Math Curriculum Evaluation

Technical Appendix

Duncan Chaplin

April 21, 2026

Executive Summary

River Forest school district in Illinois implemented two new math curricula in the last decade. The results of this evaluation suggest positive effects of both–around 1.7 points on the Illinois Assessment of Readiness (IAR) math test for the Savas Investigations curriculum and around 4.0 points for the Desmos curriculum. Bayesian analyses suggest that there is a 65 percent chance of positive impacts for Investigations and a 78 percent chance of positive impacts for Demos. This is in spite of the fact that there appear to have been short term reductions in performance for Investigations in the years immediately after it was introduced. These estimates were based on an analysis of over a decade of data by district, year, and grade covering River Forest and a set of matched comparison districts. While the data suggest positive effects, there was a substantial chance that the opposite was true and since this was not a randomized control trial, bias is possible.

Acknowledgements: I would like to thank several people for their valuable contributions to this evaluation. Eric Isenberg, secretary of the school board for River Forest Public Schools, had the original vision for this work and helped with the design of the final product. Ed Condon, the current superintendent of River Forest, oversaw the overall process and provided valuable input. Christine Trendel, Director of Curriculum and Instruction at River Forest, provided valuable data and insights into what other changes might have affected student outcomes. Maureen Font, at the Illinois State Department of Education, assisted with their FOIA process. Josh Stewart, at Rocky Mountain Research, provided a quality assurance review of the code, PowerPoint presentation, and Technical Appendix. This work was implemented using vibe coding and some assistance with writing from ChatGPT.

Table of Contents

Overview

This appendix provides details regarding the methods used to produce the results of this evaluation. The results were presented in the PowerPoint Presentation, “River Forest Public Schools, Math Curricula Evaluation,” also dated April 21, 2026 (Chaplin 2026).

This evaluation estimates effects of two math curricula as they were implemented in the River Forest school district in Illinois compared to whatever was being implemented in other school districts at the same time. During the 2017-2018 school year River Forest implemented the Investigations curriculum in grades K through 5. Concern has been expressed that Investigations may have lowered student performance because grade 3 to 5 test scores fell in 2018 and 2019 compared to the two previous years. In the spring of 2024 River Forest briefly implemented the Desmos curriculum in grades 6-8. It was fully implemented in those grades during the 2024-2025 school year, and in grade 5 in 2025-2026. This evaluation estimates the effects of Desmos on math performance in the spring of 2024 and 2025.

Research questions

This evaluation is designed to answer two key research questions.

- What is the effect of Investigations on math test scores in grades 3-5 in 2017-18 and on students who might have benefited in later years in the River Forest School District?
- What is the effect of Desmos on math test scores in grades 6-8 in the spring of 2024 and 2025 in the River Forest School District?

It was also designed to answer one exploratory question.

- Do these results vary by grade and year?

For each evaluation question, effects were estimated combining all grades and post-treatment time points. If those results had been statistically significant then effect estimates would have been presented by year and, separately, by grade and year in a table (with measures of precision) and graphically. Since the main results were not statistically significant, only a subset of the possible breakdowns by year and grade are presented. This was done to maintain alignment with the research plan approved by Isenberg (see section on Model Specification below).

Data

This evaluation uses data by district and grade from 2015 to 2025 for grades 3 through 8.[1] The outcomes are predicted math scale scores. They are predicted based on the fractions of students scoring in each performance category on the IAR tests from 2019 to 2025 and the Partnership for Assessment of Readiness for College and Careers (PARCC) tests from 2015 to 2018. Predicted scale scores are the outcome instead of actual scale scores because the latter were not available prior to 2022. Predicted scale scores were used instead of performance by category to keep the analysis more parsimonious. Data from 2020 are left out because testing was not done in that year. Data from 2021 are excluded from the main analyses because participation rates were still relatively low in that year–around 70 percent compared to over 95 percent in other years.

Most of the data used in these analyses was obtained from the website of the Illinois State Board of Education (ISBE). This includes student test scores by category, student characteristics, and student enrollment. Scale scores were not available there. Those were obtained by submitting a Freedom of Information Act (FOIA) to ISBE for student-level data from 2022 to 2025 with identifiers for district and grade. This request covered the districts used in this evaluation in grades 3 to 8. Results for district/grade combinations covering fewer than 10 students were redacted following state policy regarding confidentiality.

In 2025 the website data did not cover all 5 categories of student performance so those were calculated using the scoring cut-points from previous years combined with the student-level data. The validity of that method was verified by calculating the fractions of students scoring in each category based on the student-level data and the cut-points from 2022 to 2024. The correlation between the fractions of students scoring in each performance category based on the website data and the student-level data was almost perfect.

The covariates used in the regressions include standard student characteristics, student enrollment, grade, year, district, and the prior year's value of English Language Arts (ELA) scores. Students who are white or missing race are the omitted category for race/ethnicity. The variables used for matching were the district-level means and standard deviations in test scores by grade and year from 2015 to 2017 (before these new curricula were implemented).

[1] Throughout if a single year is used to describe a school year it refers to the spring of the year since that is when tests are administered.

Several other variables were considered for matching or regression adjustment. Gender was dropped because it was non-trivial to calculate from the website data and perhaps for this reason it was not included in previous work done for this district (Earvolino, 2024). Student characteristics by grade were dropped because they were also non-trivial to calculate. The variable identifying the percentages of students with race unknown was dropped because it was not statistically significant in the previous report for River Forest (Earvolino, 2024) and because the definitions used by districts may change over time. Variables from the American Community Survey, which had been used by Earvolino, 2024, were dropped because that data only captures 5-year averages so would likely not give us much information on variation by year. In addition, much of their variation would likely have been captured by the district fixed effects and the fraction of students in families with low income. Variables that were potentially endogenous were excluded as they could bias the results. These included variables describing staffing, the numbers of schools by type, class size, and school funding.

Three additional types of outcomes were considered. IAR student growth percentiles were dropped because they are not available before 2018 and because they might be somewhat endogenous. MAP scale scores were dropped because they were not easily available for comparison districts. ELA scores were dropped as an outcome because the curricula focused on Math rather than ELA.

Two other datasets were considered: SEDA and Ed Facts. SEDA provides scale score estimates by administrative district and grade from 2008-9 to 2018-19 (as of 11/13/2025). To justify using those data one would need to replicate their method for later years. Another option was to use Ed Facts. Those data continue up to 2023-24 (as of 11/13/2025) but only provide proficiency rates.

Evaluation Design

This model compares how outcomes changed in River Forest relative to similar districts, before and after implementation, while controlling for other factors. These curricula were implemented in a subset of grades in some years enabling use of a triple difference estimator that exploits variation by grade as well as district and year. The evaluation also uses matching, regression adjustment, and estimates of uncertainty that avoid the need for strong assumptions that are normally used, as discussed below in the section on uncertainty.

To ensure that the results did not influence the evaluation design, all decisions were approved by Eric Isenberg (on the school board of River Forest), before effects of the

new curricula in River Forest were estimated. Estimates of the standard deviations of the placebo effects described below were used to help inform final decisions.

Model Specification

Effects were estimated using regression analyses limited to a carefully selected set of matched comparison districts. Comparison districts were limited to those from suburban Chicago regular elementary districts in Cook County and the five collar counties (Lake, Will, DuPage, Kane, McHenry). There were 205 potential comparison districts in the state data at least once between 2019 and 2025. 171 of those had data for all six years (excluding 2020). Potential comparison districts are further limited to the 136 that were also present in the PARCC data from 2015 to 2018 and had complete test score data from 2015 to 2025 (again excluding 2020). The final set of comparison districts were selected from these 136 districts by matching based on the mean and standard deviation of average math scores for each district, by year and grade from 2015 to 2017 (before either of the new curriculum were implemented). The mean is used to capture overall academic performance. The standard deviation is used to capture how stable district performance is. Matching was done using Mahalanobis distance based on the matching characteristics.[2] This method was found to work well in similar situations (Van Dine et al 2021). The 108 districts with the smallest distance values were selected. The number of comparison districts was kept fairly large to ensure reasonably precise estimates of the probabilities that the effects were positive (or negative).

Effects were estimated using a standard regression model.

$Y = T'B1 + X'B2 + e$ where

Y = the outcome
T = a vector of treatment variables,
X = a vector of control variables, and
e is an error term.

The outcome (Y) is predicted math scale scores, as described below.

[2] A common alternative to matching in situations like this is to use weights as is done by Abadie et al (2010, 2015).

The vector of treatment variables (T) consists of one variable for Investigations and one for Desmos. The Desmos variable is an indicator for if the grade is 6 through 8 and the year is 2024 or 2025 where years are identified by the spring of the year. Those are the grades and years when Desmos was implemented. The Investigations treatment variable is either a dummy or a dosage variable. Dosage is described in Table 1. In most models the treatment variable for Investigations is a dummy equal to one if dosage is positive and 0 otherwise. While the Investigations curriculum is only implemented in grades 3 through 5, the main model also incorporates potential effects on grades 6 through 8. Exploratory work evaluates variation in effects by grade.

Table 1

Year\Grade	3	4	5	6	7	8
2015	0	0	0	0	0	0
2016	0	0	0	0	0	0
2017	0	0	0	0	0	0
2018	1	1	1	1	1	1
2019	1	2	2	2	2	2
2020-2025	1	2	3	3	3	3

The control variables (X) are the percentages of each district's student body who were Black, Hispanic or American Indian, Asian or Pacific Islander, Multiracial, English Language Learner (ELL), Individualized Education Plan (IEP), or Low Income as well as the log of district enrollment, that variable squared, grade dummies, year dummies, district dummies, and the prior year's value of ELA scores. In some models the prior year's value of ELA scores is dropped. The categories Hispanic and American Indian were combined in part because their outcomes were similar and in part because of concern that the definition of American Indian might have changed over time. Asian and Pacific Islander were combined for similar reasons.

Uncertainty

In addition to estimating the effects of the new curricula in River Forest, I also estimate placebo effects for the comparison districts to help gauge how unusual the River Forest results are. If River Forest's estimate was close to the center of the distribution of placebo estimates, that would suggest it was more likely due to chance rather than a true effect. Most evaluation methods use standard errors to describe uncertainty in the results. That cannot be done in this evaluation because the effects being estimated are for new curricula in a single district (River Forest). This means that the sample size of

the treatment group is constrained, and the standard assumptions about how estimates converge as the sample size grows do not apply. For this reason, an alternative method of describing uncertainty based on placebo effect estimates is used. More precisely, the regression model is estimated once for each comparison district, each time pretending the untreated comparison district received the new curriculum in the same years and grades as River Forest, which was excluded from these regressions. The standard deviation of those placebo effects is used to describe the variation in estimated effects one might expect to see under the null of no treatment. This standard deviation is similar to the standard error used in other evaluations. A similar concept is discussed by Bertrand, Duflo, & Mullainathan (2004), Abadie, Diamond, & Hainmueller (2010, 2015) and Conley & Taber (2011). Bayesian shrinkage is then used on the effect estimate. The standard deviations of the placebo effects for Investigations and Desmos are 4.35 and 5.22 respectively for research question 1. In comparison the standard deviations of student-level math scores range from 29.7 to 43.0 with a mean of 34.9 so the method appears to be precise enough to estimate effects of about 0.35 or 0.42 standard deviations for Investigations and Desmos respectively. This is similar to what was found by Dotter et al (2021) when estimating effects of school reforms in DC compared to other large urban areas using a synthetic cohort model. In that paper the minimum effect sizes were around 0.31 and 0.44 for grade 8 and grade 8 math. That paper also had about 10 years of data and a comparison sample of about 60 urban areas. The years of data were not consecutive so spread out over a longer time frame (from 1992 to 2017). Bayesian methods were used to interpret the results using methods described by Deke et al (2022) with priors for the standard deviation of effects obtained from What Works Clearinghouse data using data on elementary school math. The prior for the mean was set to 0.

Robustness tests

Across all robustness checks, the results were qualitatively similar: with small, positive, and imprecisely estimated effects except when estimating effects of Investigations without grades 6-8, as discussed below.

Dropping cases with missing race. Some observations appear to be missing data on race/ethnicity for some students. When data are limited to those observations with very little missing data, the results are similar to the main results as shown in Table 2 below. This finding is discussed further in the section on missing data below.

Dropping lagged ELA scores. The main estimation models control for lagged ELA scores. A potential benefit of controlling for these scores is that it could improve the precision of the resulting estimates. A downside is that it could introduce bias if ELA

scores were affected by the new curricula. The posterior standard deviations without ELA scores are 15 percent and 7 percent larger than those with lagged ELA scores for Investigations and Desmos respectively. Hausman-Wu tests suggest no clear differences in the results, so the main model includes that control variable.

Effects of Investigations without grades 6-8. The main model allows Investigations to improve test scores in grades 6-8 when the students who received the Investigations curriculum were in those grades. When the analyses are limited to only grades 3-5 the estimated effects of Investigations drop to be slightly negative (-0.71).[3] However, the potential for benefits in grades 6-8 is important. That is why they were included in the main model. The main model also controls for the effects of Desmos when estimating effects of Investigations, so the benefits of Investigations are those for the cohorts of students in grades 6-8 before Desmos was implemented.

Effects of Dosage for Investigations. Students may benefit more from multiple years of Investigations compared to having only 1 year. The main model includes a binary dummy identifying if there was any dosage. When that is replaced by a continuous dosage variable, the results were similar to those for the binary variable–in both cases the point estimate was positive and between 50 and 100 percent of the size of their posterior standard deviations.

Effects by Year. The Desmos intervention was only partly implemented in the 2023-2024 school year. For this reason, effects were estimated by year. The results were almost identical across years, so this does not appear to be an issue. A similar pattern holds for Investigations in that the results are similar when we exclude the 2024 or 2025 years.

Desmos and Investigations Separately. The main model estimates effects of Desmos and Investigations jointly. When they are estimated separately the results are similar to when they are estimated jointly, but the effect estimate for Investigations is somewhat larger though still smaller than that of Desmos.

[3] Desmos was dropped from this model because it only affects performance in grades 6-8.

Table 2. Robustness Tests

Robustness test	Effect estimate	Standard Deviation	Probability Effect > 0
Main Model: Investigations	1.69	4.35	0.65
Main Model: Desmos	4.04	5.22	0.78
Dropping cases with missing Race: Investigations	1.83	4.37	0.66
Dropping cases with missing Race: Desmos	3.91	5.26	0.77
Dropping lagged ELA scores: Investigations	1.80	4.99	0.64
Dropping lagged ELA scores: Desmos	2.80	5.58	0.69
Investigations without grades 6-8	-0.71	4.96	0.44
Dosage for Investigations	1.18	1.79	NA
Desmos by Year, 2024	3.96	5.77	0.75
Desmos by Year, 2025	4.32	5.61	0.78
Investigations by Year, exclude 2025	1.43	4.31	0.63
Investigations by Year, exclude 2024	1.31	4.18	0.62
Desmos and Investigations Separately, Investigations	3.04	4.12	0.75
Desmos and Investigations Separately, Desmos	3.83	5.23	0.77

Note: Effect estimates, standard deviations, and probabilities are based on the posterior distribution estimates from the Bayesian model except for dosage. In that case the point estimate is from the regression model and the standard deviation is of the placebo effects.

Graphical presentation of results: The PowerPoint presentation (Chaplin 2026) presents graphs showing performance overall (combining grades) by year for River Forest and all other districts. If that graph had suggested a very different story than the one found using the estimated effects then additional exploratory work would have been done to determine which covariates caused that difference. However, that was not the case. The graphs suggested a small but imprecisely estimated positive effect for each curriculum which is similar to the estimated effects.

Concurrent changes in River Forest school district

River Forest experienced several changes that could have affected the effect estimates presented in this evaluation. Based on their timing and content, as well as robustness checks, it appears unlikely that these changes caused substantial bias.

The current principal (Tina Steketee) of the middle school (Lincoln) started in the 2024-2025 school year which is the first year that the Desmos curriculum was fully implemented. She replaced Larry Gartski who had been the principal for 17 years. Since the Desmos intervention was for grades 6 to 8 and Lincoln is the only school that

serves those grades, there is strong overlap between this principal and that curriculum. However, Desmos was also at least partially implemented in the previous school year (2023-2024) and, as shown in the robustness checks, the effects of Desmos appear to be similar during those two years. This makes it less likely that the estimated effect of Desmos is driven by the leadership change.

Other staffing changes seem even less likely to matter. The superintendent did not change from Spring 2012 to Spring 2025. This spans the entire timeline for this evaluation so there should be no bias in these results due to changes in superintendents. Thomas Hagerman was the previous superintendent (from 2008 to 2011) which is prior to the start of the data used in these analyses. Ed Conden has been the superintendent since then. He will be stepping down in April of 2026 which is after the end of the data series used here.

Willard Elementary had one principal (Diane Wood) from 2014-2015 to 2023-2024. The current principal, Christine Gerges, started in 2024-2025 after working in this school district for 18 years prior to becoming the principal. Hence, in the robustness section we present results with and without the spring 2025 results. The results were similar, as discussed in that section.

Lincoln Elementary had one principal change during the pre-period, but none since the spring of 2016. Pamela Hyde was the principal for 12 years ending in the 2015-2016 school year. Casey Godfrey, the current principal, started that year. This should not matter much since it was during the pre-period, and the only influence of the 2014-2015 data is through the lagged ELA scores included as regressors in our model.

In addition to staffing changes there have been two significant curricula changes since 2015. All-day kindergarten was started during the 2023–2024 school year. This was too late to matter for this evaluation since those students would only be too young to appear in the data used here which only covers grades 3 through 8 and ends in the 2024-2025 school year, when those students were only in grade 1. The district also implemented a new ELA curriculum in the spring of 2025 for all teachers in grades 6 through 8 and about two thirds of those in grades K-5. Assuming no cross-over across subjects this should not matter either.

This evaluation covers the introduction of Desmos in grades 6-8 in 2023-2024. Desmos was expanded to cover grade 5 starting in the 2025-2026 school year–which is after the end of the data used in this evaluation.

Overall, none of the concurrent changes described above appears sufficient to explain the observed patterns. That said, there could be unobserved concurrent changes (for example in staff effort) that could cause bias if present.

Predicting Scale Scores

Because scale scores were unavailable before 2022, they are predicted using performance categories, which explain nearly all variation in observed scores. A key challenge for this evaluation was that scale score data were not available for the period before the spring of 2022. To address this issue data from 2022 to 2025 were used to explore the relationships between the fractions of students scoring at each level, which are publicly available in all years, and the scale scores. These fractions explained almost all the variation in scale scores. This was especially true for the larger districts without fractions equal to 0 or 1. For those districts the r-squared statistics by grade and subject were all at 99 percent or above. Even when the sample was enlarged to include all districts, the r-squared statistics remained at 90 percent or above. This justified using the fractions to predict scale scores in the earlier years. Predicted values of scale scores were used in all years for consistency and because using scale scores only in the post-period could have introduced bias, as discussed below.

A related challenge is that Illinois switched from PARCC to IAR in 2019. The data analyzed suggest no large shift in the fractions scoring in each category in that year for either River Forest or the state, so the PARCC fractions were used to estimate what the IAR scale scores would have been had the IAR tests been implemented in those earlier years. More precisely, the percentage scoring by performance level is very similar for PARCC in 2018 compared to IAR in 2019. This is true both for the state, and for district 90 in Math and ELA. The largest difference was 2.9 percentage points, and the average absolute difference was under 1 percentage point. See Table 3.

Table 3: Differences in Percentages of Students Scoring in Each Performance Category–PARCC in 2018 vs IAR in 2019

Entity	Subject	Year	% not met	% partial	% approach	% met	% Exceeded
District 90	ELA	2019	3.2	8.1	19.5	47.2	22
District 90	ELA	2018	2.9	6.9	19.7	45.5	24.9
Illinois	ELA	2019	16.4	19.4	26.3	31.5	6.3
Illinois	ELA	2018	16.1	20.2	26.8	31	5.9
District 90	Math	2019	3.9	11.2	21.6	53.3	10.1
District 90	Math	2018	3.1	9.5	24.3	50.9	12.1
Illinois	Math	2019	16.3	25.3	26.6	27.2	4.6
Illinois	Math	2018	16.3	25.1	27.2	26.7	4.6
District 90	ELA	Difference	0.3	1.2	0.2	1.7	2.9
Illinois	ELA	Difference	0.3	0.8	0.5	0.5	0.4
District 90	Math	Difference	0.8	1.7	2.7	2.4	2
Illinois	Math	Difference	0	0.2	0.6	0.5	0

District 90 is River Forest in Illinois.
Average difference is 1 percentage point; maximum is 2.9 percentage points.

Other authors have used predictions when data are missing on scale scores. For example, Potamites et al (2009) estimated the standard deviation of scale scores using a similar method (See footnote 12 in that report), though they do not address non-linearities in the functional form that are addressed below. Reardon et al (2017) comes closer to doing what is done here in that they estimate mean scale scores using the fractions of students who score in each category. However, they rely on relatively strong functional form assumptions. In contrast, the method used here may allow for a wider variety of functional forms. On a related note, this evaluation uses parameters based on the empirical distribution of mean scale scores when available while Reardon et al (2017) only use mean scale scores to validate their method.

The prediction model uses the probabilities of scoring above each of the 4 cut-points used to divide students into 5 performance categories. One could regress scale scores on those probabilities. This would be analogous to what was done by Potamites et al (2009). However, the relationships between those probabilities and scale scores are unlikely to be linear. For this reason, the probabilities are transformed using the inverse of a normal cumulative density function. The

resulting numbers can be thought of as z-scores. They aren't on the same scales as each other or the scale scores but, if the scale scores are normally distributed then the relationships between each of the z-scores and mean scale scores would be linear. Since there are four z-scores (one for each probability), the model has more information than would be needed in theory to predict scale scores. The mean scale scores are regressed on the four z-scores. A separate regression is run for each grade and subject, combining the data from 2022 to 2025 (the only years where scale scores are available). Dummies for years are included in these regressions, but their coefficient estimates are generally not statistically significant and are omitted when calculating predicted scores. The coefficient estimates on the z-scores are used to predict scale scores in all years.

It is not possible to take the inverse of the normal cumulative density function when the probability is equal to 0 or 1. For this reason, a small number (0.000001) was added or subtracted from the probability to keep it away from that bound. This creates a type of measurement error in the z-scores since the true value of that small number is unknown. Smaller districts are also likely to have more measurement error in those variables than larger districts. Measurement error can reduce the magnitudes of the coefficients on those variables. For this reason, the main regressions for predicting test scores are run using only districts with above median enrollment and no probabilities equal to 0 or 1. The z-scores all have positive coefficient estimates with very large t-statistics. The resulting R-squared statistics are all at 99 percent or above. As noted above, the R-squared statistics fall to be closer to 90 percent when the full sample is used. As might be expected, given measurement error, the coefficient estimates decrease by around 10 percent in the full sample compared to the one used for predictions.

There are three other ways of validating this method. First, the resulting scale scores across all observations (including those from smaller districts and with 0/1 probabilities) are highly correlated (from 0.948 to 0.970) with the actual scale scores by grade and subject. Second, when these prediction regressions are run separately by year and then each year's coefficient estimates are used to predict scores in each year, the resulting predicted scores in a given year are highly correlated with each other regardless of which year was used to predict them. Indeed, all those correlations (by grade and subject) are at 0.99 or above. Finally, these correlations do not appear to vary systematically depending on how many years apart the prediction equations are. In other words, predictions based on a regression ran using 2022 data are just as highly correlated with predictions based on a regression ran using 2025 data as are predictions based on regressions

using data fewer years apart. This suggests that using regressions based on data from 2022 to 2025 to predict scores in earlier years may work well.

One could imagine using observed mean scale scores when available instead of the predicted scores. Given the high R-squared statistics the results would likely have been similar. However, this method was not used because of a concern that the predicted values in River Forest might differ systematically from the actual scale scores for River Forest. This could happen if the mean values of scale scores for River Forest within a category differed from the typical mean for other districts. This seems most likely to be a probably in the bottom and top categories. The lowest category is small for River Forest suggesting that differences there are less likely to matter. Hence, the upper tail seems more likely to be potentially problematic. Suppose, for example, that River Forest had lower test scores among students in the top category than other districts with a similar percent of students in that category. If this is the case and one used predicted values pre-treatment but actual values post-treatment, the estimated effects would be biased downwards.

Since the predicted scale scores are based only on the fractions of students scoring in each category, they do not capture changes in the distribution of performance within categories. Capturing such changes is not possible due to a lack of pre-treatment data on actual scale scores. However, the results do capture increases in the percentages of students who made it into any category and can be thought of as a way of summarizing results across categories.

Missing Data

The test score data were very complete in the most recent years (2022–2025), as well as in 2017 and 2018. There are no missing data in 2025 and only a single district with missing data in each of the other years in this group. In contrast, missingness is more substantial in 2015, 2016, 2020, and 2021. In 2015 and 2016, between 52 and 54 districts have missing data, depending on the grade. No data were reported in 2020. In 2021, between 2 and 18 districts have missing test score data, again depending on the grade. Districts with missing test score data in any year are excluded from all analyses.

The student characteristics data (Black, White, Asian, Hispanic, American Indian, Pacific Islander, Multiracial, ELL, IEP, and Low Income) are also generally complete once districts with missing test scores are dropped. There are no missing values for these variables from 2015 to 2017 in this set of districts. In 2018, a small amount of data is missing for these variables and is imputed using 2017 values. In later years,

missingness increases, but this appears to be driven primarily by the suppression of small subgroup counts rather than true data gaps. Accordingly, these missing values are imputed as zero. This was done in place of using lagged values since in most cases if a given variable was missing it was missing for many years in a row. Also, standard imputation methods would not have worked well because the data available for the years with substantial missing data would be highly selected (missing all the observations close to 0). We could impute using data from other years but that would not account for trends over time.

If this method did not work well then we would expect the sum of the race/ethnicity variables to be well below 100 percent in many cases. This did not happen. Following these imputations, the sum of race/ethnicity shares (Black, White, Asian/Pacific Islander, Hispanic/American Indian, and Multiracial) exceeds 90 percent for more than 99 percent of observations and exceeds 97 percent for 90 percent of observations. Approximately 3 percent of districts fall below 90 percent in at least one year. Excluding these districts has minimal effect on the results as shown in the robustness section above.

Two outlier values are also adjusted. One district, North Palos SD 117, reports Multiracial enrollment of 28.7 percent in 2025, and another, Cicero SD 99, reports American Indian enrollment of 24.7 percent in 2017. All other district–year observations, including for these districts, are at or below 16 percent for these categories. These outliers are therefore replaced with the corresponding values from the prior year.

In general, the trends look as expected. For example, academic performance drops from 2019 to 2021, when COVID hit, and then climbs back up often reaching the pre-2019 levels by 2025. However, there were some odd patterns in the data. The American Indian and Pacific Islander populations in River Forest dropped by more than half from 2019 to 2021 (from 0.206 to 0.0947 percent for AI and from 0.08 to 0.008 percent for PI). One possible explanation- enrollment drops from 2,310 to 2,158 during that period and remains at the lower level after. Hence, it is possible that the American Indian and Pacific Islander populations were disproportionately likely to leave these public schools and their districts. Since these are very small subgroups, American Indians were combined with Hispanics and Pacific Islanders with Asians.

The original plan was to exclude districts if the sum of the scores by category across all five categories was less than 95 percent in any grade, year, or subject. That sum was only 94.5 percent in grade 8 math in 2018 for River Forest so the cut-off for excluding districts was lowered to 94 percent. All the other grade/year combinations for River Forest have values of 99 percent or above, as do 99 percent of the grade/year observations for other districts.

Twelve grade-year observations were between 1.01 and 1.05 in Math as is one observation for ELA. These are likely errors. However, they do not affect the calculation of predicted scores which are based only on the percentages of students scoring above each cut point. Those values are never greater than 1.01 in these data for Math or ELA. Values between 1 and 1.01 are likely due to rounding errors. They are set to a value slightly below 1 in the calculations when taking the inverse of the cumulative normal density function of these probabilities.

In the remaining cases where the sum of the categories was below one, each of the probabilities of scoring above each cut point were adjusted to account for the possibility that the scores of some students were omitted from the numerators of these categories even though those students were counted in the denominators. This was done by dividing the probabilities by the sum of the probabilities across all 5 categories. This can be thought of as treating the students whose scores were missing as if they had scores similar to those that were reported.

Limitations

These estimates are not causal in the strict experimental sense; they reflect associations after adjusting for observable differences. In other words, the estimated effects may be biased by factors that were associated with the new curricula but not controlled for in the current model. For example, it is possible that staff changed their levels of effort at the same time the new curricula were introduced. This evaluation cannot control for those types of changes. However, if one interprets the results as capturing the joint effects of additional effort and the new curricula then this is not a problem.

In addition to being causal, these estimates are not designed to estimate the effects of Investigations and Demos outside of River Forest. Indeed, it is possible that Investigations and Demos were being implemented in other school districts at roughly the same time as River Forest. If this is the case then the evaluation is still relevant for how well River Forest is implementing these curricula.

Main Model Full Regression Results

The coefficients on the control variables in the main model are presented below in Table 4. This table uses standard errors that assume independently distributed error terms. These are not used when developing the Bayesian results that are used in this evaluation. Interestingly, the regular standard errors are also large relative to the coefficient estimates. In this way the results are like those produced by the placebo

effect estimates and Bayesian shrinkage. Perhaps not surprisingly, the standard errors are smaller than the standard deviations of the placebo effect estimates and the posterior standard deviations. Most of the coefficient estimates are not statistically significant.

Table 4: Regression Output for Main Regression

Variable	Coefficient Estimate	Standard Error	t-value	Probability
Investigations Dummy	2.008	2.414	0.832	0.405
Desmos Dummy	5.209	3.536	1.473	0.141
Black	**0.263**	**0.086**	**3.078**	**0.002** **
Hispanic or American Indian	**-0.136**	**0.040**	**-3.413**	**0.001** ***
Asian or Pacific Islander	0.120	0.073	1.643	0.100
Multiracial	0.142	0.086	1.658	0.097 .
ELL	**-0.174**	**0.044**	**-3.967**	**<0.001** ***
IEP	0.066	0.070	0.949	0.343
Low Income	-0.005	0.016	-0.296	0.767
Log enrollment	-8.636	13.471	-0.641	0.521
Log enrollment^2	0.115	0.857	0.135	0.893
Lag Predicted ELA score	**0.262**	**0.011**	**23.362**	**<0.001** ***

Outcome: Predicted Math score
Observations: 5,232 (excludes 2015, 2020, and 2021)
Weights: Enrollment
Fixed Effects: Grade (6), Year (8), District (109)
RMSE: 5.79 Adjusted R^2: 0.864 Within R^2: 0.113
Bold indicates statistically significant at the 0.05 level or below.
'***' indicates statistically significant at the 0.001 level,
'**' at the 0.01 level, '*' at the 0.05 level, and '.' 0.1 ' at the 0.10 level.
Coefficient estimates and standard errors are without Bayesian shrinkage.
Probability is the probability of observing a coefficient estimate this large if the true parameter were 0.

Descriptive Statistics for Control Variables

Table 5 presents the mean, standard deviation, minimum, and maximum for each variable used in the regressions. The figures following show how the means for each control variable compare for River Forest versus the comparison districts. Some of the differences are quite large—for example, on the percent low income. Readers should keep in mind that the evaluation design adjusts for these overall differences by controlling for district fixed effects. One way to implement these fixed effects is to calculate a residual value for each variable by subtracting the overall mean value of each variable across years from the annual values. After this adjustment the residual values for each district are equal to 0 across years. This is done both for the outcome variable (predicted test scores in math) and each of the control variables. Thus, all variables start with equal values on average. In addition, the model further adjusts for the residuals to account for the remaining effects of any changes in these controls over time.

The percent Black and percent low income were lower for River Forest and falling over time relative to the comparison districts. The percent Hispanic or American Indian and the percent Asian or Pacific Islander were also lower for River Forest than for the comparison districts with mixed changes over time. The percent mixed race was higher for River Forest and rose over time. The percent ELL was lower than for the comparison districts but stable. The percent IEP started out higher for River Forest but fell and ended up being lower suggesting it may have been a particularly important control variable to include in the model. Enrollment rose and fell while the lagged ELA scores fell during COVID and then rose later.

Table 5: Descriptive Statistics for Continuous Variables used in Regressions

Variable	Mean	Standard Deviation	Minimum	Maximum
Predicted Score Math	737	18.7	637	823
% Black	10.6	20	0	92.8
% Hispanic or American Indian	25.2	19.1	0.4	88.1
% Asian or Pacific Islander	7.8	10.9	0	56.7
% Multiracial	3.62	2.33	0	13.5
% ELL	15.4	12.3	0	55.4
% with an IEP	14.4	2.87	6.6	25.5
% Low Income	36.6	24.2	0	99.8
Log enrollment	7.39	0.886	5.07	9.64
Log enrollment squared	55.4	13	25.7	92.9
Enrollment	2,341	2,283	159	15,328
Lag of the Predicted ELA score	743	18.5	646	826

The data in this table cover 109 districts (including River Forest), 6 grade levels (3 through 8), and 10 years (2015 to 2025 excluding 2020 when there was no test score data). 2021 is included here but was excluded from the main regressions due to low test score participation rates. 2015 was excluded from the main regressions since there was no lag data available for that year. The regressions also included indicators for district, year, and grade. Enrollment was used as a weighting variable while log enrollment and log enrollment squared were included as covariates.

Figures Showing Student Characteristics used as Control Variables in Regressions

Figure A.1 Percent Black: River Forest versus Comparison Districts

Figure A.2: Percent Hispanic or American Indian: River Forest versus Comparison Districts

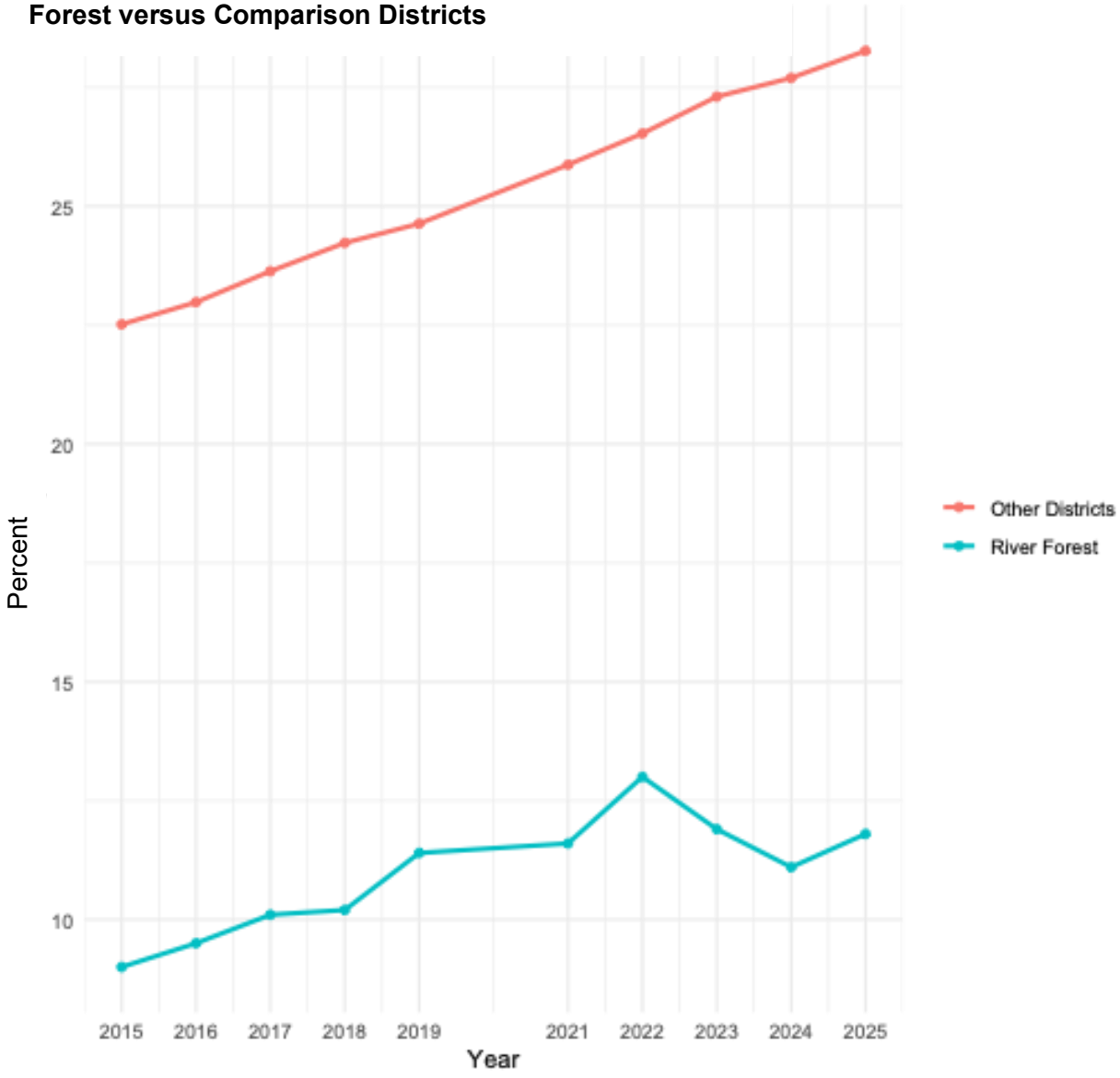


Figure A.3: Percent Asian or Pacific Islander: River Forest versus Comparison Districts

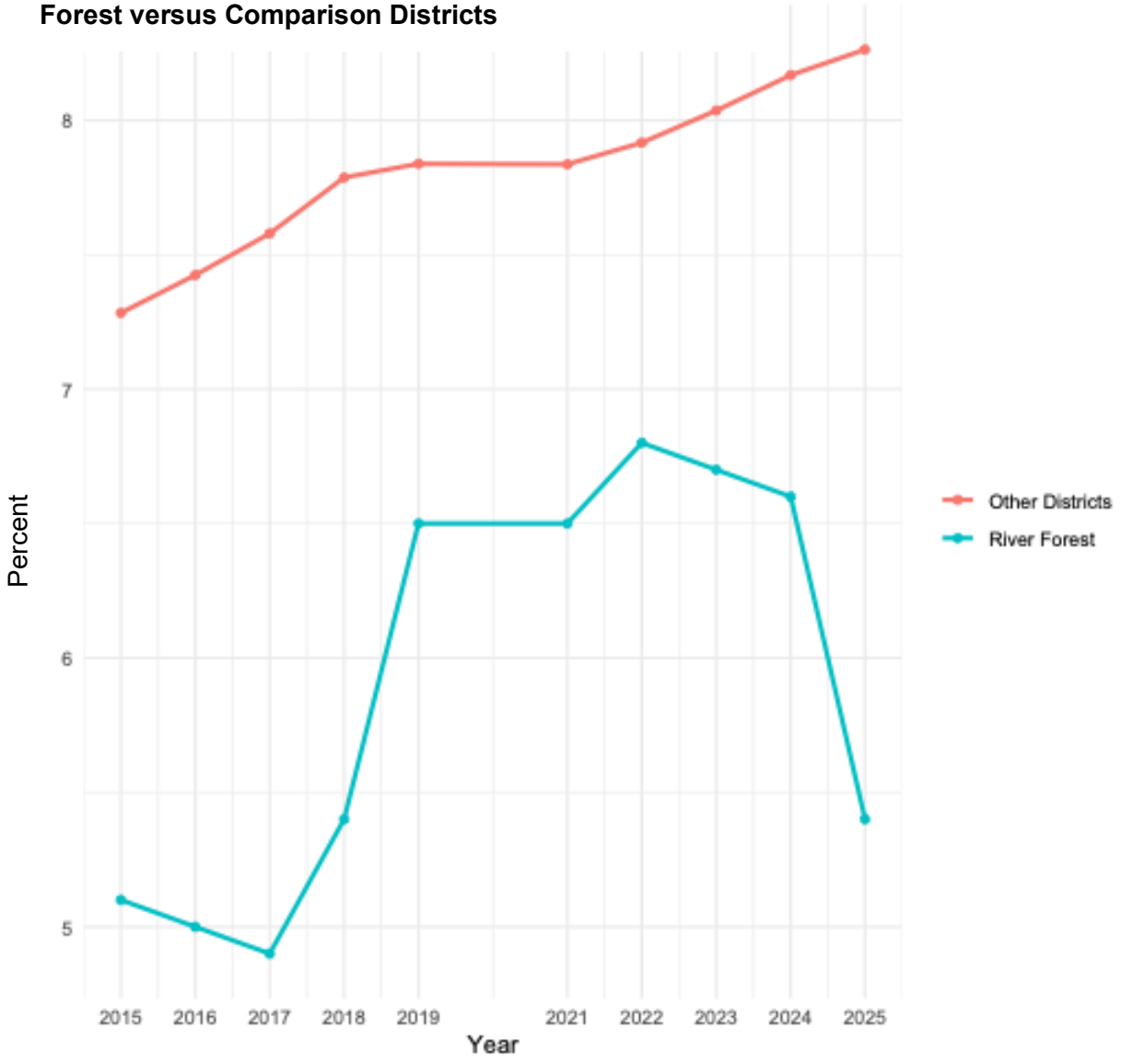
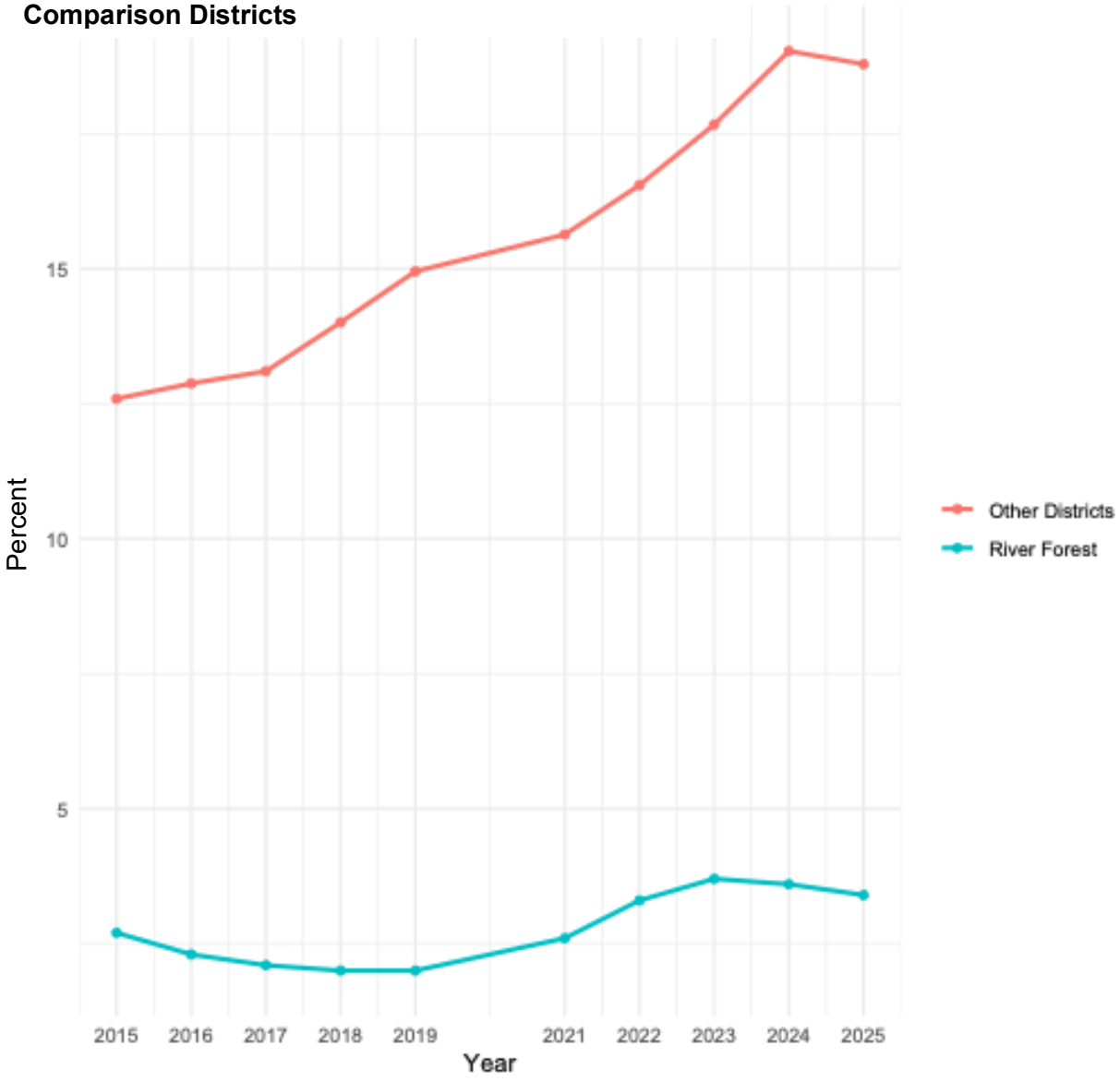


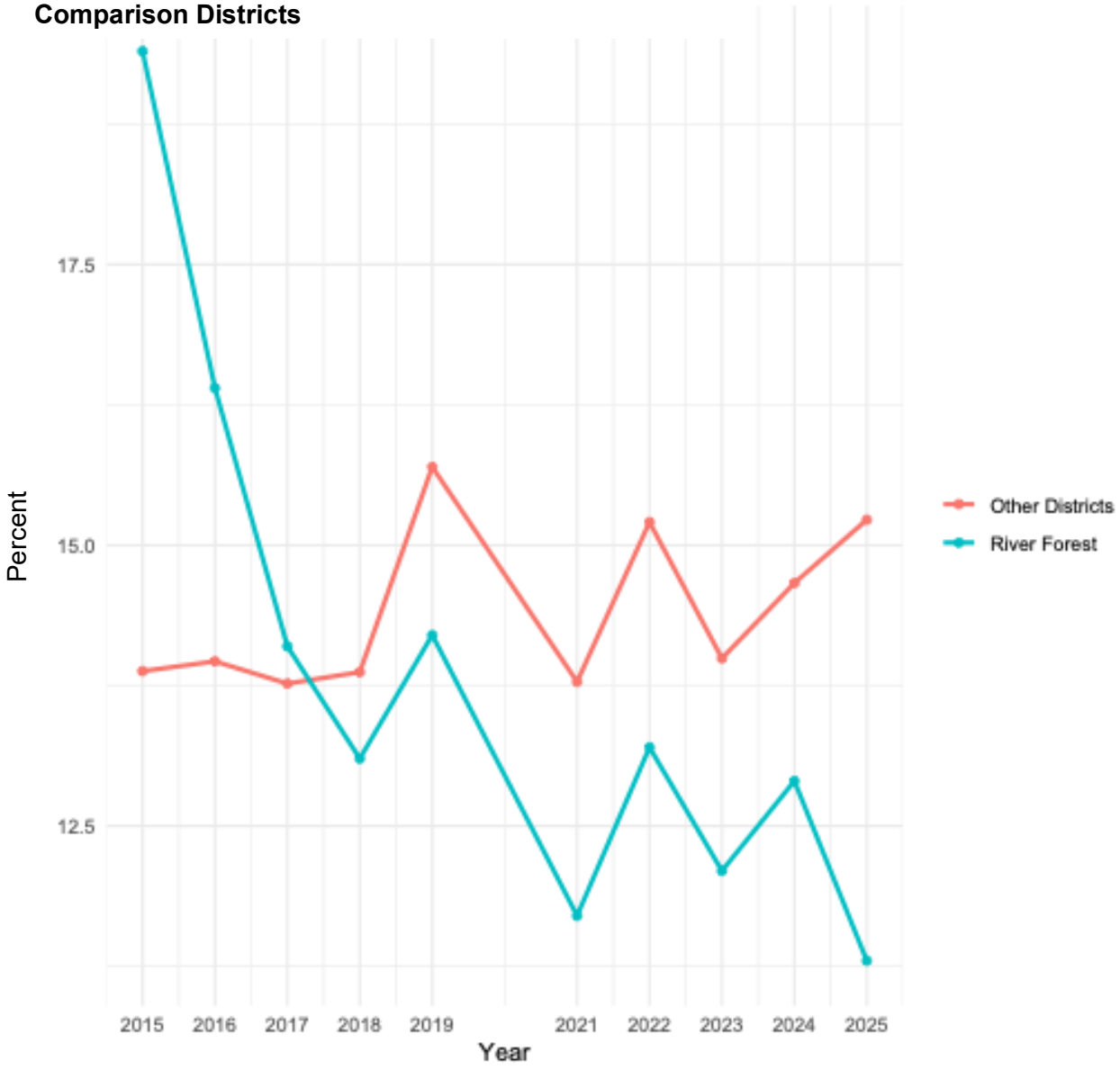
Figure A.4: Percent Multiracial: River Forest versus Comparison Districts

Figure A.5: Percent ELL: River Forest versus Comparison Districts



ELL means English Language Learner

Figure A.6: Percent IEP: River Forest versus Comparison Districts



IEP means Individualized Education Plan

Figure A.7: Percent Low Income: River Forest versus Comparison Districts

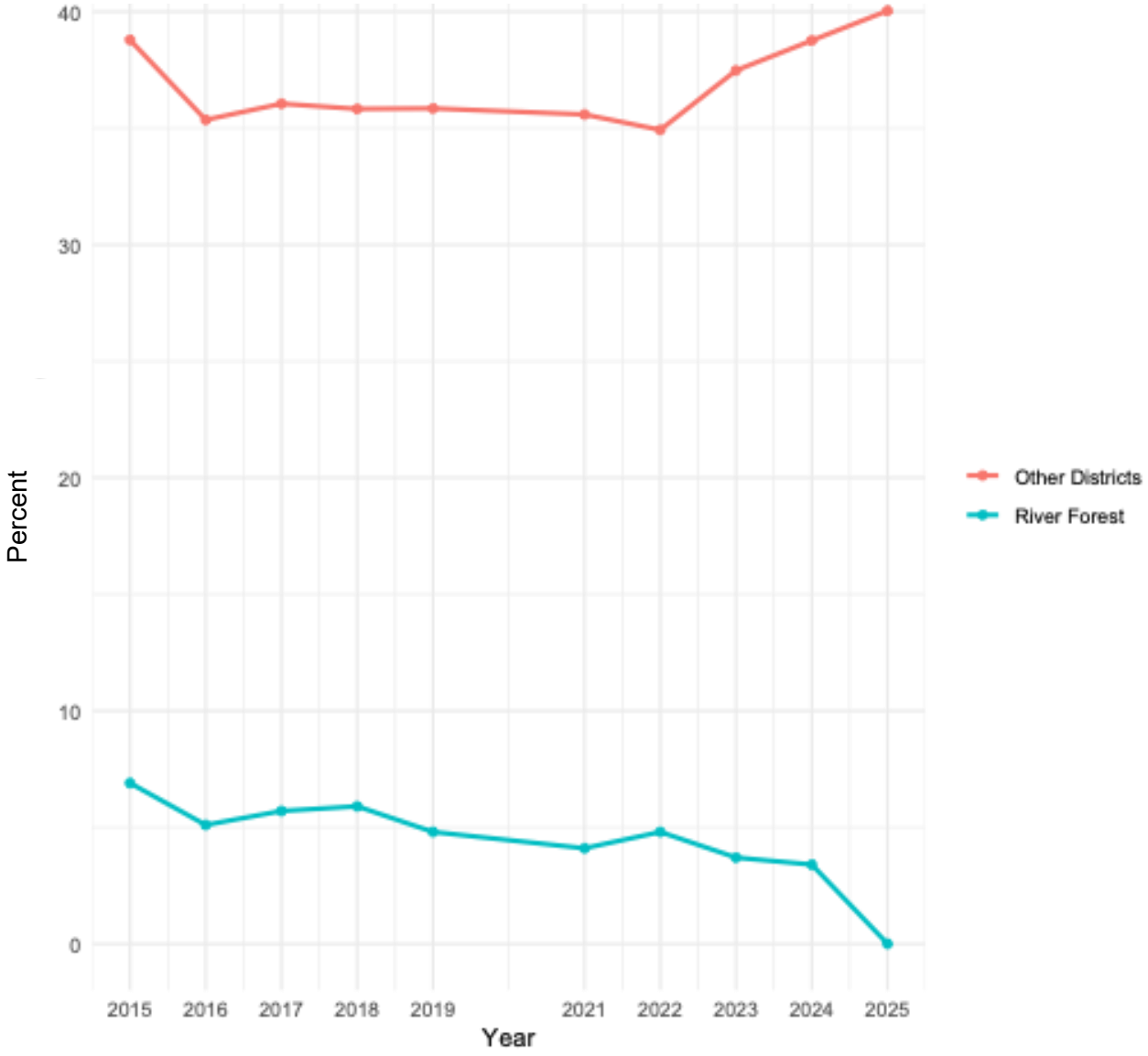


Figure A.8: District Enrollment: River Forest versus Comparison Districts

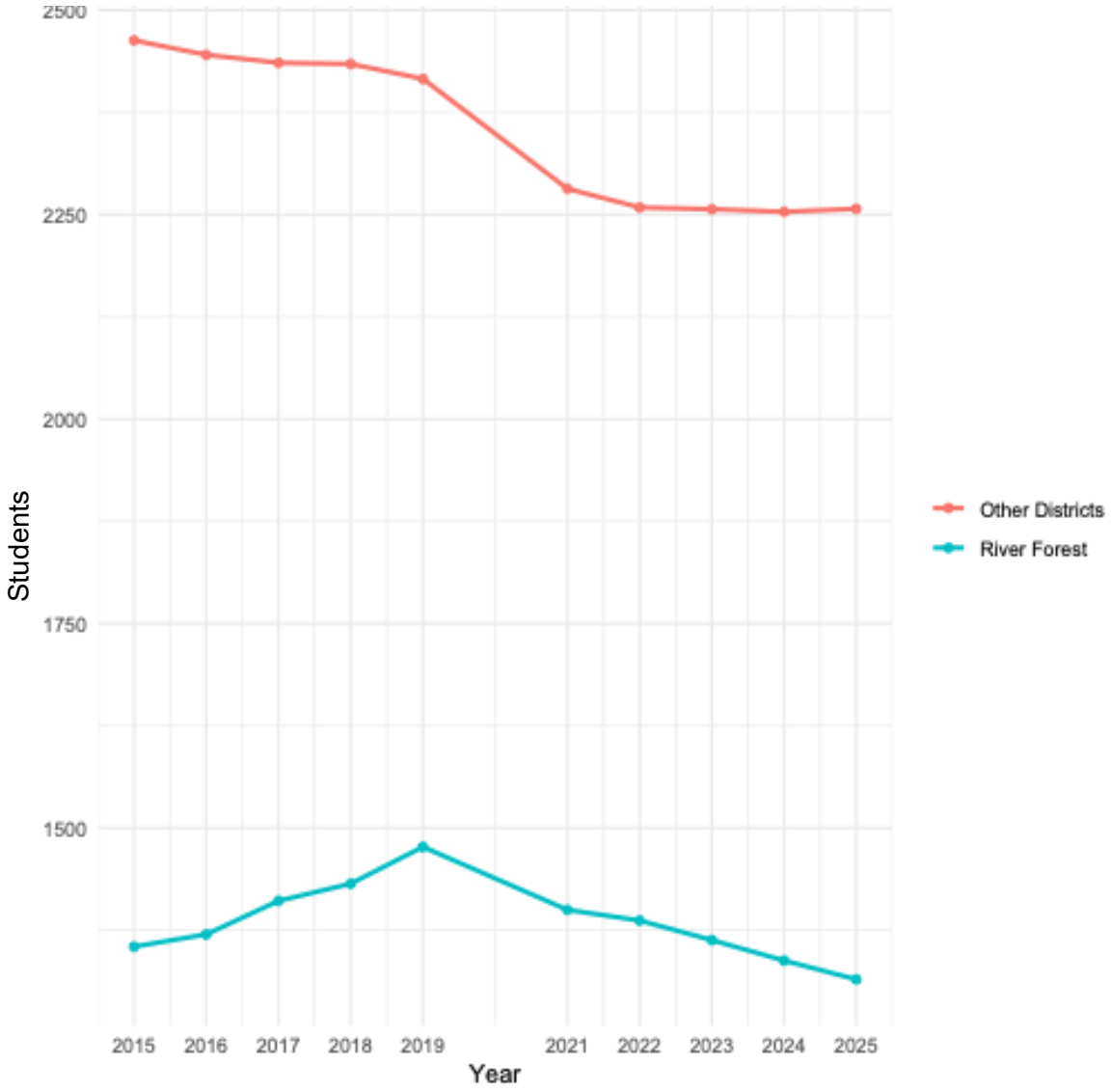
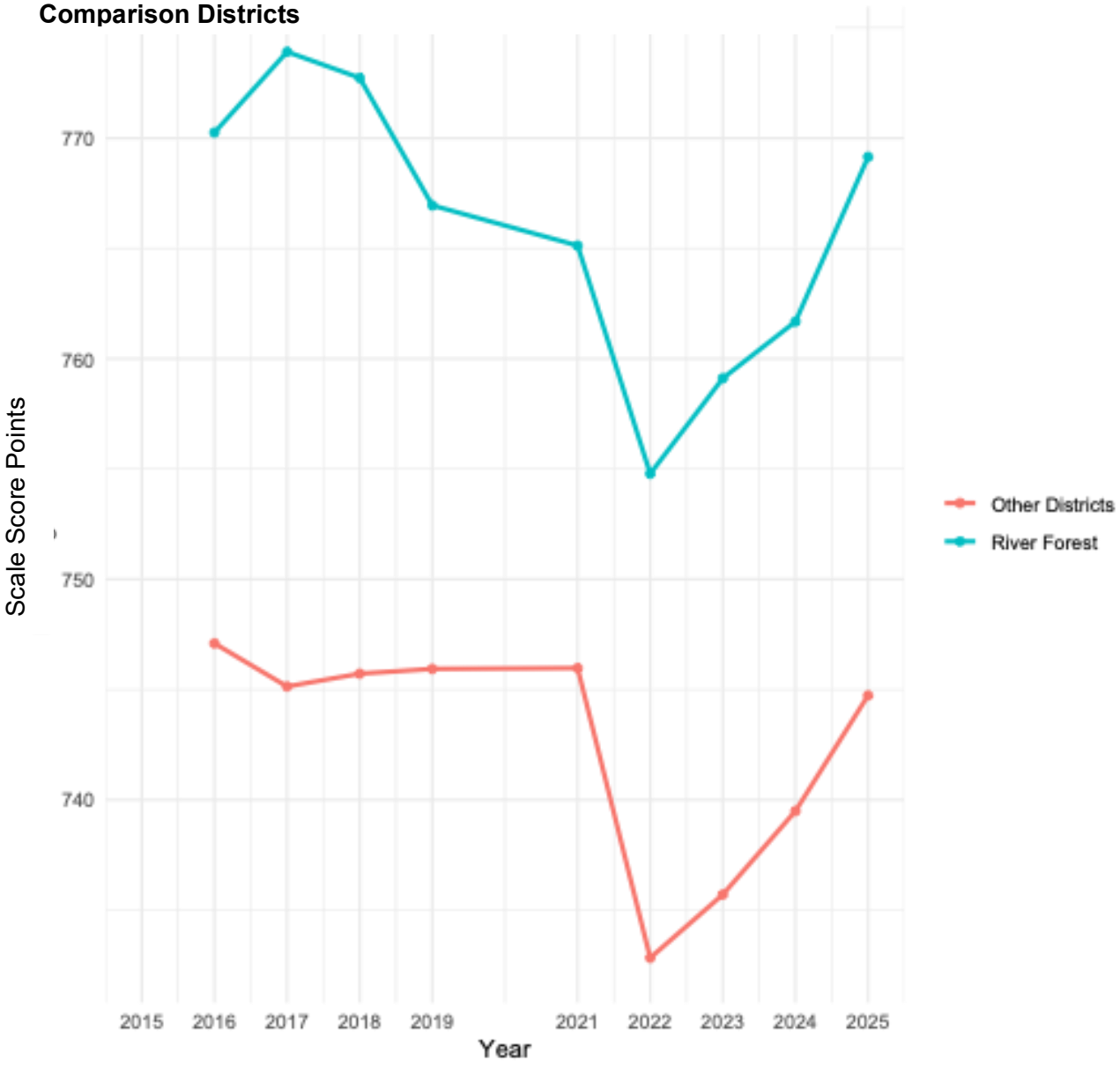


Figure A.9: Lag of Predicted ELA Scores: River Forest versus Comparison Districts



ELA means English Language Arts

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105(490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. Comparative politics and the synthetic control method. *American Journal of Political Science* 59(2): 495-510.

Bertrand, Marianne & Esther Duflo & Sendhil Mullainathan, 2004. "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, President and Fellows of Harvard College, vol. 119(1), pages 249-275.

Chaplin, Duncan (2026). "River Forest Public Schools, Math Curricula Evaluation." Presentation to the school board of the River Forest Public Schools district in Illinois, April 21.

Conley, Timothy, and Christopher Taber, 2011. "Inference with 'Difference in Differences' with a Small Number of Policy Changes." *The Review of Economics and Statistics*; 93 (1): 113–125.

Deke, John, Mariel Finucane, and Dan Thal (2022). "The BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations: A practical guide for education researchers." National Center for Education Evaluation and Regional Assistance, NCEE 2022-005, U.S. Department of Education.

Dotter, Dallas, Duncan Chaplin, and Maria Bartlett (2021). "Impacts of School Reforms in Washington, DC on Student Achievement." Report submitted to the Arnold Foundation. Washington, DC: Mathematica, August.

Earvolino, Phil (2024), "Similar District Modeling Results with WLS," Report produced for the River Forest School District.

Potamites, Liz, Kevin Booker, Duncan Chaplin, and Eric Isenberg (2009). "Measuring School and Teacher Effectiveness in the EPIC Charter School Consortium—Year 2 Final Report." Mathematica Policy Research, Washington, DC, October 23.

Reardon, Sean F., Benjamin R. Shear, Katherine E. Castellano, Andrew D. Ho (2017). "Using Heteroskedastic Ordered Probit Models to Recover Moments of Continuous Test Score Distributions From Coarsened Data." *Journal of Educational and Behavioral Statistics, 2017, Vol. 42, No. 1, pp. 3–45.*

Van Dine, Douglas, Bruce Randel, and Mary Klute (2021). "A Guide to Identifying Similar Schools to Support School Improvement." REL 2021-096, US Department of Education, Regional Educational Laboratory Central, Washington, DC, July.