

Between-School Variability in Achievement Gaps

Daniel Anderson<sup>1</sup>

<sup>1</sup> University of Oregon

Author Note

Correspondence concerning this article should be addressed to Daniel Anderson, 175 Education, 5262 University of Oregon, Eugene, OR, 97403. E-mail: daniela@uoregon.edu

## Abstract

Achievement gaps are well documented by income and race/ethnicity. Comparatively little research, however, has investigated between-school differences in these gaps. Using publicly available data from California, Oregon, and Washington from the 2014-15 to 2017-18 school years, we estimate school-level racial and economic achievement gaps by grade and content area. We find the majority of the variance in these estimates lies between schools (within district). We then investigate the extent to which these estimates depend upon geography, with visual displays (maps) providing clear evidence of spatial clustering. Finally, we build a computational model using the physical location of the school (longitude and latitude) to predict the achievement gap. We find approximately 18-61% of the total variability accounted for by location, depending on the specific model. We further use the residualized estimates to identify schools with unexpected achievement gaps. Implications for future research and educational systems-level reform are discussed.

*Keywords:* Achievement Gap, Effect Size, Spatial Variance, Machine Learning

Word count: X

### Between-School Variability in Achievement Gaps

Racial and economic achievement gaps are pervasive and persistent, having existed throughout the United States educational system “for at least as long as we have thought to ask the question and had sufficient data to verify the answer” (Reardon, 2013a). A wealth of previous research has investigated achievement gaps, including multiple large-scale longitudinal studies. These investigations generally suggest that racial achievement gaps narrowed considerably in the 1970’s and 1980’s (Hedges & Nowell, 1999; Konstantopoulos, 2005; Lee, 2002), remained stable through the 1990’s and early 2000’s (Lee, 2002; Reardon, 2013b), and have perhaps begun to slowly narrow again within the last decade (Center for Educational Policy Analysis, 2019). In terms of economic achievement gaps, the picture is considerably less clear. Reardon (2011) found differences in achievement between students from families in the 90th versus 10th percentiles of the income distribution to have increased relatively steadily since the 1940’s, which largely replicated findings from Konstantopoulos (2005). Hanushek, Peterson, Talpey, and Woessmann (2019), however, arrived at a different conclusion, stating “any increase in the disparities in wealth, earnings, and income that may have occurred over the past half century do not translate into an increased connection between students’ family backgrounds and their achievement levels” (p. 11). At best, however, economically-based achievement gaps have remained stable over time, which itself is a sign of a lack of progress.

Despite the wealth of previous research investigating achievement gaps, surprisingly little has investigated between-school differences. The determinants of achievement gaps are complex, relating to both communities and educational systems (Todd & Wolpin, 2007). Given the interplay of these factors, it is logical to assume that achievement gaps would differ by school, depending on the specific community in which the school is located and the specific practices employed by school personnel. At the school district level, recent evidence from Owens (2018), as well as Reardon, Kalogrides, and Shores (2019), both suggest strong variation between districts in terms of the magnitude of achievement gaps, with both studies

finding the variation to highly correlated with geography. Reardon et al. (2019) found achievement gaps “ranging from nearly zero in some places to larger than 1.5 SD in others” (p. 1166), with metropolitan areas generally having higher achievement disparities.

The purpose of this paper is to extend previous research by focusing specifically on school-level achievement gap variability, with a particular emphasis on geographic variance. Our analysis includes the estimation and exploration of school-level Hispanic-White, Black-White, and economically-based achievement gaps in California, Oregon, and Washington across the 2014-15 to 2017-18 school years. We estimate the proportion achievement gap variability attributable to between-school differences, as opposed to districts, counties, and states. We further provide screenshots of, and a link to, an interactive mapping application allowing for the exploration of between-school and geographic differences in achievement gaps. Finally, we report the results a computational model built predict school-level achievement gaps based on the physical location of the school (longitude and latitude). In what follows, we first discuss prior research on achievement gaps, including evidence of community- and school-inputs. We then discuss the specific methods and results of our study, before concluding with a discussion of the implications of our findings for practice, policy, and future research.

### **Community-Level Inputs**

Evidence from Reardon (2011) and Konstantopoulos (2005), among others (e.g., Dills, 2006) has indicated that economically-based achievement gaps have grown over time. This trend could be attributed, in part, to the overall rise of income inequality observed in the general population over the same time period (Saez, 2018; Saez & Zucman, 2016). Reardon (2011), however, cautions that the issue is more nuanced, stating “it is not rising income inequality per se that has caused the income achievement gap; rather, a dollar of income (or factors correlated with income) appears to buy more academic achievement than it did several decades ago.” (p. 17). As mentioned previously, however, recent evidence from Hanushek et al. (2019) indicates economic achievement gaps as having been largely stable

over time, which complicates the attribution of rising income inequality being related to economically-based achievement gaps. Making the issue more complex, race and income are correlated, with long-standing racial gaps in income having been observed (Blau & Graham, 1990; Hoover, Compton, & Giedeman, 2015).

Recent research by Owens (2018) and Reardon et al. (2019) indicate that neighborhood-level economic and racial segregation play a large-role in district-level achievement gaps. Owens (2018) found higher levels of income segregation were related to stronger effects of income on achievement gaps (i.e., family income mattered more in predicting achievement gaps when neighborhood segregation was high). Similarly, Reardon et al. (2019) found racial and economic segregation levels correlated with the White-Black achievement gap between 0.41 and 0.65 in metropolitan areas and 0.20 to 0.38 across districts, depending on the specific measure of segregation. For the White-Hispanic achievement gap, measures of segregation correlated between 0.37 and 0.71 in metropolitan areas, and 0.31 to 0.42 in districts.

Reardon et al. (2019) also found higher achievement gaps generally within metropolitan areas. Urban and economically distressed communities often face much higher rates of crime and poor health outcomes (Geronimus, 2000), both of which have been found to relate to achievement. For example, Milam, Furr-Holden, and Leaf (2010) found that higher perceived neighborhood violence related to significantly lower reading and math scores. Bowen and Bowen (1999) found increased neighborhood violence exposure related to lower attendance and poorer grades. Basch (2011a) outlines the reciprocal relation between health, poverty, and education in a series of articles, arguing that reducing health disparities will reduce the achievement gap. The author describes disproportionality in vision problems (Basch, 2011b), asthma (Basch, 2011c), and inattention and hyperactivity (Basch, 2011d) among other issues impacting school-aged urban youth at higher rates, which have been found to relate to achievement. A similar argument is made by Fiscella and Kitzman (2009). While little research to date has evaluated systematic patterns in these factors, it logically

follows that if neighborhood demographic segregation relates to achievement gaps, so too would stark differences in community safety and health.

### School-Level Inputs

Given the strong influence of community features on achievement gaps, it is worth considering the extent to which achievement gaps can be influenced by school systems. Although it is logical that school systems play a role, quantifying the magnitude of this effect is not straightforward. As Fryer and Levitt (2006) note, “it is not obvious how to separately identify the effect of school quality from one in which the influence of neighborhood quality on student outcomes grows with age” (p. 271).

Fryer and Levitt (2004) found evidence that the Black-White achievement gap grew from students’ time of school entry in kindergarten to the end of first grade. The authors speculated that school quality may play an important role in these trends, which was supported by preliminary evidence comparing achievement trajectories within versus across schools. Including a set of fixed effects for schools reduced approximately two-thirds of the differences observed in Black-White trajectories. In 2006, Fryer and Levitt extended their study to examine the first four years of schooling, rather than the first two (i.e., extending to the end of Grade 3). The authors found that the Black-White achievement gap grew, on average, approximately 0.10 standard deviations per year. Importantly, however, this trend was not observed for the Hispanic-White achievement gap in either study, which was instead found to narrow over time, complicating the attribution to a general “school quality” indicator. The 2006 study also showed far less support for between-school differences explaining the divergent trajectories, with the authors concluding that “all of the ground lost between first grade and third grade by Blacks is within rather than across schools” (p. 16). In both studies, however, the authors did not account for neighborhood or community inputs. It is therefore unclear whether the reduction was related to between-school or between-community differences.

In contrast to Fryer and Levitt (2004, 2006), Hanushek and Rivkin (2006) found

substantial differences in the Black-White achievement gap corresponding with estimates of school quality. The authors reanalyzed the ECLS-K data used by Fryer and Levitt, but also included evidence from the Texas Schools Project to examine differences between schools in Grades 3-8. The authors not only found that “the majority of the expansion of the achievement gap with age occurs between rather than within schools” (p. 4) for both sets of data (essentially counter to the results found by Fryer & Levitt, 2006), but also that identifiable characteristics of the school could account for much of the growth in achievement disparities, including student turnover, teacher experience, and the racial composition of students’ within the school. Unfortunately, the authors did not include Hispanic-White or economically-based achievement gaps in their investigation. Figlio and Karbownik (2017), did investigate economic achievement gap variation in Florida, and had a similar finding to Hanushek and Rivkin (2006), with more within-district (between-school) variation than between-district variation. Hanushek and Rivkin (2006) attribute the differences in their findings to a more precise decomposition of within and between school differences. Perhaps equally important, however, was that the model fit by Hanushek and Rivkin (2006) included zone-by-year fixed effects, which the authors argue “remove in a very general way all variation over time in neighborhood and local economic conditions... [including] the myriad changes documented to occur in ‘transitional neighborhoods’” (p. 19). The authors were therefore able to better parse between-school differences from between-community differences.

### **Summary**

Achievement gaps are long-standing and prevalent across the country. Prior research suggests that, at the school district level, there is strong geographical variance, with community-level segregation in particular playing an important role in district-level achievement gaps. The current study extends this line of research by focusing on the school level, rather than the district level. Although factors such as community-level variance are outside the direct control of schools, understanding the impact of these variables is critical to understanding and parsing the impact of schools, and how changes in practice or policy may

relate to school-level achievement disparities between student groups. Given that prior research has focused primarily on the district level, however, this manuscript focuses on baseline estimates of between-school variability, including geographic variance, in achievement gaps. The following three research questions are addressed.

1. To what extent do Hispanic-White, Black-White, and economically-based achievement gaps vary between schools, as opposed to between districts or counties?
2. To what extent do school-level achievement gaps vary spatially (geographically)?
3. How much of the variance in achievement gaps between schools can be accounted for by the physical location (longitude and latitude) of the schools?

## Method

### Data sources

Publicly reported percent proficient data were collected from the state websites of California, Oregon, and Washington by grade and student subgroups in English/Language Arts and Mathematics for the 2014-15 to 2017-18 school years (see California Department of Education (2018), Oregon Department of Education (2019), Washington Department of Education (2019)). In each state, these files also reported the total number of students in each grade level. Data files were standardized within each state, and then combined across states, such that student subgroups had common labels. In each state, the data were reported as the percentage of students in the corresponding year/grade/subgroup/content area across four performance categories, where a value of 1 represented the lowest performance category and a value of 4 represented the highest.

The minimum reporting requirements differed across states. In Washington, for example, any school with fewer than 10 students in the corresponding “cell” (year/grade/subgroup/content combination) reported missing data, while Oregon reported missing data for any cell with fewer than six students. In California, however, the minimum reporting requirements were based on the number of students tested, which did not necessarily correspond to the number of students with scores. In fact, 26 schools reported

percent proficient data on a single student (where the values were 0% in all performance categories except the category the student scored in, which was 100%). To make the data more consistent with other states, and ensure reasonably stable achievement gap estimates, we restricted the data to align with the reporting requirements used in Oregon, with the school-level data included only if greater than five students were included in the corresponding cell. Finally, data were restricted to only schools that reported data on students coded White and Black, White and Hispanic, or economically disadvantaged.

Following the construction of the dataset across states, we linked the data to files from the National Center for Education Statistics (NCES) Education Demographics and Geographic Estimates (NCES, 2018), which provided latitude and longitude values for approximately 99.9% of schools in our sample across states. These data were linked via the NCES school identifiers, which were obtained by linking the statewide identifiers to the NCES Common Core of Data files (NCES, 2019)<sup>1</sup>. The raw data included 10,962, 1,255, and 2,345 schools in California, Oregon, and Washington, respectively, while the final data included 9,631, 1,202, and 1,732 students across states, respectively.

**Achievement gap estimation.** As outlined by Ho and Reardon (2012), standardized mean differences (i.e., effect sizes) can be estimated from ordinal percent proficient data by using the cumulative sum of the percentages reported in each category to approximate the overall empirical cumulative distribution function (ECDF) for each group. The ECDFs for each group are then paired, and the area under the paired curve represents the probability that a randomly selected student from Group A (on the x-axis) will have a higher score than a randomly selected student from Group B (on the y-axis). This area can then be transformed into an effect size estimate, interpreted in standard deviation units, by

$$V = \sqrt{2}\Phi^{-1}(AUC) \quad (1)$$

---

<sup>1</sup> For a complete description of all files used, including the names of the files and the code used to produce the final dataset, please see the README file within the data folder of the online GitHub repository.

where  $\Phi^{-1}$  represents the inverse normal distribution. Under the assumption of respective normality,  $V$  is equivalent to Cohen's  $d$ . We used this approach to estimate school-level achievement gaps for each academic year by grade and content area.

Hispanic-White, Black-White, and economically based achievement gaps were estimated for every school with a sufficient number of students in the corresponding cell (i.e., at least 6 in California and Oregon, and 10 in Washington). Both of the racial/ethnic achievement gaps were direct comparisons between the corresponding groups. For economically-based achievement gaps, however, we compared the distribution of students coded as economically disadvantaged for a given school to the distribution for *all students*, which included students coded as economically disadvantaged. A more direct comparison would be between students who were and were not coded as economically disadvantaged, but states do not report separately for the latter group. Thus, while the economically-based achievement gap estimate provides a general indication of the difference in achievement for students coded economically disadvantaged, it is not a direct comparison like the other measures, and should be interpreted cautiously. A total of 9,279, 4,191, and 12,460 schools were included in evaluations of the Hispanic-White, Black-White, and economically based achievement gaps, respectively, across states.

## Analyses

**Decomposition of Achievement Gap Variance.** Research Question 1 concerned obtaining a baseline estimate of the degree to which achievement gaps vary between schools versus other factors. We addressed this question through the use of a multilevel model, with separate models fit for each achievement gap. The model was specified as

$$\begin{aligned} \text{gap}_{isdct} = & \beta_0 + \beta_1(\text{Math}_i) + \beta_2(g4_i) + \beta_3(g5_i) + \beta_4(g6_i) + \beta_5(g7_i) + \beta_6(g8_i) + \\ & \beta_7(g11_i) + \beta_8(g13_i) + S_{0s} + D_{0d} + C_{0c} + St_{0t} + \epsilon_i \end{aligned} \quad (2)$$

where  $\text{gap}_{isdct}$  represents achievement gap estimate  $i$  in school  $s$ , district  $d$ , county  $c$ , and state  $t$ . The  $\beta_1$  term represents the average achievement gap difference between

English/Language Arts and Mathematics, while  $\beta_2$  through  $\beta_8$  represent dummy-coded fixed effects comparing the average achievement gap in the corresponding grade,  $g$ , to the achievement gap in Grade 3. Finally, the  $S$ ,  $D$ ,  $C$ , and  $St$  terms represent random effects (achievement gap variability) for schools, districts, counties and states, respectively. Note that preliminary models were also fit including pseudo-cohort and academic year as random effects, but the proportion of the variability associated with these variables was very small ( $< 0.1\%$  combined) and the terms were therefore removed in favor of model parsimony.

Confidence intervals for all estimates were obtained by profiling the log-likelihood.

**Spatial Variability.** The second and third research questions both addressed geographic/spatial variability in achievement gaps, with the third specifically focused on the proportion of the total variability that can be accounted for by school location. We address these questions through both visual displays, using an interactive mapping application, and through a computational model with a weighted  $k$ -nearest neighbor algorithm. For all spatial analyses, we computed a weighted achievement gap average across grades for each school within a given content area and school year for each estimated achievement gap, where the weights were defined by the total enrollment within the given grade level. These weighted averages were computed because we were primarily interested in school-level achievement gaps, rather than differences in achievement gaps between grades within schools (and this later variance was modeled as part of Research Question 1). Thus, each school was represented by a single value for each content area/academic year/achievement gap combination.

In terms of the visual display, we built an interactive map using the *leaflet* package within the R statistical computing environment (Cheng, Karambelkar, & Xie, 2018; R Core Team, 2019), in combination with *shiny* (Chang, Cheng, Allaire, Xie, & McPherson, 2019). Points representing schools were overlaid on the map, and colored according to the estimated achievement gap with a diverging color palette. The application allows for the exploration of different achievement gaps (Hispanic-White, Black-White, or Economically

Disadvantaged-All Students) across different content areas (English/Language Arts or Mathematics) and school years (2014-15 to 2017-18). Note that, because not all schools reported data on all groups, the number of points appearing on the map changes based on the achievement gap selected.

For the computational model, we first built a model on the 2016-17 data using only latitude and longitude to predict the school level achievement gaps. The predictions from this model were then evaluated against the 2017-18 data both to evaluate the model against data not used in the model estimation (as a validation/testing dataset) and to evaluate the stability of achievement gaps from one year to the next. Following this evaluation, we modeled data across the 2014-15 to 2016-17 school years, again reserving the 2017-18 data for model evaluation. This first model was therefore a more “pure” measure of geographic variability, while the second also included prior achievement gap data in the estimation. In each model, we split the data (from the 2016-17 school year or 2014-15 to 2016-17 school years) into *training* and *validation* datasets. This split was conducted randomly, stratified by school district (to ensure schools within all districts were represented in both datasets), with approximately 75% of all cases being represented in the training dataset and the remaining 25% reserved for the validation set. The purpose of this splitting was to allow the model to be fit (trained) on one set of data, but validated against another (which was not included in the estimation of the model). This process provides an indication of the model performance and an estimate of the *out-of-sample* predictive accuracy of the model. The 2017-18 data served as the final validation dataset, which was reflective of what could be expected in terms of model stability and accuracy against new (annual) data collections.

We used a weighted  $k$ -nearest neighbor algorithm to predict school achievement gaps. As mentioned, the model including only 2016-17 data included latitude and longitude as the only predictors in the model. The second evaluation, including data across the 2014-15 to 2016-17 data, included one model with data pooled across years (with latitude and longitude again being the only predictor variables, but multiple observations represented for each

school), and a second including a linear predictor for time (representing an overall average increasing/decreasing trend in achievement gaps).

Standard  $k$ -nearest neighbor models are non-parametric and relatively intuitive, with the prediction for a given observation estimated as the average of the  $k$  nearest observations in the prediction space. This implies that a measure of distance is needed, as well as an optimal  $k$ . In our application, we used Euclidean distance which, in geospatial analyses, relates to the physical distance (as the crow flies) between two points (schools). The weighted  $k$ -nearest neighbor algorithm extends this algorithm by including weights based on distance (e.g., points located closer in the prediction space are weighted more heavily than those further away; see Hechenbichler & Schliep, 2004).

The optimal  $k$  was determined using the training data for each evaluation with 5 repeats of 10-fold cross-validation. This implies first splitting the training data into 10 randomly selected subsets, or folds, of roughly equal size, stratified by school district. The model for candidate  $k$  was then fit to 9 of the 10 folds, with predictions made on the left out fold. This process was then repeated for every  $k - 1$  combination of folds for each candidate  $k$ . Note that the five repeats implies that the training data were randomly split into 10 folds five times (which is different from 50-fold cross validation). The repeated  $v$ -fold cross validation procedure helps reduce the variance in predictions Kuhn (2014). Each candidate,  $k$  was then evaluated relative to the accuracy of predictions on the left-out folds across the repeated sampling, the using differences in root mean square error (RMSE) to evaluate model performance (for more information on cross-validation procedures, see Zhang & Yang, 2015).

We fit models with 1 to 50 neighbors and evaluated the predictive accuracy of each candidate  $k$  relative to the of the left out data. Separate models were fit to each achievement gap outcome for each content area. For each model, we identified the  $k$  that minimized the RMSE for the left out set, fit this model to the full training data, and evaluated its predictive accuracy against the test dataset. Finally, we compared the predictions from the final model, built on data from the 2016-17 or 2014-15 to 2016-17 school years, to the

achievement gaps data estimated from the 2017-18 school year. We used these models to evaluate the proportion of variance accounted for by geography alone (2016-17 data only), versus geography and prior achievement gap values. All data were prepared using the *tidyverse* suite of R packages, with all plots produced using the *ggplot2* package (Wickham, 2016, 2017), while all modeling was conducted using the *tidymodels* suite of packages (Max & Wickham, 2018). The *k*-nearest neighbor model specifically was fit using the *kknn* package (Schliep & Hechenbichler, 2016).

## Results

### Decomposition of Achievement Gap Variance

Our first research question addressed the total variability in achievement gaps lying between schools versus districts, counties, and states. Table ?? reports the fixed effects and variance components (in standard deviation units) for each model fit. As can be seen, the average hispanic\\_white achievement gap across states was estimated at -0.41, indicating that students coded as Hispanic scored, on average, about 0.4 standard deviations below students coded as White, while the average Black-White achievement gap was estimated at -0.55 and the average economically-based achievement gap was estimated at -0.25. The achievement gap in mathematics was slightly larger for both racial/ethnic achievement gaps, relative to English/Language Arts, while the achievement gap was virtually identical across content areas for the economically based achievement gap. Grade level was generally not a meaningful source of predictor (despite confidence intervals occasionally not including zero in the interval).

The average achievement gap varied between schools with a standard deviation of 0.23 for the Hispanic-White achievement gap, 0.20 for the Black-White achievement gap, and 0.22 for economically-based achievement gaps. Achievement gaps universally varied less between districts, than between schools within districts, with county-level variance being of roughly equal size to district-level variance. State-level variance differed across achievement gaps, being the second largest source of variability for economically-based achievement gaps, and

the smallest for each of the racial/ethnic achievement gaps. Given that only three states were represented, however, the uncertainty of these estimates was rather broad, as can be seen by the confidence intervals. Figure 1 provides a visual representation of the percentage of total variability across these factors for each achievement gap. Note that residual variance represents within-school (between-grade) variability.

### **Geographic variability in achievement gaps**

The second research question addressed the extent to which school-level achievement gaps vary spatially. We began our investigation with visual explorations through the development of an interactive mapping application. Figure 2 displays images from the application<sup>2</sup> of the bay area and Seattle for the Hispanic-White achievement gap in English/Language arts during the 2014-15 school year (the default settings). As can be seen, there is clear evidence of geographical clustering of schools according to the magnitude of the estimated achievement gap. The map of the bay area displays schools generally having much larger achievement gaps in San Francisco, Berkley, and San Rafael areas than in the Vallejo, Concord, or Hayward areas. Similarly, schools near the central of Seattle generally had larger achievement gaps than schools on the outskirts of the city. Although not presented as figures in this paper, similar trends can be observed when exploring mathematics, alternative academic years, and alternative achievement gaps. Of note, the patterns did change somewhat based on the achievement gap. For example, achievement disparities between students classified as economically disadvantaged as compared to all students were generally low in San Francisco, but became much more broad further south around around Palo Alto and Mountain View. Patterns in the Black-White achievement gap, however, were generally similar to those observed for the Hispanic-White achievement gap, albeit, with far fewer schools represented (and the vast majority of schools represented being in California).

---

<sup>2</sup> To view the application and explore geographic variance in achievement gaps, please visit the following link (which has been anonymized for peer-review):

<https://anono-for-peer-review.shinyapps.io/ach-gap-variability-shiny/>

Following the visual exploration of achievement gaps, we addressed our third research question, which sought to evaluate the proportion of the total variability between schools that could be accounted for by geography, using a  $k$ -nearest neighbor approach. We began by modeling only data from the 2016-17 school year, with separate models fit for each content area and each achievement gap. Figure 3 displays the the change in the average RMSE (across the repeated folds) for each achievement gap and content area. The value that minimized the RMSE is displayed by a point on the given line. When evaluating these plots, we noted that some of the neighbor values that minimized RMSE were very high (e.g., the number of neighbors that minimized RMSE for the Black-White achievement gap was 40). Yet, across models, the decreases in RMSE slowed rapidly after about 10 neighbors (shaded gray background). Inspecting the average  $R^2$  of the left out folds for a model with 10 neighbors, versus the model that minimized RMSE, confirmed minimial differences, with a maximum difference across models of 0.01. Further, when considering the context of the research question, using greater than 10 schools to predict the achievement gap for a given school begins to move beyond the immediate community (in most cases) that influence achievement disparities. We therefore chose a  $k$  of 10 as our optimal model for 2016-17 data.

Following our evaluation of the optimal  $k$ , we fit the model to the full training data, and evaluated the predictive accuracy of the resulting fit to the test data (i.e., the 25% removed during the initial split). We also evaluated the model predictions against the entire 2017-18 dataset (which was not used in model estimation) and, somewhat surprisingly, found the model did a marginally better job of predicting the 2017-18 data than the test set from 2016-17 (higher  $R^2$  values), indicating the estimates were highly stable across years. The RMSE and  $R^2$  values on each respective test dataset are reported for each achievement gap in the top half of Table ???. As can be seen, between 0.18% and 0.59% of the total variability was accounted for by the models.

The model fit described above provides an estimate of the degree to which school-level achievement gaps can be predicted by the achievement gaps of the closest 10 schools.

Presumably, if the achievement gap of a given school can be predicted by the achievement gap of schools in the same geographic proximity, then at least part of what drives achievement gaps are related to larger-scale factors beyond the individual school. However, we were also interested in cases in which the model failed, with specific schools having a much different achievement gap than would be expected based on the surrounding schools. We therefore reproduced the map with the 2017-18 data, but with the points colored according to the residualized estimates rather than the raw achievement gap, as displayed in Figure 4. It is worth noting that many schools in this representation appear to “fade away”, indicating the model did a good job of predicting the achievement gap for the individual school. This was a purposeful design choice, with the intention to highlight schools that were discrepant in either direction (i.e., larger or smaller achievement gaps than would be expected by the surrounding schools). These residualized estimates provide a critical source of information when considering school- versus community-level features relating to achievement gaps. While there are likely both school system and community features relating to the spatial variation in achievement gaps, schools that are highly discrepant from the overall trends may be functioning in a systematically different way. Follow-up qualitative investigations in these schools and communities may help form deeper understandings of this functioning.

Finally, we evaluated the extent to which the predictive accuracy of the model changed when including achievement gap estimates from prior years. Two separate models were fit, with the first including school year pooled over time (i.e., longitude and latitude remaining as the only predictors, but each school including multiple observations), and the second including a *wave* term coded as a linear function of time (i.e., 0, 1, 2, 3). Note that, for each model, the nearest neighbor in the predictor space could be the same school estimate from a prior year. The data were fit to the 2014-15 to 2016-17 data and again evaluated against the 2017-18 data. Results of the repeated 10-fold cross validation indicated considerably lower  $k$  values across models, ranging from XX for BLAH to XX for BLAH, suggesting at least some within-school data being used in the predictions. The model including *wave* universally had

a higher RMSE and was therefore not evaluated further.

Overall, the model including prior achievement gap estimates accounted for more variance than the model from a single year, and the variance accounted for by the model using the testing data from 2014-15 to 2016-17 was again very similar to the variance accounted for by the model in the 2017-18 data, as reported in the bottom half of Table ???. The models accounted for 0.29% to 0.59% of the total variability in between-school achievement gaps.

## Discussion

Achievement gaps are complex, deriving from a multitude of factors. In this paper, we evaluated differences in achievement gaps between schools, with a particular focus on geographic variation. Similar to Owens (2018) and Reardon et al. (2019), we found strong evidence of variation in achievement gaps being linked to geography. In particular, Reardon et al. (2019) found district-level achievement gaps represented in metropolitan areas to generally be larger than those represented in other areas. The current study generally aligns with these results, as illustrated through visual representations of school-level achievement gaps across California, Oregon, and Washington. Urban communities frequently have higher rates of crime and poor health outcomes, each of which have been shown to impact achievement (Basch, 2011a; Bowen & Bowen, 1999; Geronimus, 2000; Milam et al., 2010). These community-level variables alone, however, would not necessarily explain disparate outcomes between student groups, unless they were found to impact one group of students more than another (i.e., they may explain differences in absolute achievement, but not necessarily achievement gaps). Rather, it is likely that schools within urban communities are more likely to include students who have been impacted by these conditions at differing degrees. Poor health outcomes are correlated with neighborhood-level poverty (e.g., see Mode, Evans, & Zonderman, 2016), and racial/ethnic minorities are highly over-represented within these communities (Economic Innovation Group, 2018). The most likely group of students to be negatively impacted by community-level variables are therefore students from

a racial/ethnic minority and from economically disadvantaged backgrounds, which ultimately leads to higher achievement gaps.

In our study, we found more between school (within-district) variation in achievement gaps than between districts (although the latter remained an important source of variation). This finding highlights the need for additional research at the school level. For example, explorations of the residualized achievement gap estimates revealed some schools that were highly discrepant from those surrounding it. Understanding why these schools differ from others in their community may help to understand systems-level practices in education that relate to achievement gap variability which, if manipulated, could begin to help close achievement gaps at scale. Certainly the compendium of factors relating to achievement gaps go beyond the direct control of school personnel. Owens (2018) and Reardon et al. (2019), for example, both found indicators of community-level segregation to have a strong relation with achievement gaps in the area. While potentially a long-term policy and public planning goal, teachers and administrators cannot directly influence neighborhood segregation levels. However, that some schools are highly discrepant from those around them also provides promising evidence that *something* different is happening in those schools, which may relate to school practices, policies, climate, etc., all of which can be influenced by school personnel. Further, Hanushek and Rivkin (2006) did find evidence that differences in school quality related to changes in the Black-White achievement gap, providing further evidence that schools are an important factor in achievement gap variation.

Many (Carter & Welner, 2013; Flores, 2007; Milner IV, 2012) have discussed reframing achievement gaps as opportunity gaps. Differential opportunities lead to differential achievement. Unfortunately, there is strong evidence that part of these differential opportunities are baked into the fabric of typical school systems. Disproportionality in suspension and expulsion data by race/ethnicity and socioeconomic status is well documented and is perhaps as prevalent as achievement gaps themselves (e.g., see noltemeyer10; Skiba, Michael, Nardo, & Peterson, 2002; Skiba et al., 2011). In the largest

study on the topic to date, Morris and Perry (2016) found approximately 20% of the Black-White achievement gap was accounted for by disproportionality in out-of-school suspension rates. Conducting follow-up investigation with schools found in our study to be discrepant from those around them may reveal differences in these opportunity variables, which could ultimately lead to promising principles of practice for school systems that are committed to reducing achievement disparities and large-scale inequities.

### **Limitations**

Perhaps the primary limitation of this study was the lack of a direct comparison for students coded economically disadvantaged. Comparing this group of students to *all students* provides an incomplete picture of the achievement differences, and these analyses should be interpreted highly cautiously. Preliminary inspections of the overall distribution of these achievement gaps also revealed a skewed and essentially truncated (at the upper tail) distribution. As a general indicator, however, they are still useful in understanding differences in achievement between this group and the total population.

There are also various methods for calculating effect sizes from ordinal data, including methods that essentially smooth the approximated paired ECDF (e.g, through maximum likelihood with a receiver operating characteristic curve; see Ho & Reardon, 2012). This study used the simplest method, which did not include any smoothing, but has been found to perform well and represent the full continuous data well (Anderson & Stevens, 2018). Similarly, when evaluating between-school differences we estimated mean achievement gaps for each school across the grade bands. This is part of the limitation of not having student-level data. Had we of had student-level data, we could have calculated the difference in achievement for every student in each group at the corresponding school, which perhaps would have been a better indicator of the school-level achievement gap. Instead, we calculated weighted averages across grade levels (weighted by the number of students within the grade), which may have introduced a small amount of bias into the estimates.

Finally, it is worth mentioning that a variety of alternative methods could have been

used to address geographic variability, which may have produced different results. It would be interesting to evaluate, in future research, the difference in estimates between a  $k$ -nearest neighbor model and other computational models, such as a neural net. One of the advantages of the latter method is that a broad array of community-level variables could be included in the model, with “automatic” regularization implemented to reduce or eliminate the impact of any variable that does not relate to the outcome, and a more precise residualized estimate could be obtained. The residualized estimates used here were only based the expectation based on the surrounding schools, but if neighborhood-level segregation, safety, and health outcomes could be included in the model we may be able to better identify schools that are truly functioning in different ways that impact achievement gaps.

## Conclusions

It is possible that many of the community-level factors relating to achievement gaps are reciprocal, including the relation between economic prosperity and academic attainment. The 2018 Distressed Communities Index (Economic Innovation Group, 2018), for example, found that zip codes that were relatively prosperous had nearly six times the number of residents with a Bachelor’s degree or higher than zip codes that were economically distressed, while residents whose highest educational attainment was a high school diploma were over-represented in economically distressed zip codes. The report concludes “educational attainment has emerged as the sharpest fault-line separating thriving communities from struggling ones” (p. 2). Higher educational outcomes may therefore lead to higher economic prosperity, which in turn leads to higher educational outcomes. This findings is particularly important given that, as mentioned previously, the same report found racial/ethnic minorities were highly over-represented within economically distressed communities.

From a public-policy perspective, understanding the myriad factors that contribute to inequality is vitally important. From a school systems perceptive and educational research perspective, however, it is critical that we begin to parse this community-level variance related to inequality and achievement from school systems-level variance. If a more complete

understanding of the unique contribution specific practices on achievement gaps could be obtained, such as disproportionate school-wide discipline, school and district personnel could begin to address these specific practices. Given their historic prevalence and persistence, achievement gaps may feel overwhelming and intractable. Obtaining better understandings of actionable steps educators can take to reduce achievement disparities at scale, and the expected benefit of specific practices, could help schools and district better target their limited resources. Understanding the separate (and related) impacts from the community and the school system could toward this goal.

## References

- Anderson, D., & Stevens, J. (2018). Exploring and visualizing school achievement and school effects. Paper presented at the annual meeting of the National Council of Measurement in Education.
- Basch, C. E. (2011a). 1Healthier students are better learners: A missing link in school reforms to close the achievement gap. *Journal of School Health, 81*(10), 593–598.
- Basch, C. E. (2011b). 2Vision and the achievement gap among urban minority youth. *Journal of School Health, 81*(10), 599–605.
- Basch, C. E. (2011c). 3Asthma and the achievement gap among urban minority youth. *Journal of School Health, 81*(10), 606–613.
- Basch, C. E. (2011d). 4Inattention and hyperactivity and the achievement gap among urban minority youth. *Journal of School Health, 81*(10), 641–649.
- Blau, F. D., & Graham, J. W. (1990). Black-white differences in wealth and asset composition. *The Quarterly Journal of Economics, 105*(2), 321–339.
- Bowen, N. K., & Bowen, G. L. (1999). Effects of crime and violence in neighborhoods and schools on the school behavior and performance of adolescents. *Journal of Adolescent Research, 14*(3), 319–342.
- California Department of Education. (2018). *Test results for english language arts/literacy and mathematics*. Retrieved September 25, 2019, from <https://caaspp.cde.ca.gov/sb2018/ResearchFileList>
- Carter, P. L., & Welner, K. G. (2013). *Closing the opportunity gap: What america must do to give every child an even chance*. Oxford University Press.
- Center for Educational Policy Analysis. (2019). *Racial and ethnic achievement gaps*. Retrieved September 24, 2019, from <https://cepa.stanford.edu/educational-opportunity-monitoring-project/achievement-gaps/race/>
- Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2019). *Shiny: Web application framework for r*. Retrieved from <https://CRAN.R-project.org/package=shiny>

- Cheng, J., Karambelkar, B., & Xie, Y. (2018). *Leaflet: Create interactive web maps with the javascript 'leaflet' library*. Retrieved from <https://CRAN.R-project.org/package=leaflet>
- Dills, A. K. (2006). Trends in the relationship between socioeconomic status and academic achievement. *Available at SSRN 886110*.
- Economic Innovation Group. (2018). *From great recession to great reshuffling: Charting a decade of change across american communities*. Retrieved September 25, 2019, from <https://eig.org/wp-content/uploads/2018/10/2018-DCI.pdf>
- Figlio, D., & Karbownik, K. (2017). Some schools much better than others at closing achievement gaps between their advantaged and disadvantaged students. *Washington, DC: Brookings Institution, July, 20*.
- Fiscella, K., & Kitzman, H. (2009). Disparities in academic achievement and health: The intersection of child education and health policy. *Pediatrics, 123*(3), 1073–1080.
- Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap? *The High School Journal, 91*(1), 29–42.
- Fryer, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics, 86*(2), 447–464.
- Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review, 8*(2), 249–281.
- Geronimus, A. T. (2000). To mitigate, resist, or undo: Addressing structural influences on the health of urban populations. *American Journal of Public Health, 90*(6), 867.
- Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2019). *The unwavering ses achievement gap: Trends in us student performance*. National Bureau of Economic Research.
- Hanushek, E. A., & Rivkin, S. G. (2006). *School quality and the black-white achievement gap*. National Bureau of Economic Research.
- Hechenbichler, K., & Schliep, K. (2004). Weighted k-nearest-neighbor techniques and ordinal classification.

- Hedges, L. V., & Nowell, A. (1999). Changes in the black-white gap in achievement test scores. *Sociology of Education*, 111–135.
- Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal “proficiency” categories. *Journal of Educational and Behavioral Statistics*, 37(4), 489–517.
- Hoover, G. A., Compton, R. A., & Giedeman, D. C. (2015). The impact of economic freedom on the black/white income gap. *American Economic Review*, 105(5), 587–92.
- Konstantopoulos, S. (2005). Trends of school effects on student achievement: Evidence from nls: 72, hsb: 82, and nels: 92.
- Kuhn, M. (2014). Retrieved from <http://appliedpredictivemodeling.com/blog/2014/11/27/vpuig01pqbklni72b8lcl3ij5hj2qm>
- Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, 31(1), 3–12.
- Max, K., & Wickham, H. (2018). *Tidymodels: Easily install and load the 'tidymodels' packages*. Retrieved from <https://CRAN.R-project.org/package=tidymodels>
- Milam, A., Furr-Holden, C., & Leaf, P. (2010). Perceived school and neighborhood safety, neighborhood violence and academic achievement in urban school children. *The Urban Review*, 42(5), 458–467.
- Milner IV, H. R. (2012). Beyond a test score: Explaining opportunity gaps in educational practice. *Journal of Black Studies*, 43(6), 693–718.
- Mode, N. A., Evans, M. K., & Zonderman, A. B. (2016). Race, neighborhood economic status, income inequality and mortality. *PloS One*, 11(5), e0154535.
- Morris, E. W., & Perry, B. L. (2016). The punishment gap: School suspension and racial disparities in achievement. *Social Problems*, 63(1), 68–86.
- NCES. (2018). *Education demographic and geographic estimates*. Retrieved September 25, 2019, from <https://nces.ed.gov/programs/edge/Geographic/SchoolLocations>
- NCES. (2019). *Common core of data: America's public schools*. Retrieved September 25,

- 2019, from <https://nces.ed.gov/ccd/>
- Oregon Department of Education. (2019). *Assessment group reports*. Retrieved September 25, 2019, from <https://www.oregon.gov/ode/educator-resources/assessment/Pages/Assessment-Group-Reports.aspx>
- Owens, A. (2018). Income segregation between school districts and inequality in students' achievement. *Sociology of Education*, 91(1), 1–27.
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither Opportunity*, 91–116.
- Reardon, S. F. (2013a). No rich child left behind. *New York Times*, 4(28), 13.
- Reardon, S. F. (2013b). The widening income achievement gap. *Educational Leadership*, 70(8), 10–16.
- Reardon, S. F., Kalogrides, D., & Shores, K. (2019). The geography of racial/ethnic test score gaps. *American Journal of Sociology*, 124(4), 1164–1221.
- Saez, E. (2018). Striking it richer: The evolution of top incomes in the united states. In *Inequality in the 21st century* (pp. 39–42). Routledge.
- Saez, E., & Zucman, G. (2016). Wealth inequality in the united states since 1913: Evidence from capitalized income tax data. *The Quarterly Journal of Economics*, 131(2), 519–578.
- Schliep, K., & Hechenbichler, K. (2016). *Kknn: Weighted k-nearest neighbors*. Retrieved from <https://CRAN.R-project.org/package=kknn>
- Skiba, R. J., Horner, R. H., Chung, C.-G., Rausch, M. K., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of african american and latino disproportionality in school discipline. *School Psychology Review*, 40(1).
- Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review*,

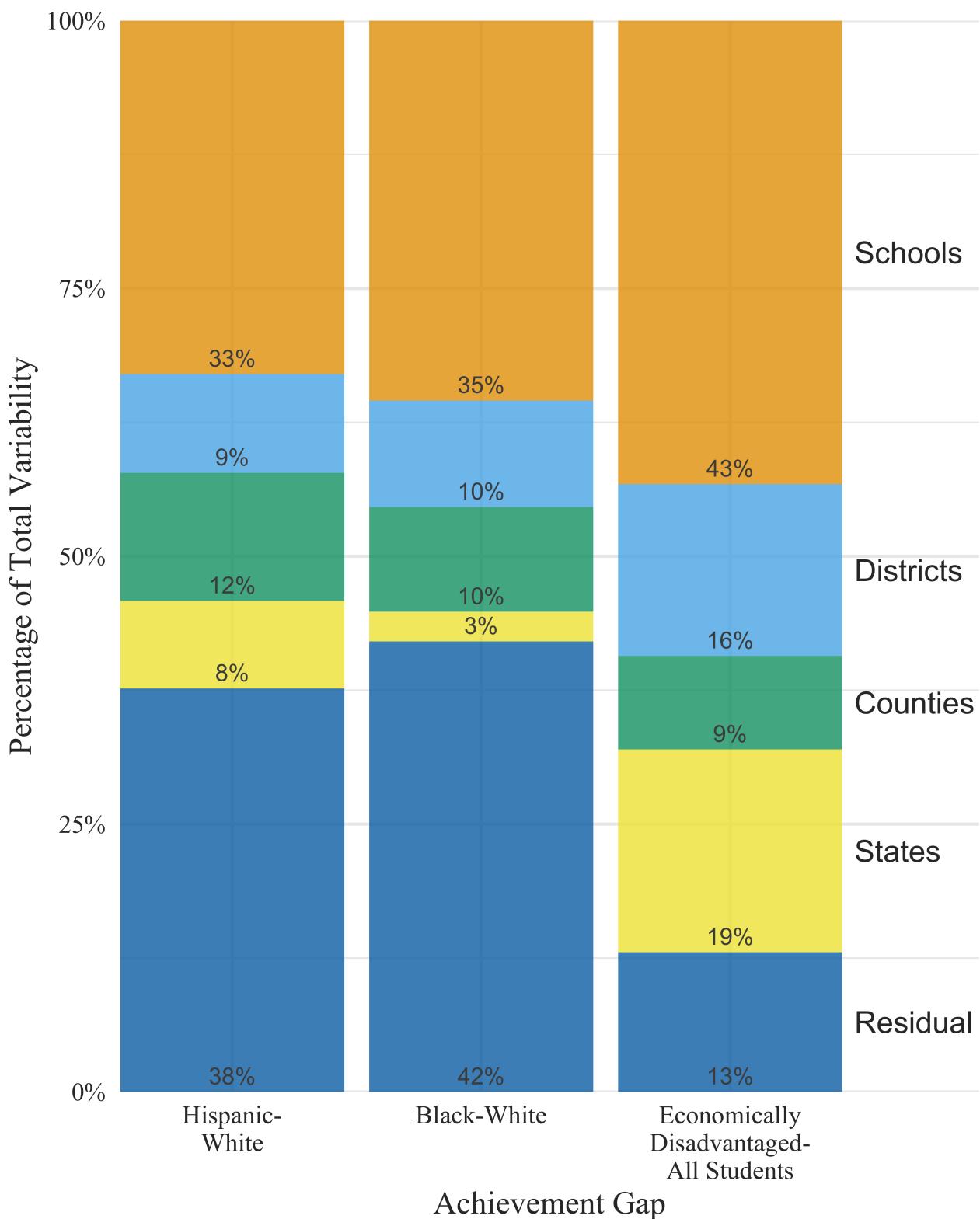
- 34(4), 317–342.
- Todd, P. E., & Wolpin, K. I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital*, 1(1), 91–136.
- Washington Department of Education. (2019). *Data portal*. Retrieved September 25, 2019, from <https://www.k12.wa.us/data-reporting/data-portal>
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>
- Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from <https://CRAN.R-project.org/package=tidyverse>
- Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1), 95–112.

Table 1. *Parameter estimates from multilevel models.*

Parameter	Hispanic-White			Black-White			Ec. Dis.-All Students		
	Est.	LB	UB	Est.	LB	UB	Est.	LB	UB
<b>Fixed effects</b>									
Intercept	-0.41	-0.58	-0.25	-0.55	-0.68	-0.44	-0.25	-0.44	-0.07
Math	-0.03	-0.03	-0.03	-0.06	-0.06	-0.06	-0.00	-0.00	0.00
Grade 4	-0.01	-0.01	-0.00	-0.02	-0.03	-0.00	-0.00	-0.00	-0.00
Grade 5	-0.01	-0.01	-0.01	-0.02	-0.03	-0.00	-0.00	-0.00	-0.00
Grade 6	-0.00	-0.01	0.00	0.00	-0.01	0.02	-0.00	-0.00	0.00
Grade 7	-0.01	-0.02	-0.01	-0.01	-0.02	0.01	0.00	-0.00	0.00
Grade 8	0.01	0.00	0.01	0.02	0.00	0.03	0.01	0.00	0.01
Grade 11	-0.01	-0.02	-0.01	0.01	-0.01	0.02	-0.00	-0.01	-0.00
Grade 13	0.00	-0.00	0.01	0.02	0.00	0.03	-0.00	-0.00	-0.00
<b>SD</b>									
Schools	0.23	0.23	0.24	0.20	0.19	0.19	0.22	0.21	0.22
Districts	0.12	0.11	0.14	0.12	0.11	0.13	0.12	0.11	0.13
Counties	0.12	0.09	0.17	0.09	0.07	0.10	0.13	0.11	0.16
States	0.07	0.00	0.21	0.13	0.06	0.34	0.11	0.05	0.30
Residual	0.26	0.25	0.26	0.11	0.11	0.11	0.24	0.24	0.24

Table 2. *Model performance for optimal k-nearest neighbor.*

Model/Test Dataset	Achievement Gap	ELA			Mathematics		
		k	RMSE	R^2	k	RMSE	R^2
<b>Geography only</b>							
2016-17	Black-White	10	0.34	0.22	10	0.31	0.18
2016-17	Ec. Dis-All	10	0.17	0.47	10	0.18	0.47
2016-17	Hispanic-White	10	0.28	0.23	10	0.29	0.21
2017-18	Black-White	10	0.31	0.25	10	0.30	0.27
2017-18	Ec. Dis-All	10	0.14	0.57	10	0.14	0.59
2017-18	Hispanic-White	10	0.26	0.30	10	0.25	0.31
<b>Geography &amp; prior estimates</b>							
2014-15 to 2016-17	Black-White	7	0.30	0.29	6	0.29	0.35
2014-15 to 2016-17	Ec. Dis-All	4	0.18	0.60	5	0.18	0.61
2014-15 to 2016-17	Hispanic-White	5	0.26	0.38	5	0.25	0.39
2017-18	Black-White	5	0.30	0.29	5	0.29	0.35
2017-18	Ec. Dis-All	4	0.18	0.60	6	0.18	0.61
2017-18	Hispanic-White	3	0.26	0.38	6	0.25	0.39



*Figure 1.* Proportion of achievement gap variance attributable to various sources. Note the residual represents within-school (between grade) variation.

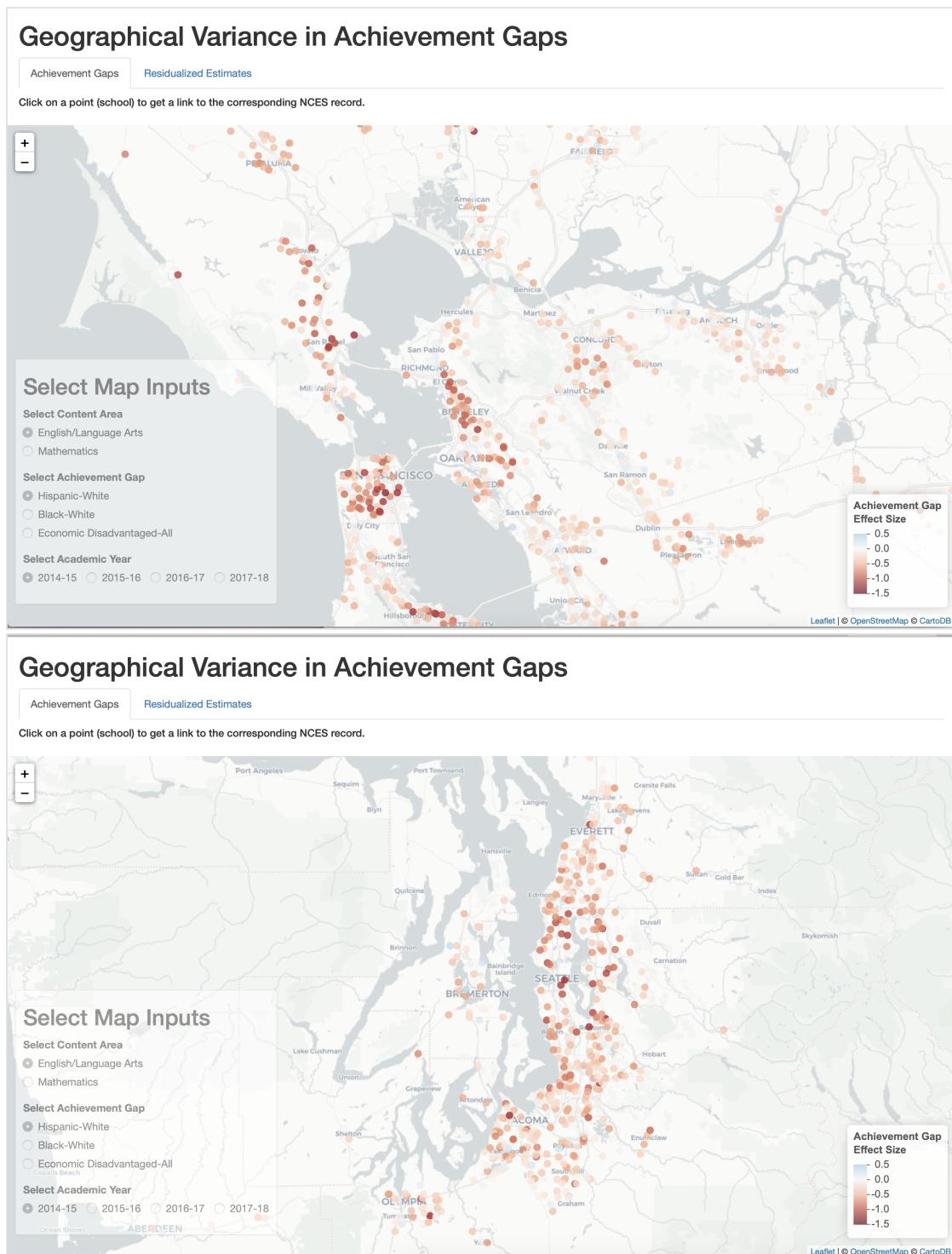
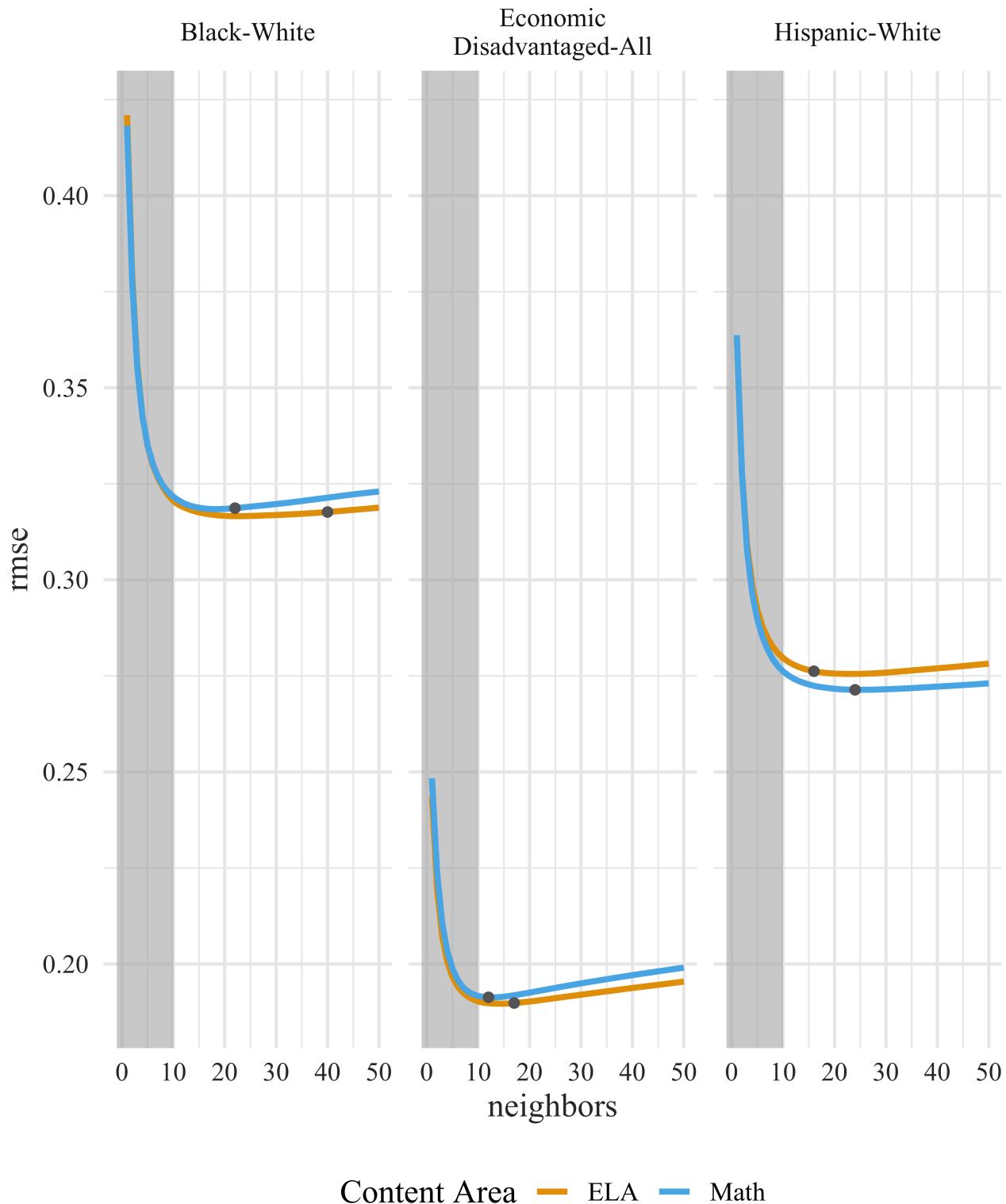
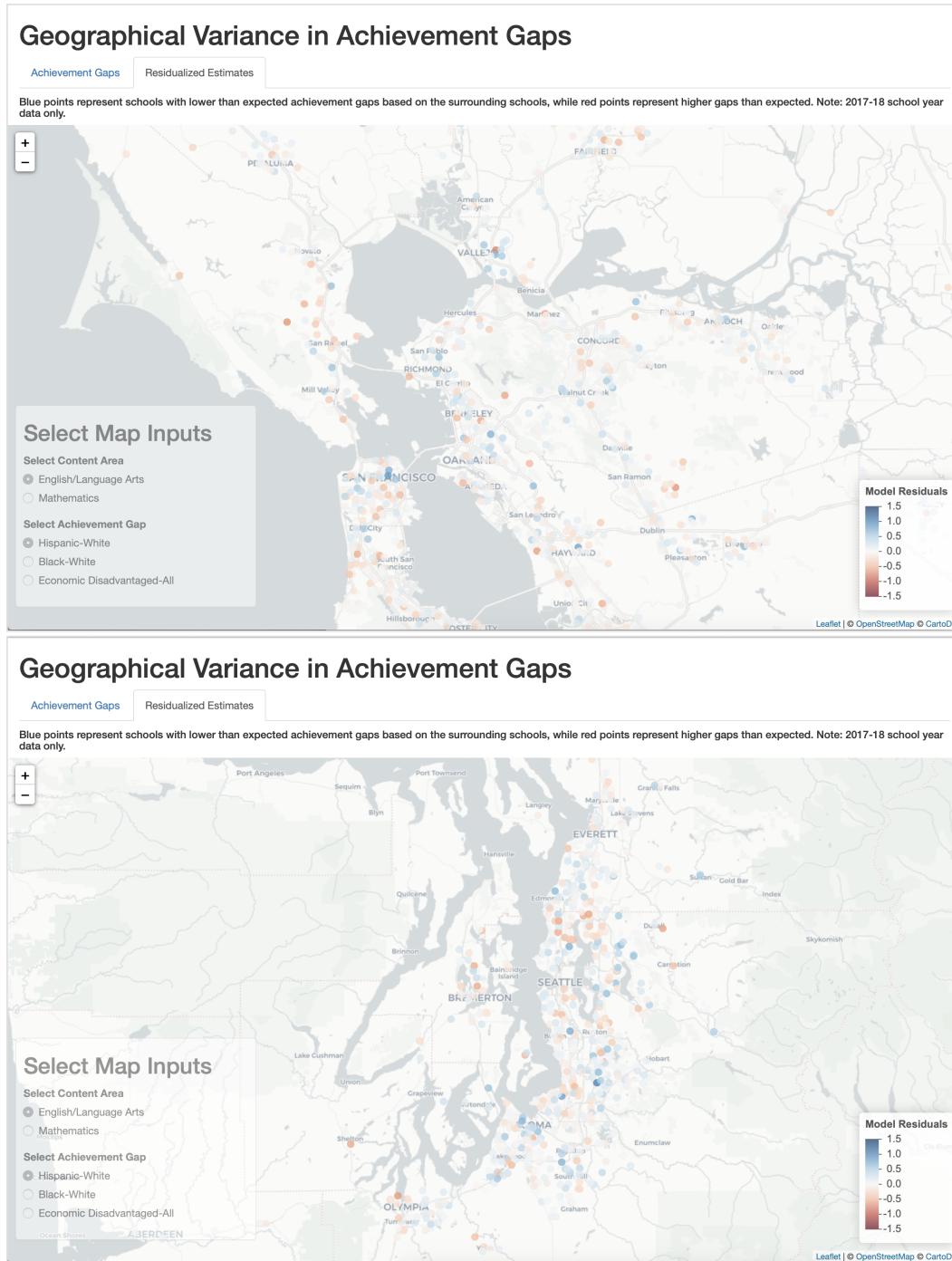


Figure 2. Screenshots of interactive web application. To explore the full application, please visit the following link: <https://anono-for-peer-review.shinyapps.io/ach-gap-variability-shiny/>



*Figure 3.* Average root mean square error (RMSE) over five repeats of 10-fold cross-validation for k-nearest neighbor models with 1 to 50 neighbors. Models included latitude and longitude as a predictor of the 2016-17 school-level achievement gap data.



*Figure 4.* Screenshots of interactive web application with residualized estimates shown. Blue points represent schools with lower achievement gaps than would be expected by the model, while red points represent schools with higher achievement gaps than would be expected by the model. To explore the full application, please visit the following link: <https://anono-for-peer-review.shinyapps.io/ach-gap-variability-shiny/>