IES Institute of Education Sciences

APPENDIXES
October 2021

Regional Educational Laboratory Mid-Atlantic
At Mathematica

Predicting Early Fall Student Enrollment in the School District of Philadelphia

See https://go.usa.gov/xMQhG for the full report.

Appendix A. Methodology

This appendix provides details on the study data, sample, and methodology.

Data

The study used student-level data, aggregated at the grade level for each school and year, for students enrolled in the School District of Philadelphia's (SDP) neighborhood K–8 schools during the 2015/16–2018/19 school years. The administrative datasets provided by SDP included a unique identifier to track students over time and across data files. Data files provided demographic information, school enrollment information, discipline information, and student achievement information (table A1).

Table A1. Administrative data used to predict incoming cohort sizes

Administrative data category	Predictor
Grade size	Current and previous grade sizes for the same student cohort and other cohorts. Lagged attrition from the target school-by-grade combination.
Attendance and suspensions	Aggregate and lagged attendance and suspension.
Test scores	Aggregate and lagged performance on the Pennsylvania System of School Assessment (standardized tests of literacy and numeracy), as well as three measures of English language arts (the Reading-Curriculum Based Measurement, Nonsense Word Fluency, and Oral Reading Fluency) in grades 1 and 2.
Demographics	Aggregate and lagged race/ethnicity, gender, economically disadvantaged status, English learner student status, and Individualized Education Program status.
Geographic data	Aggregate and lagged distance to neighborhood and alternative schools.
Structure	Grade and school indicator variables.

Source: Administrative data for 2015–19 provided by the School District of Philadelphia

The study team created variables to indicate whether a student had missing data on each predictor. These dichotomous (taking on values of either zero or one) variables were aggregated at the cohort level. Thus, each predictor in the final dataset has an associated missing data variable to indicate how many students had missing data for each specific variable. This is sometimes referred to as a "missingness incorporated into attributes" strategy (Twala et al., 2008).

Sample

The study used data on students who were in grades 1–8 in 174 SDP neighborhood schools in any year from 2015/16 to 2018/19. In each year there were about 1,000 grade-by-year units. The sample included all neighborhood schools in the district that enrolled any students in kindergarten through grade 8 and that were represented in at least two adjacent school years.

Table A2. School and cohort configuration in the 2017/18 school year

	Grade span in schools			Total in neighborhood schools
	K–5	6–8	K–8	
Schools	40	13	105	158
Cohorts (school-by-grade combinations)	201	24	711	936

Source: Administrative data for 2015–19 provided by the School District of Philadelphia.

The analytic sample contained 149,154 unique students, 50 percent of whom were Black, 21 percent of whom were Hispanic, 14 percent of whom were White, and 61 percent of whom economically disadvantaged (table A3).

Table A3. Demographic composition of the study sample, 2015–19

Characteristic	Frequency	Percent
Race/ethnicity	149,154	100.00
Black	74,325	49.83
Hispanic	31,990	21.45
White	21,117	14.16
Asian	11,204	7.51
Multiracial	10,092	6.77
American Indian/Alaska Native	294	0.20
Native Hawaiian/Pacific Islander	114	0.08
Unknown	18	0.01
Economically disadvantaged	91,012	61.02
English learner students	19,742	13.24

Source: Administrative data for 2015–19 provided by the School District of Philadelphia.

The demographic composition of the analytic sample was similar to that of SDP as a whole but markedly different from that of the overall U.S. student population (table A4). The share of English learner students is slightly larger in the study sample than in the entire district, which is to be expected given that the study did not include students in grades 9–12, who are less likely to be English learner students. The difference in the proportion of economically disadvantaged students between the analytic sample and the district as a whole is due to the fact that the study team calculated economic disadvantage based on individual student records, whereas the information reported by SDP is based on the U.S. Department of Agriculture's Community Eligibility Provision, which allows schools' economic disadvantage rate to be counted as 100 even if the true rate is less than that (Food and Nutrition Service, 2019).

Table A4. Demographic composition of students in grades K–8 in the School District of Philadelphia and the United States, 2017/18 school year (percent of total)

Characteristic	School District of Philadelphia	United States
Race/ethnicity		
Black	52	15
Hispanic	21	27
White	14	48
Asian	7	5
Multiracial	6	4
American Indian/Alaska Native	0.20	1
Native Hawaiian/Pacific Islander	0.07	—
Unknown	—	—
Economically disadvantaged	91	53
English learner students	11	10

— is not available.
Source: Administrative data for 2015–19 provided by the School District of Philadelphia; National Center for Education Statistics, United States Department of Education.

Methodology

This study addressed one primary research question and three subquestions

1. How well do machine learning algorithms predict incoming cohort sizes for grades 1–8 in SDP's neighborhood schools?

   1a. How well does each algorithm predict incoming cohort sizes for the following fall semester?

   1b. Does the precision of each algorithm differ for cohorts with a larger proportion of Black students, economically disadvantaged students, or English learner students?

   1c. Which administrative variables contribute most to the predictions?

The first two research tasks proceeded along similar lines. First, the study team built prediction models by applying one common prediction algorithm—ordinary least squares (OLS)—and three machine learning prediction algorithms—least absolute shrinkage and selection operator (LASSO), elastic net, and random forest—to retrospective data from the 2016/17 and 2017/18 school years. Those models were then used to predict incoming grade sizes in the 2018/19 school year. Performance was assessed by comparing predicted grade sizes to actual grade sizes.

Cross-validation was used to generate predictions from the modeling set. The study team randomly partitioned the data into 10 sets of roughly equal size. In each of 10 runs, one set was held out as the testing set, and the algorithms used the remaining nine sets to generate predictions of incoming grade size in the testing set. This process was repeated until each set had served as the testing set, after which predictions were averaged across all 10 runs. The same partitions were used by all algorithms: OLS, LASSO, elastic net, and random forest.

This analysis was conducted for the analytic sample as a whole (research question 1a) as well as by cohort level (school by grade by year), demographic characteristics (research question 1b), and grade level. Year-to-year changes in grade size might be more variable in some grades, especially those at natural phases of transition between school levels (for example, grades 5 and 6). The demographic characteristics examined were the cohort-level proportions of Black students, economically disadvantaged students (also known as low socioeconomic status students), and English learner students. Performance was assessed separately for cohorts in which the proportion of Black students was above the median for SDP (roughly 74 percent), cohorts in which the proportion of economically disadvantaged students was above the median (61 percent), and cohorts in which the proportion of English learner students was greater than 10 percent (which is near the 60th percentile). These demographic characteristics took priority because of concern about achievement gaps between these students and their peers (Fry, 2008; Hansen et al., 2018). Researchers have recently expressed concern that machine learning tasks undertaken to address social policy could interact with, and even exacerbate, inequality along dimensions of race/ethnicity, socioeconomic status, and the like (Buolamwini, 2018). For example, if an algorithm can more accurately predict incoming grade size for majority White cohorts than for majority Black cohorts, any resulting differences in teacher allocation might reduce instability more for White students. Although such algorithm-enabled inequality is hypothetical and might even be justified by overall improvements to classroom stability, the study team separated performance metrics by cohort demographic characteristics for the sake of transparency.

The analysis for research question 1c relied on the variable importance scores produced by the random forest algorithm, which slightly outperformed the other approaches. These scores, which are applied to all variables considered by the model, range from 0 for variables that are discarded as providing insufficient predictive power to 1 for the most important predictor.

For any machine learning endeavor, researchers must make many choices about the type of function (for example, regression trees versus LASSO) and which, if any, constraints are placed on the level of complexity (formally referred to as a regularizer). These choices are discussed below.

OLS. OLS regression models are widely used across many disciplines, and they are relatively easy to implement and interpret. One complication of using OLS is that it does not have built-in features to avoid overfitting—that is, making predictions that are accurate only for the sample data and that would not perform well for new observations—as the machine learning algorithms do. Avoiding overfitting is especially important with the large sets of potential predictors used in this study. The study team thus chose a reasonable set of regressors based on what SDP would likely use given the data available and the restrictions on number of variables the model could accommodate. The grade size and structure datasets (see table A1) were selected for the OLS models in the belief that these basic data capture some of the latent forces associated with student mobility, but the choice of variables is ultimately subjective. Another set of predictors, such as geographic and attendance variables, might have been preferable.[1]

LASSO. The LASSO algorithm extends OLS for cases in which the primary goal is prediction and there are many potentially correlated predictors (Tibshirani, 1996). The extension to OLS includes constraints (also known as regularization) on the magnitude of the regression coefficients by including a penalty to the model likelihood. This penalty forces coefficients of variables with low predictive power to 0, reducing the likelihood of overfitting the model to small signals in the training data. The magnitude of the penalty is controlled by a tuning parameter, or regularizer. The Stata command lasso was used to perform the LASSO analysis. The study team used 10-fold cross-validation to tune the lambda penalty parameter.

Elastic net. The elastic net algorithm extends the LASSO algorithm by including a second penalty term known as a ridge penalty (Zou & Hastie, 2005). As with the LASSO, the ridge penalty reduces the magnitude of the regression coefficients, but it does not reduce them to 0. By including both penalties, the elastic net algorithm attempts to

[1] When this analysis was being planned, the study team was advised by the district that students' home addresses would be rife with missing data, so geographic variables were not seriously considered as part of the OLS variable set. There was also concern among the study team that the highly skewed nature of attendance variables would lead them to be poor predictors on average, even if they were excellent at identifying a handful of cohorts that were likely to dramatically change size from year to year.

optimize the trade-off between the LASSO and the ridge. An additional tuning parameter controls the trade-off. The Stata command elasticnet was used to perform the elastic net analysis. The study team used 10-fold cross-validation to tune lambda (the LASSO penalty) and alpha (the mixing parameter).

Random forest. This algorithm splits the data into successive partitions based on values of the predictors until the divided sets accurately predict the outcome (Breiman, 2001). Specifically, the random forest algorithm is a collection of classification trees in which each tree consists of a series of binary decisions based on predictors. For example, first check current grade size of target cohort; if current grade size is above a certain threshold, check size of previous cohort at target grade level; and if size of previous grade level is below a certain threshold, check a certain demographic. After an observation passes through all these decisions, a prediction is made. The random forest algorithm consists of many such classification trees, in which each tree is built with a small amount of randomness, and the final predictions are averaged across the trees. The advantage of the random forest algorithm over the OLS, LASSO, and elastic net algorithms in that the random forest algorithm implicitly considers more complicated relationships with predictors (such as interactions and nonlinear effects), but the results of random forest algorithms are more difficult to interpret.

The Stata command rforest was used to perform the random forest analysis. The study team used 400 trees, did not explicitly restrict tree depth or leaf size, and set the number of variables to investigate in each tree to 155 (this includes school fixed effects). The number of trees and the number of variables to investigate in each tree were both chosen through a process called tuning, in which the study team tested the random forest model on a wide array of values and selected the number of trees and variables that minimized prediction error in the modeling set. This was done for the number of trees first; then, with the value of 400 trees having been chosen, the study team tuned the number of variables in models that used 400 trees.

Table A5 lists the predictors used in each algorithm. Aside from the limited set used for OLS, all predictors were used for the remaining algorithms.

Table A5. Administrative data used by each algorithm

	Prediction data set and task			
Administrative data category	OLS	LASSO	Elastic net	Random forest
Grade size	X	X	X	X
Attendance and suspensions		X	X	X
Test scores		X	X	X
Demographics		X	X	X
Geographic data		X	X	X
Structure	X	X	X	X

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' compilation based on administrative data for 2015–19 provided by the School District of Philadelphia.

References

Breiman, L. E. O. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Buolamwini, J. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*(1), 1–15.

Food and Nutrition Service. 2019. "Community eligibility provision." U.S. Department of Agriculture. Retrieved August 9, 2021, from https://www.fns.usda.gov/cn/community-eligibility-provision.

Fry, R. (2008). *The role of schools in the English language learner achievement gap.* Pew Hispanic Center.

Hansen, M., Mann Levesque, E., Quintero, D., & Valant, J. (2018). *Have we made progress on achievement gaps? Looking at evidence from the new NAEP results*. Brown Center Chalkboard, Brookings Institution. Retrieved August 9, 2021, from https://www.brookings.edu/blog/brown-center-chalkboard/2018/04/17/have-we-made-progress-on-achievement-gaps-looking-at-evidence-from-the-new-naep-results/.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*(1), 267–288.

Twala, B. E. T. H., Jones, M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950–956. https://doi.org/10.1016/j.patrec.2008.01.010.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix B. Supporting analyses

This appendix presents full results overall and for each demographic subgroup across which performance was assessed, followed by the top 15 predictors from the random forest analysis on the entire dataset.

Across the testing and extrapolation datasets, the range of median absolute deviation (MADs) is small. The random forest algorithm has the lowest MAD in the extrapolation set at 5.74, which is roughly 10 percent of the median cohort (school by grade) size in the extrapolation year (2018/19; table B1). The root mean squared errors (RMSEs) in the extrapolation set are similar, ranging from 11.21 for the elastic net algorithm to 11.70 for the OLS algorithm. The skewed nature of the cohort sizes drives the divergence between the MAD and the RMSE: a typical cohort is 60–70 students, but a handful are more than 300.

The proportion of school-by-grade cohorts subject to reallocation varies across algorithms and analytic sets. The elastic net results in the least amount of reallocation in the testing set (19 percent), whereas the random forest performs best in the extrapolation set (22 percent of cohorts reallocated; table B2).

Table B1. Predictive accuracy for incoming cohort sizes in neighborhood schools, by algorithm

	Median absolute deviation		Root mean squared error	
Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	6.31	5.99	16.21	11.70
LASSO	6.13	6.21	17.97	11.39
Elastic net	6.08	6.12	17.63	11.21
Random forest	5.91	5.74	14.16	11.40

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

Table B2. Proportion of school-by-grade cohorts subject to reallocation, by algorithm

Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	.26	.29
LASSO	.22	.23
Elastic net	.19	.23
Random forest	.24	.22

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

The pattern of results for cohorts in which the proportion of Black students is above the median for the School District of Philadelphia (SDP) is similar to that of cohorts in the analytic sample as a whole (tables B3 and B4). For the extrapolation set the random forest algorithm has the lowest MAD and results in the least amount of reallocation. The random forest algorithm has the lowest RMSE in the extrapolation set, unlike the RMSE patterns in the full dataset, in which the elastic net performs best. The RMSEs also appear smaller in general for cohorts in which the proportion of Black students is above the median for SDP (about 9) than for the analytic sample as a whole (about 11).

Table B3. Predictive accuracy for incoming cohort sizes in neighborhood schools among cohorts in which the proportion of Black students is above the median for the School District of Philadelphia, by algorithm

	Median absolute deviation		Root mean squared error	
Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	5.56	5.66	10.49	9.64
LASSO	5.11	6.33	12.61	9.05
Elastic net	5.13	6.29	12.14	9.03
Random forest	5.49	5.27	12.37	8.85

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia administrative data.

Table B4. Proportion of school-by-grade cohorts subject to reallocation among cohorts in which the proportion of Black students is above the median for the School District of Philadelphia, by algorithm

Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	.26	.27
LASSO	.22	.24
Elastic net	.18	.24
Random forest	.23	.21

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

The pattern of results for cohorts in which the proportion of economically disadvantaged students is above the median for SDP is somewhat similar to that of cohorts in the analytic sample as a whole (table B5). For the extrapolation set the random forest algorithm has the lowest MAD but—in contrast to the results for the analytic sample as a whole—does not result in the least amount of reallocation. The random forest algorithm has the lowest RMSE in the extrapolation set, unlike the results for the full analytic sample, in which the elastic net algorithm performs best. The RMSE also appears smaller in general for cohorts in which the proportion of economically disadvantaged students is above the median for SDP (about 9) than for the analytic sample as a whole (about 11). The LASSO and elastic net algorithms perform slightly better in amount of reallocation than the random forest algorithm in the extrapolation set, leading to 21 percent reallocation compared with 22 percent with the random forest algorithm (table B6).

Table B5. Prediction of incoming cohort sizes in neighborhood schools among cohorts in which the proportion of economically disadvantaged students is above the median for the School District of Philadelphia, by algorithm

	Median absolute deviation		Root mean squared error	
Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	5.99	5.55	10.58	9.59
LASSO	5.99	5.90	12.87	9.05
Elastic net	6.03	5.93	12.42	8.98
Random forest	5.62	5.22	11.91	8.93

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

Table B6. Proportion of school-by-grade cohorts subject to reallocation among cohorts in which the proportion of economically disadvantaged students is above the median for the School District of Philadelphia, by algorithm

Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	.25	.26
LASSO	.20	.21
Elastic net	.17	.21
Random forest	.22	.22

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

Predictive performance is slightly worse for cohorts in which the proportion of English learner students is greater than 10 percent. The pattern of results for these cohorts is similar across algorithms, with the random forest algorithm having the lowest MAD for the extrapolation set, tying for lowest amount of reallocation, and having a slightly higher RMSE than the elastic net algorithm (tables B7 and B8).

Table B7. Predictive accuracy for incoming cohort sizes in neighborhood schools among cohorts in which the proportion of English learner students is greater than 10 percent, by algorithm

	Median absolute deviation		Root mean squared error	
Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	7.14	6.67	17.22	14.47
LASSO	6.86	6.33	21.96	14.27
Elastic net	7.02	6.26	21.63	13.97
Random forest	6.27	6.25	15.52	14.29

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

Table B8. Proportion of school-by-grade cohorts subject to reallocation among cohorts in which the proportion of English learner students is greater than 10 percent, by algorithm

Algorithm	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
OLS	.27	.32
LASSO	.21	.25
Elastic net	.19	.25
Random forest	.25	.25

OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.
Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

No obvious patterns emerge by grade level under the random forest algorithm (table B9).[2] The ranking of grades with respect to cohort reallocation differs in the testing and extrapolation sets, probably because of the smaller number of cohorts used to calculate reallocation rates in each grade rather than some underlying pattern. It might

[2] Because of the use of lagged information on grade sizes and cohort, as well as information on past patterns of attrition from cohorts, the predictions can be made only for grades 2–7 rather than for grades 1–8.

be that the degree of fluctuations in cohort size varies across grades; these results merely mean that such fluctuations are predicted equally well (or poorly) across all grades.

Table B9. Proportion of school-by-grade cohorts subject to reallocation under the random forest algorithm, by grade level

Grade level	Testing set (2016/17, 2017/18)	Extrapolation set (2018/19)
Grade 2	22	30
Grade 3	32	23
Grade 4	25	28
Grade 5	33	19
Grade 6	31	32
Grade 7	30	27

Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

The variable importance scores separate naturally into four tiers. The random forest algorithm discards 118 predictors, largely consisting of fixed effects for individual schools, as irrelevant to prediction. The next 126 predictors have importance scores near 0 (median of .007, all less than .10) and are a mix of all types of predictors. The remaining two tiers stand out from the prior two, with 11 predictors having modest importance scores (between .10 and .30) and four having the greatest impact on accuracy (all with importance scores of .65 or higher; table B10). The discarding of most school indicators (school fixed effects) might not be surprising given the additional variables from which the models could choose, but the random forest algorithm does not find any of them to be particularly helpful. Only eight school indicators have scores of greater than .01; the highest score was .06. Likewise, the lagged attrition variable, with an importance score of .003, does not meaningfully improve predictions. With one exception, test score performance is not a strong predictor of incoming cohort size either. Most of the trichotomized test variables (literacy and numeracy scores) have importance scores of less than .1, with many near .01. Grade-level indicators all have scores near 0 or, in the case of grades 1 and 2, exactly 0.

Table B10. Relative importance of top 15 predictors of school-by-grade enrollment under the random forest algorithm

Predictive strength rank	Predictor	Predictor importance	Data category
1	Base-year enrollment (prior cohort) in same grade as prediction year	.97	Grade size
2	Number of students with more than five in-school suspensions	.91	Attendance and suspensions
3	Number of students with more than five out-of-school suspensions	.83	Attendance and suspensions
4	Number of students with fewer than six absences	.65	Attendance and suspensions
5	Number of students with Pennsylvania System of School Assessment English language arts scores between the 30th and 60th percentile	.27	Test scores
6	Number of Asian students	.24	Demographics
7	Number of students with missing Nonsense Word Fluency scores in the under 30th percentile category	.24	Test scores
8	Number of students with missing Nonsense Word Fluency scores in the 30th–60th percentile category	.22	Test scores
9	Number of students with missing Oral Reading Fluency scores in the 30th–60th percentile category	.21	Test scores
10	Number of students with missing Oral Reading Fluency scores in the under 30th percentile category	.20	Test scores
11	Number of students with missing Nonsense Word Fluency scores in the over 60th percentile category	.20	Test scores
12	Number of students with missing Oral Reading Fluency scores in the over 60th percentile category	.16	Test scores
13	Number of economically disadvantaged students	.15	Demographics
14	Number of students with absences between 6 and 10	.10	Attendance and suspensions
15	Base-year enrollment in the grade level below prediction grade (same cohort as prediction grade)	.10	Grade size

Source: Authors’ calculations based on administrative data for 2015–19 provided by the School District of Philadelphia administrative data.