811, 10.695]*	0.410

Note. * Confidence band does not cover 0

significant across all frameworks: OLS ($p = .021$, $d = .10$), CRSE ($p = .004$, $d = .12$), and LME ($p = .001$, $d = .15$). Again, the LME model provided the most precise estimates ($SE = 1.82$).

The staggered DiD estimator provided a more nuanced evaluation of these findings (Table 3). The Simple Average ATT was 4.846 and was statistically significant ($SE = 1.482$, 95% CI [1.94, 7.75], $d = .345$), confirming the overall positive and significant effectiveness of the intervention. While group-specific effects for Cohorts 1 and 3 did not reach significance, Cohort 2 demonstrated a substantial and significant impact (ATT = 10.873, $p < .05$).

Crucially, the event study analysis for Vocabulary revealed a 'sleeper effect' or delayed impact. While the immediate effect upon treatment was positive but not quite significant (Post-period 0: ATT = 4.125, $p > .05$), the intervention's impact became robust and significant in the subsequent period (Post-period 1: ATT = 5.753, 95% CI [0.81, 10.70], $d = .410$). This dynamic trajectory, coupled with a non-significant pre-trend (Pre-period 1: ATT = 3.450, $p > .05$), reinforces the causal claim that the small-group instruction led to sustained vocabulary growth that intensified over time.

5. Discussion

While linear regression provides a foundational model for DiD, the nested and longitudinal structure of educational data requires more sophisticated analytical frameworks to minimize the impact of independence assumption violations. As hypothesized, Linear Mixed Effects (LME) and Cluster Robust Standard Error (CRSE) methods were better suited to account for the data's hierarchical structure (Bertrand et al., 2004). Consistent with previous methodological literature, the LME and CRSE models yielded higher precision and lower standard errors compared to the baseline OLS model (Rokicki et al., 2018; Huang & Li, 2021).

Initial diagnostic tests for the staggered DiD estimator on reading scores indicated a violation of the unconditional parallel trends assumption in alignment with Corral and Yang's (2024) concern about violations to the parallel trends assumption in practice. These discrepancies may stem from divergent academic backgrounds between the treatment and comparison groups, leading to distinct baseline growth trajectories (Roth et al., 2023). Furthermore, during data cleaning, we determined that grade retention among certain students also introduced additional longitudinal noise. These findings underscore the necessity for researchers to collect comprehensive demographic and student-history data, enabling the use of conditional parallel trends tests via covariate adjustment to mitigate pre-existing group differences (Callaway & Sant'Anna, 2021; Roth et al., 2023).

Cohort 1, 2, and 3 data exhibited similar overall trends in outcome score changes among pre and post intervention across all traditional DiD analytic methods. While standard error estimates and confidence intervals varied, the superior precision of the higher-level models enabled a more reliable detection of treatment effects.

However, considering the staggered design of the present study, using traditional DiD methods often risk "forbidden comparisons" when aggregating cohorts across different calendar years (Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfoeuille, 2020). By conducting the staggered DiD approach, we obtained a more robust simple average ATT for reading (5.812) and vocabulary (4.846) when controlling for students' ethnicity, special education, gender and age, with effect sizes ranging from $d = 0.35$–0.49. This corrects the initial OLS assessment of "negligible" effects, revealing that the small-group intervention produced practically meaningful growth in student achievement (Callaway & Sant'Anna, 2021; Sun & Abraham, 2021).

Although traditional and staggered DiD methods both indicated a positive intervention effect, the precise estimates differed significantly. Traditional methods produced consistent ATEs but suffered from higher standard errors and lower effect sizes, with only Cohort 1 reaching significance under the LME model. This suggests that traditional DiD may lack the power to detect effects in smaller sub-samples or may misestimate the ATE by failing to account for heterogeneous treatment timing. In contrast, the staggered DiD estimator prevented forbidden comparisons by using only never-treated or yet-to-be-treated units as a comparison group, resulting in lower standard errors and greater statistical power (Callaway & Sant'Anna, 2021; Sun & Abraham, 2021).

Furthermore, the staggered DiD results offer a dynamic perspective on student growth that static models fail to capture. The overall Reading ATT of 5.812 indicates a significant average increase in LEP student achievement after adjusting for students' ethnicity, special education status, gender, and age. A temporal analysis reveals that reading scores increased by 4.825 one semester into the intervention ($e = 0$), escalating to 7.054 by the second semester ($e = 1$). This trajectory suggests that the small-group instruction exerts a cumulative and continuous influence on reading achievement (Callaway & Sant'Anna, 2021).

In contrast, the intervention's impact on Vocabulary exhibited greater heterogeneity and a delayed onset. While the simple average ATT remained significant at 4.846 when controlling for covariates, the effect was predominantly driven by Cohort 2, with no statistically significant influence detected for Cohorts 1 and 3 individually. Moreover, the vocabulary gains were only realized during the final semester of third grade ($e = 1$, ATT $= 5.753$, $p < .05$), indicating a 'sleeper effect' where linguistic gains required longer immersion to manifest. Collectively, these findings suggest that the intervention was more immediately and consistently effective for reading proficiency than for vocabulary development.

Although individual analyses of Cohorts 1, 2, and 3 showed similar overall trends, Cohort 4 varied significantly. This cohort was characterized by a disproportionately large comparison group ($n = 565$) and a minimal treatment group ($n = 8$), leading to high standard errors ($SE = 8.00$). Unlike the earlier cohorts, Cohort 4 treatment students had higher average scores at pretreatment than the comparison group. Contextual factors, including an influx of Venezuelan immigrant students and a delayed intervention start in January (rather than September), resulted in fewer intervention hours. Consequently, Cohort 4 was excluded from the combined analysis to maintain the integrity of the data set.

In response to research question 2, treatment effect estimates were identical across traditional analytical methods, however standard errors were smaller with LME, followed by OLS with CRSE, and lastly OLS alone had the highest SE for all analysis except for Cohort 4. Results indicate that when students' time points (pre and post) are clustered within each student, the unexplained variance is reduced (Huang & Li, 2021; Rokicki et al., 2018; Snijders & Bosker, 1993). Researchers should be cautioned that in an educational research context, observations are likely to be grouped, hence, errors are correlated across individuals within groups; so, models that do not account for this correlation can result in misleading standard errors (SE) and may lead to incorrect inferential statistics (Abadie et al., 2023; Bertrand et al., 2004). Also, the same student score may be measured repeatedly, which will lead to independence assumption violation. The linear mixed effects model does not require either of the above assumptions to be met for it to be valid. Also, unequal sample sizes are not as problematic with mixed linear modeling (Huang & Li, 2021; Snijders & Bosker, 1993). OLS with cluster robust SE is calculating standard error similarly to LME except with the Cohort 4 data analysis.

When conducting DiD analyses in educational research or evaluation studies, researchers must consider both the nested structure of the students and the staggered structure of multi-year data collection to ensure robust results (Corral & Yang, 2024). The staggered DiD estimator is highly recommended for educational evaluation when: (a) ample student outcome data are available, (b) longitudinal datasets are collected during different time points for both treatment and control groups, (c) multiple pre-intervention time points allow for the evaluation of pre-trends, and (d) the analytical approach adequately addresses clustered data (Callaway & Sant'Anna, 2021; Corral & Yang, 2024).

Schools regularly administer standardized tests multiple times during the school year, so obtaining such data is often possible. Moreover, school administrators and teachers may be able to glean valuable information regarding campus intervention initiatives using current or past student data. Norm-referenced scaled scores can be reliably compared from year to year over many years, particularly when those scores are based on underlying latent trait analysis with IRT (item response theory) based metrics such as the RIT scores in the present study. Further, as demonstrated in the present study, while prior parallel trends are important to rule out confounds, researchers must use caution and holistically interpret these trends with DiD within the context of the study.

5.1. Conclusions

Both the linear mixed effects (LME) model and the OLS with cluster robust standard errors (CRSE) produced similar parameter estimates for each cohort data analyzed separately. When cohorts were combined, the increased sample size and statistical power yielded equivalent results across the LME and CRSE frameworks. For the present study, the convergence of these analytical methods strengthens the reliability of the substantive conclusions and causal inferences. Furthermore, the integration of the staggered DiD estimator provided a robust validation of these findings by effectively addressing the risk of heterogeneous treatment timing and potential bias from staggered intervention adoption. As demonstrated, DiD—when coupled with LME, CRSE, or staggered estimators—serves as a powerful and practical tool for deriving causal insights from observational educational data in a straightforward and replicable manner. (Abadie et al., 2023; Bertrand et al., 2004).

Standardized student test scores are available for the beginning and end of each school year. Consequently, intervention's impact on student achievement may be assessed using a DiD framework. This methodology provides nuanced insights to effective strategies, policies, or teachers by connecting student outcome scores to pre-post growth trajectories. The implications for applied research and evaluation are significant: the average impact of teacher-led treatments or district-wide policies can be systematically compared against comparable student groups or historical performance benchmarks, facilitating more data-driven decision-making in school settings.

5.2. Limitations

Despite the robust findings, several limitations warrant consideration. First, the lack of detailed teacher demographic data precluded a quantifiable verification of group comparability regarding years of experience or certification levels. Second, because teacher participation was based on self-selection, the potential for selection bias cannot be entirely ruled out, as participating teachers may possess unobserved characteristics (e.g., higher initial motivation) that influenced student outcomes. Furthermore, the study lacked sufficient covariate data at the teacher level to implement advanced matching or adjustment techniques to further control for pre-existing differences between groups. Additionally, the current study employed a list-wise deletion approach to handle missing data to satisfy the balanced panel requirements of the staggered DiD estimator. Future research could employ more functional imputation methods, such as Multiple Imputation or Full Information Maximum Likelihood (FIML), to further enhance the utilization of the available longitudinal data.

Another significant limitation was the potential confounding impact of the COVID-19 pandemic on the educational landscape, which introduced unforeseen noise and variability into the longitudinal data, particularly for the later cohorts. Finally, while pre-treatment trends were validated using two time points, future research should aim to secure more extensive historical data to establish pre-intervention parallel trends with greater confidence and over a longer duration.

Recommendations for future research

To improve the precision of future longitudinal evaluations, researchers should collect both student demographic data and teacher-level data, including years of experience and specific certification levels. Such data would facilitate the use of advanced matching techniques to better control for the selection bias often present in voluntary teacher participation, while also accounting for baseline variation in student achievement prior to intervention. Furthermore, while list-wise deletion is a common strategy for maintaining panel balance, future studies should consider employing robust imputation methods, such as Multiple Imputation or Full Information Maximum Likelihood (FIML), to preserve statistical power and minimize data loss.

We also recommend that future research extend the observation window to include multiple pre-intervention time points, which would allow for a more rigorous validation of parallel trends over a longer duration. Given the "sleeper effect" identified in this study regarding vocabulary gains, longitudinal tracking should persist well beyond the immediate post-intervention phase to evaluate the long-term sustainability of the treatment effects. Finally, evaluators working with staggered intervention rollouts should try to use more robust estimators, such as the Callaway and Sant'Anna (2021) method, to ensure that causal inferences are not confounded by temporal or cohort heterogeneity.

CRediT authorship contribution statement

Yan Zhang: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Princy Sebastian:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Darrell M. Hull:** Supervision, Project administration.

Funding

This publication was supported by the Office of English Language Acquisition (OELA) of the U.S. Department of Education PR/Award #T365Z170181. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by the U. S. Department of Education.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Gemini (Google) to improve the clarity, coherence, and language of the manuscript, as well as to assist in structuring responses to peer review comments. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of Competing Interest

We have no known conflict of interest to disclose.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.stueduc.2026.101594.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics, 138*(1), 1–35.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press. https://doi.org/10.1515/9781400829828-010

Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics, 144*(2), 370–395.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics, 119*(1), 249–275. https://doi.org/10.1162/003355304772839588

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies, 91*(6), 3253–3285.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Card, D., & Krueger, A. B. (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: reply. *American Economic Review, 90*(5), 1397–1420. https://doi.org/10.1257/aer.90.5.1397

de Chaisemartin, C., & D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review, 110*(9), 2964–2996.

Cook, T. D., Campbell, D. T., & Shadish, W. R. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Corral, D., & Yang, M. (2024). An introduction to the difference-in-differences design in education policy research. *Asia Pacific Education Review, 25*(3), 663–672.

Furquim, F., Corral, D., & Hillman, N. (2019). A primer for interpreting and designing difference-in-differences studies in higher education research, . In *Higher Education: Handbook of Theory and Research: 35*. *In M. B. Paulsen (Ed.* (pp. 1–58). Springer. https://doi.org/10.1007/978-3-030-11743-6_5-1

Gardner, J. (2022). Two-stage differences in differences⟨https://arxiv.org/abs/2207.05943⟩.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*, 709–722. https://doi.org/10.3758/bf03192961

Heck, R. H., Thomas, S., & Tabata, L. (2013). *Multilevel modeling of categorical outcomes using IBM SPSS*. Routledge. https://doi.org/10.4324/9780203808986

Horii, S., Nguyen, C. T. M., Pham, H. T. T., Amaike, N., Ho, H. T., & Aiga, H. (2021). Effectiveness of a standard clinical training program in new graduate nurses' competencies in Vietnam: A quasi-experimental longitudinal study with a difference-in-differences design. *PLOS ONE, 16*(7), Article e0254238. https://doi.org/10.1371/journal.pone.0254238

Huang, F. L., & Li, X. (2021). Using cluster-robust standard errors when analyzing group-randomized trials with few clusters. *Behavior Research Methods*, 1–19. https://doi.org/10.3758/s13428-021-01627-0

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent comparison group design. *Psychological Bulletin, 82*(3), 345–356. https://doi.org/10.1037/0033-2909.82.3.345

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics, 4*(3), 165–224. https://doi.org/10.1561/0800000014

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114–140.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, Article 104092. https://doi.org/10.1016/j.jml.2020.104092

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., & Sekhon, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology, 16*, 1–21. https://doi.org/10.1007/s10742-016-0146-8

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: Basic concepts and examples. *Mixed-effects Models in States and States-Plus*, 3–56. https://doi.org/10.1007/978-1-4419-0318-1

Rietveld, J., Schilling, M. A., & Bellavitis, C. (2019). Platform strategy: Managing ecosystem value through selective promotion of complements. *Organization Science, 30*(6), 1232–1251. https://doi.org/10.1287/orsc.2019.1290

Rokicki, S., Cohen, J., Fink, G., Salomon, J. A., & Landrum, M. B. (2018). Inference with difference-in-differences with a small number of groups: A review, simulation study, and empirical application using SHARE data. *Medical Care, 56*(1), 97–105. https://doi.org/10.1097/mlr.0000000000000830

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics, 235*(2), 2218–2244.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics, 29*(3), 343–367.

Sieweke, J., & Santoni, S. (2020). Natural experiments in leadership research: An introduction, review, and guidelines. *The Leadership Quarterly, 31*(1), Article 101338. https://doi.org/10.1016/j.leaqua.2019.101338

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*(3), 237–259. https://doi.org/10.3102/10769986018003237

Snow, J. (1856). Cholera and the water supply in the South Districts of London in 1854. *Journal of Public Health and Sanitary Review, 2*(7), 239–257. https://doi.org/10.1136/bmj.s4-1.42.864

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2), 175–199.

Wing, C., Yozwiak, M., Hollingsworth, A., Freedman, S., & Simon, K. (2024). Designing difference-in-difference studies with staggered treatment adoption: key concepts and practical guidelines. *Annual Review of Public Health, 45*.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

What Works Clearinghouse. (2022). W*hat works clearinghouse Procedures and standards handbook* (version 5.0). U.S. Department of Education, Institute of Education Sciences. National Center for Education Evaluation and Regional Assistance (NCEE). https://ies.ed.gov/ncee/wwc/handbooks.