Research Report

Diversity Associated With Reductions in the Achievement Gap:

Data Mining the 2010–2011 New Jersey Assessment of Skills and Knowledge (NJ ASK)

Stephen José Hanson, Stephanie Saunders,
Arcelio Aponte, Robert Copeland, and Michael Nettles

POLICY INFORMATION CENTER

This Policy Information Report was written by:

Stephen José Hanson, Stephanie Saunders, Arcelio Aponte, Robert Copeland, & Michael Nettles

Policy Information Center
Mail Stop 19-R
Educational Testing Service
Rosedale Road Princeton, NJ 08541-0001
(609) 734-5212
pic@ets.org

Copies can be downloaded from:www.ets.org/research/pic

About ETS

At ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research. ETS serves individuals, educational institutions and government agencies by providing customized solutions for teacher certification, English language learning, and elementary, secondary and postsecondary education, and by conducting education research, analysis and policy studies. Founded as a nonprofit in 1947, ETS develops, administers and scores more than 50 million tests annually — including the *TOEFL*® and *TOEIC*® tests, the *GRE*® tests and *The Praxis Series*® assessments — in more than 180 countries, at over 9,000 locations worldwide.

RESEARCH REPORT

Diversity Associated With Reductions in the Achievement Gap: Data Mining the 2010–2011 New Jersey Assessment of Skills and Knowledge (NJ ASK)

Stephen José Hanson,[1,2] Stephanie Saunders,[3] Arcelio Aponte,[4] Robert Copeland,[5] & Michael Nettles[3]

1 Psychology Department, Rutgers University, Newark, NJ
2 Rutgers University, New Brunswick, NJ
3 Policy Evaluation Research Center, Educational Testing Service, Princeton, NJ
4 New Jersey State Board of Education, Trenton, NJ
5 Lower Merion School District, Ardmore, PA

The academic achievement gap is a persistent and pernicious educational challenge confounded with race and socioeconomic status. The achievement gap persists despite over three decades of interventions and federal, state, and local policies and initiatives meant to close it. We examine the achievement gap in one of the most diverse states in the United States, New Jersey. Across the 480 school districts in New Jersey, the school district racial makeup varies from almost complete segregation to high diversity. Using data-mining techniques on the statewide standardized test, the 2010–2011 New Jersey Assessment of Skills and Knowledge (NJ ASK) standard language arts literacy and mathematics, we found that test scores increased with increasing diversity. In particular, the achievement gap between Black and White third grade students was lower by more than 60% in racially diverse districts when compared to racially homogeneous districts. This result is consistent with theories of peer spillover effects that implicate peer racial diversity in reducing the achievement gap. A cluster analysis of New Jersey school districts revealed eight common profiles, which could inform school policy and practice aimed at improving school performance and closing the achievement gap.

Keywords Achievement gap; diversity; school reforms; data mining; big data

doi:10.1002/ets2.12292

The achievement gap is one of the great economic and educational challenges that the United States faces in the 21st century. Despite decades of federal and state interventions, changes in policies, and increasing resources for failing school districts, the achievement gap continues to persist (American Institutes for Research, 2013). In this paper, we introduce a diversity hypothesis that results from student mixture distributions and their dynamics (Fruehwirth, 2013). In particular, the mixing dynamics of various student peer groups may produce achievement improvements that are associated with a smaller achievement gap when students from different racial groups interact within diverse school districts.

However, there are observed differential achievement gaps based on the distribution of the overall scores. In particular, smaller achievement gaps appear to be more prominent within school districts with standardized test scores at the lower and mid-end of the score distributions (Fruehwirth, 2013). At low to midlevel score distributions, modeled achievement gap reductions based on peer spillover effects can be substantial, but at higher scores the estimated effects tend to be small to nonexistent. Thus, from a policy perspective, the diversity hypothesis is best examined in the context of those school districts that are neither in the floor of the score distribution (due to persistently poor performance) nor in the ceiling (related to socioeconomic advantage or selective segregation) but rather in the larger bulk of districts with student scores in the midlevel of the distribution, which are theorized to have larger increases or decreases in score tendency. In this paper, we use data-mining techniques to explore these effects in a large data set from 480 school districts and over 600,000 students in New Jersey. We tested the implications of school district racial diversity on the achievement gap in one particular year (2010) and how this effect varies across school district mean-level achievement.

Defining the Achievement Gap

The achievement gap at the most basic level is a statistically significant difference on standardized scores between races or social-economic groups or even between second- and first-language learners (Vanneman et al., 2009). These differences

Corresponding author: S. J. Hanson, E-mail: jose@rubic.rutgers.edu

can be quite large, as much as 2 – 3 standard deviations, leaving students as much as 4 years behind in academic reading. These effects for the individuals persist from an early age to high school and beyond, affecting dropout rates, long-term college graduation, and lifetime earnings (Noguera et al., 2015).

Closing the Achievement Gap

There has been considerable investment in interventions and targeted programs to try to close the achievement gap (Ferguson et al., 2009). Unfortunately, specific interventions tend not to generalize over time or are just ineffective. For example, there has been considerable focus on preschool reforms that aim to increase the percentage of disadvantaged children provided preschool education (Barnett, 2011; Cascio & Schanzenbach, 2014). On the face of it, these sets of reforms should likely be effective in equilibrating starting points for all children, thus encouraging the same educational trajectory over time and subsequently reducing the achievement gap. Unfortunately, almost all the research on preschool programs tends to show early gains in achievement, but these early gains tend to diminish over time (Puma et al., 2012). The inability of preschool reforms to reduce the achievement gap suggests strong environmental (peer group), family, or other uncontrolled factors that are more dominating (Dearing et al., 2006) compared to potential positive effects of preschool interventions.

There has also been a considerable and long-term focus on standards-based reforms, which define and assess students' skills and competencies with the assumption that higher standards are a core component of school improvement and student success (Silver, 2004). Standards-based reform has been with us for 20 years — first at the state level and then at the federal level in the form of the No Child Left Behind Act of 2002 and the Every Student Succeeds Act. It has been argued for nearly a decade (Hamilton et al., 2008) that we would be seeing sustained improvements, but they have just not been generally apparent.

Teacher reform is a common-sense approach that has the potential to reduce the achievement gap. Reform efforts aim to improve teachers' classroom skills and the quality of teacher preparation. Education research has shown the critical role of teacher quality in student performance, while at the same time, evidence has also mounted that teachers with less experience and lower qualifications are assigned to school districts with high proportions of students of color (Peske & Haycock, 2006). Thus, a potential confound is created with the presence of the achievement gap in a given school district. Although it is clear that improving teacher quality can help reduce the achievement gap (Harris & Sass, 2011), there seems to be little evidence that more global, systemic interventions compared with specific targeted interventions per school can or could eventually close the achievement gap altogether (Hamilton et al., 2008).

In this paper, we present evidence that school district diversity can have a dramatic and progressive relationship with the racial achievement gap. First, we provide an overview of the specific theoretical mechanism that may explain the diversity effect, and then we provide justification for the focus on New Jersey as a state with an extremely diverse set of school districts. This is followed by analysis and discussion of our evidence in support of the diversity hypothesis.

A Diversity Hypothesis

There has been a steady examination for many decades of whether desegregation would narrow the achievement gap (Guryan, 2004; Tumin, 1959). Many groups have looked at the effects of segregation on test scores and unfortunately have found mixed results (Johnson, 2011). In particular, Card and Rothstein (2007) provided evidence using specific controls for the family background of individual test takers, school-level controls for selective participation in the test, and city-level controls for racial composition, income, and region and have shown that the Black – White test score gap was higher in more segregated cities. Other research has shown that other sources of the achievement gap include family environment (Lee & Bowen, 2006), school resources (Ladson-Billings, 2006), and youth health (Basch, 2011). However, models on peer effects imply that the racial composition of students' peers (Vigdor & Nechyba, 2007) and the achievement of their same-race peers (Sacerdote, 2011) can influence student achievement and test scores. In particular, student utility models have been shown to mediate peer effects (Cooley, 2010; Fruehwirth, 2013, 2014). Fruehwirth argued that peer "spillover" affects selective peer group pressure, in which the presence of academically high-achieving students has an effect on low-achieving peers by increasing their performance and, in an asymmetric way, showing no deficit or effect on performance in high-achieving peers, resulting in a net reduction in achievement gap. To further clarify spillover effects, Fruehwirth considered peer effects in which one peer group (e.g., the lower end of the achievement gap) interacts with another is likely to produce interactions that have a positive effect on the low-achieving peer group. Diversity alone, through peer mixing, can provide opportunities otherwise not available in peer ecologies that are highly segregated and generally not linked to positive achievement outcomes.

Assuming that various exogenous factors (e.g., poverty, violence) tend to differentially affect Black students, particularly in large urban environments (Patton et al., 2012), we might expect an increase in diversity or race mixtures to therefore create positive peer spillover effects that, even when offset with the absence of negative peer spillover effects, produce a net reduction in the achievement gap. We expect that this would not be due merely to decreasing all groups toward a floor or bottom of performance. Reardon (2016) found that across various bivariate and multivariate models, the single most powerful pattern correlated with Black–White achievement gaps was the disparity in average school poverty rates between Black and White students. However, Reardon found that in urban areas where school racial segregation was higher than predicted from racial disparities in socioeconomic conditions, average achievement gaps were significantly larger, which is suggestive, though not definitively, of a causal link between segregation and achievement gaps.

The question we pursue here in the context of the peer effects spillover theory—a large, diverse academic ecology—is this: To what extent would these hypothesized spillover effects be implicated in smaller observed achievement gaps? Would there be differential changes in the achievement gap based on school district similarity or cluster structure? Policy strategies can arise from the individual differences between clusters that may be identified, thus concentrating the policy impact that may not fit all school districts, regardless of geography, property taxes, demographics, or race.

Why New Jersey?

Due to demographic shifts, the diversity of school-aged children has increased over the last several decades throughout the United States and in New Jersey in particular (Orfield et al., 2017). According to the 2010 census, New Jersey has the nation's fourth largest Asian population, seventh largest Hispanic population, and at 40.7%, a substantially above-average state proportion of minority population (Wu, 2011).Thus, we chose this highly diverse state as a laboratory to examine diversity and its relationship to the achievement gap.

Interestingly, the achievement gap across the United States has little to do with geographic location (McKinsey & Company, 2009). Neighboring states can diverge widely on achievement gaps; for example, consider Oklahoma and Arkansas, where Oklahoma has half the achievement gap of Arkansas. Similarly, we can compare Connecticut and New Hampshire, where Connecticut has an achievement gap three times higher than New Hampshire. Note that even the states with comparatively high overall test scores (i.e., Massachusetts) can still possess a very high achievement gap. New Jersey students on average have high overall national test scores (National Assessment of Educational Progress) but a significant and average-sized achievement gap when compared with other states (McKinsey & Company, 2009). Throughout New Jersey, there is a highly complex set of school districts that range from very high diversity to those that are almost completely segregated, causing various researchers to describe the New Jersey school system as having a kind of educational "apartheid" (Tractenberg et al., 2013). Between 1989 and 2015, the portion of New Jersey schools serving a majority non-White population increased from 22% to 46%, while the portion of intensively segregated schools also nearly doubled (Orfield et al., 2017). These extremes will, in fact, be instrumental in examining the diversity hypothesis and test score variations, as we can compare pairs of school districts matched overall in relevant variables while examining the effects of diversity on a single variable.

New Jersey also has a range of school district "ecologies," including urban, suburban, and rural school districts as well as those along the shore, inland, or near large urban centers such as New York City or Philadelphia. Consequently, we believe New Jersey could act as a microcosm of the United States, in terms of the school districts throughout the state, and has therefore a potential for the greatest generalization as we examine this diversity as it relates to student performance. Finally, the large data set size should enable detection of this potential diversity effect due to increase in power, or signal to noise ratio, thus enhancing effects that may be missed in smaller or more planned studies, ignoring what might be considered only nuisance variables.

Data Mining Versus Planned Experimental Design

Some data collection and archiving are created with some unknown hypotheses and the hope that future data analysis might be able to leverage effects in the data that were never planned for or expected. Modern data science has been partly responsible for the new artificial intelligence. For example, in their work on deep learning, LeCun et al. (2015) started with the proposition that there are covariates in large multivariate data sets that can be "mined," as in a form of a discovery of effects that might not have otherwise been considered or thought likely to be found. The fundamental principle is that a search in a large data space is often more likely to find structure than engineering or design.

In the present case we are focused on a large-scale data set of individuals with fewer than 10 variables of interest. Although relatively small compared to other large archives used in many data science domains that include 1 million to 100 million entities (Wickham & Grolemund, 2017), there are still $(10 \times 9)/2 = 45$ possible relations. We are focused on discovering one identified variable (diversity, to be defined later) and a properly sampled achievement gap as a robust t-test in sampled school districts. Once the sample is selected (say, from theoretical considerations), it is important to establish some descriptive aspects of the selected variables, first in terms of their relative frequency distributions and their shape and possible source probability density function. This allows some initial observations on the nature of random variables underlying the distribution or mixture of effects creating the observed distribution. Latent structure modeling, principal component analysis (PCA), factor analysis, multidimensional scaling (MDS), canonical correlations, and so on can be used to establish new derived variables that are lower dimensional and higher in variability and signal. Aggregation can first help define entity (e.g., subject) clusters (agglomerative hierarchical clustering), thus allowing latent variables to have higher impact per cluster if diagnostic. Multiple regression and linear discriminate analysis can be used to establish relationships between aggregated and condensed variables, providing a novel look at relationships among the derived variables that might not have been discovered or analyzed previously. In the next section, we provide details on the New Jersey Assessment of Skills and Knowledge (NJ ASK) sample and its descriptive structure.

Participation and Testing

The NJ ASK 2010 data were collected over a 2-day period in Spring 2010. Measures of mathematics and language arts literacy (LAL) grade-level proficiency were collected from all Grade 3–8 students in all school districts in the state. The assessment was designed to provide information about each student's achievement in the areas required by New Jersey's Core Curriculum Content Standards (CCCS). The test serves as an indicator to ensure that local instructional programs are aligned to those content standards and that students are mastering the knowledge and skills required by the end of each grade (New Jersey Department of Education, 2011). In 2015, the NJ ASK was replaced by the national Partnership for Assessment of Readiness for College and Careers (PARCC) assessment.

Fundamentally, the tests focused on two apparently independent measures. The first involved an LAL assessment, which included reading passages, multiple-choice questions, and writing. The second measure involved a mathematics assessment including multiple-choice questions and extended construction questions without use of a calculator.

All testing results and materials were provided by the New Jersey Department of Education in raw form. The tests were constructed by the New Jersey Department of Education and validated and shown to be highly reliable since 2004 (New Jersey Department of Education, 2004). State regulations (N.J.A.C. 6A8-2.1(a)5i) stipulate that the CCCS must be reviewed for possible revision every 5 years. Thus, the CCCS constitutes a dynamic entity, not a fixed, final set of standards (New Jersey Department of Education, 2016). Similarly, New Jersey's assessments reflect continuous refinements and evolving understandings of the CCCS while using assessment instruments that are highly standardized for the purposes of ensuring validity, reliability, and comparability. Cronbach's alpha for the 2010–2011 NJ ASK for overall student responses ranged from 0.82 to 0.91 for LAL, 0.90 to 0.92 for mathematics, and 0.84 to 0.90 for science, indicating that the tests are highly reliable (New Jersey Department of Education, 2011).

Variables in the basic testing include raw scale scores for LAL and mathematics scales, with adjusted and rescaled scores, aimed at creating comparability across years in which testing criteria may have changed. Normalization transformations can allow for comparability, given that the transformation provides probabilistic or scale-free measures, as in z-scores, or the use of principal components, as done in this analysis. Validity is a convergent process between NJ ASK tests and other known tests that provide for individual independent scores in similar tasks or skill sets that the state of New Jersey routinely tests.

The comprehensive data sets provided by the New Jersey Department of Education included 220 variables and roughly 100,000 students per grade. In total, from Grades 3–8, we had data for approximately 600,000 students. In this data set, more than half (53%) of the students were male across all six grades. Within the 220 variables, there were descriptors, school district codes, and demographics such as race (i.e., Black, Hispanic, White, Asian, Native Hawaiian, American

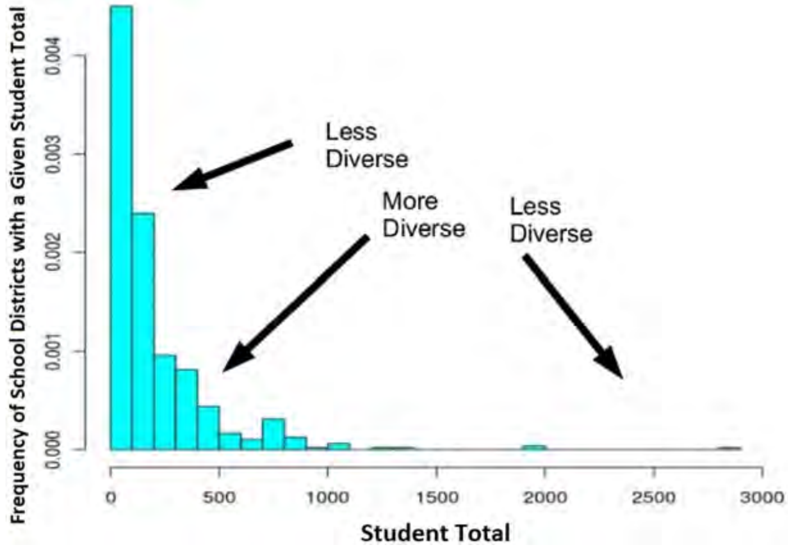

Figure 1 Relative frequencies of school districts with given student totals.

Indian),[1] gender, and socioeconomic category (at the school district level). Many of the variables included the variations on transformations of raw scores to reportable scores. The proportions of race over the whole student and grades sample broke down as follows: White, 55%; Black, 16%; Hispanic, 19%; Asian, 8%.

In this data set were 480 school districts (out of 590) with variable school district size from fewer than 10 to thousands of students per district. Districts were excluded from the analysis for two reasons: (a) missing data or (b) district size below threshold sampling entity (where school district size was fewer than 10 students). These data aggregate over more than 2,500 schools in the state of New Jersey.

Relative socioeconomic status (SES) at the school district level was indicated on a scaled categorical economic designation where district factor group codes (FGs) were assigned based on U.S. Census data from low to high in eight category steps. With this, we constructed a socioeconomic variable that was weighted by student variables. School district size (or the number of students in a school district, NSD) was also added to the set of variables for the present analysis and data exploration due to the huge variation in size of over 10 standard deviations (for orders of magnitude, see Figure 1). Interestingly, we discovered that the smallest and the largest school districts tended to have the least diverse student bodies.

Data Mining: Exploring the Data With Minimal Statistical Assumptions

Data mining is an exploratory data analysis approach aimed at discovering novel and potentially useful information from large amounts of data. According to Baker (2010), mining data from thousands of students with broadly similar learning experiences (such as being in the same grade within a state with common learning standards but in very different contexts) allows leverage that was never before possible in studying the influence of contextual factors on learning and learners. Data mining attempts to provide low-impact data analyses, ones with a small number of assumptions, especially concerning the underlying distributions (Gaussian) or potential models (e.g., linear) appropriate for the raw variables (Wickham & Grolemund, 2017). This allows the methods to reveal underlying structure without creating potential distortions of the raw data by constructing composite or highly preprocessed scores and using those as substitutes for the raw data.

We therefore started with descriptive analyses to examine the raw data distributional forms (e.g., relative frequency distributions). This allowed us to understand whether the scores per school district have functional relationships to the policy and contextual forces that now exist. Next, we examined the clustering of school districts using nonparametric clustering both to test the homogeneity of New Jersey school districts and to systematically control for some variables. The purpose of the cluster analysis was twofold: (a) Do school districts cluster together with similar types of challenges, and therefore, could the existing achievement gap be affected by similar policy? and (b) How can we characterize the individual differences that emerge from a cluster identification that might allow us to focus on smaller clusters of school districts that could be treated and viewed analytically in a similar way?

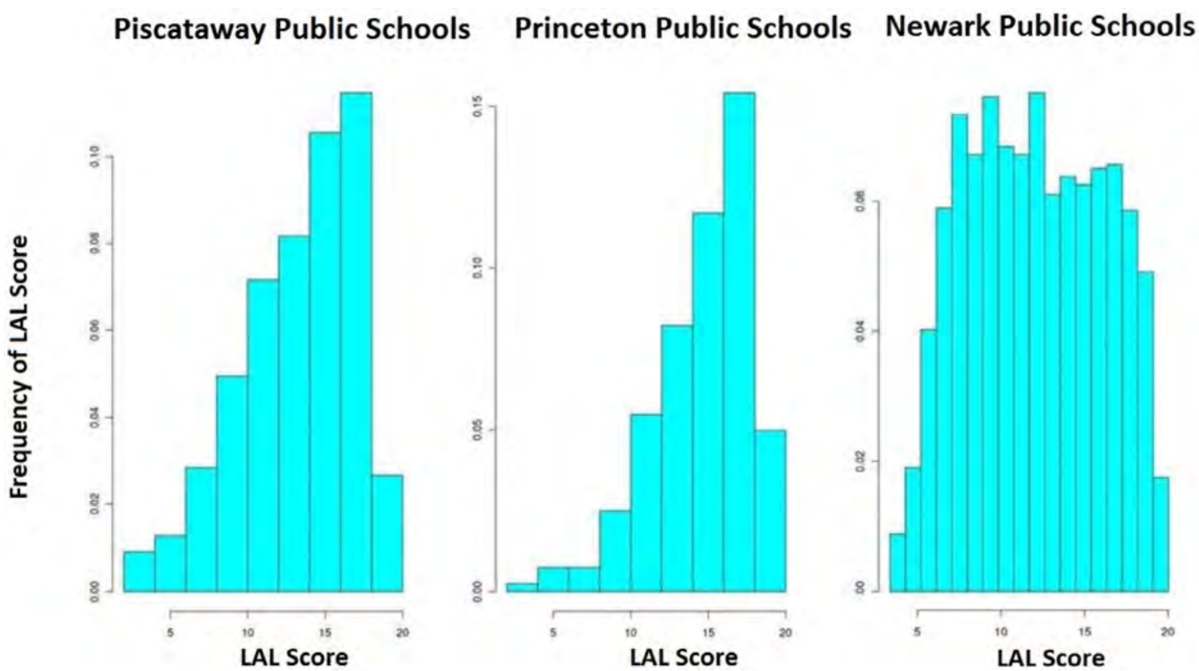

Figure 2 Relative frequency distribution of students' language arts literacy scores by three districts.

The Nature of the Frequency Distributions of Scores

Raw scores were extracted and examined for distributional properties (relative frequency distributions). First, we constructed the relative frequency distributions over all scores and all students per grade. This produced a "shape" profile for interpretation. If the distribution shows a flat profile, it means that the scores are fairly random, much as if the student got to spin a "score lottery wheel" and received a score randomly (see Figure 2, Newark distribution). This suggests that there is little academic or management pressure to reduce poor scores from students or to increase student's ability to achieve higher scores. A second kind of distributional outcome shows scores biased toward the higher end (*negative skew*) and suggests that students are under some "ecological pressure" to improve, either from local policy, teaching, or management in the school district itself (although other endogenous factors may be at play). The more extreme slope from left to right is indicative of increased pressure (see Figure 2, from Piscataway to Princeton).

Many schools, particularly those designated as Abbott—high-poverty districts with New Jersey court-ordered increased state funding for potential equalization (Sciarra & Hunter, 2015)—were in fact almost flat and what could be characterized as a uniform type (see Figure 3), which could indicate a random mixing of score outcomes (despite significant resources put into these high poverty districts). The negative skew of the non-Abbott districts, on the other hand, may indicate that some organized structure at the school district contributes to score-increase pressure, such as local leadership, policies, or other unknown peer factors.

LAL and mathematics scores were examined for correlation in order to determine test independence for further analysis. As shown, scores were correlated at Grade 3 at $r = .87$, indicating that more than 75% of the variance was in common between the two measures of subject matter proficiency. This correlation weakens over grade progression and by Grade 8 is .78. Nonetheless, because of the high correlation, the variation in one is reflected strongly in the other. Consequently, it would be acceptable to treat these two scores as one independent score, much like applying Roy's largest root, first principal component when finding a correlation between height and weight and recoding the corresponding first principal component as size (Hanson et al., 2009).

Heterogeneity and School District Clusters

For further mining analysis and to establish properties of the samples that provided an underlying structure, we assessed (a) the homogeneity of the scores over the student samples and (b) whether it is possible to analyze these groups for

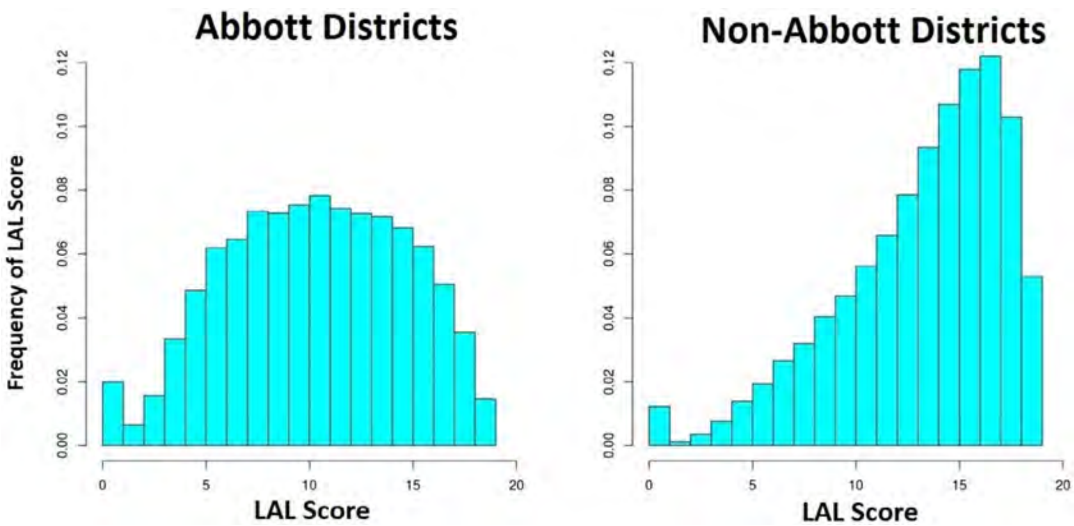

Figure 3 Relative frequency distribution of language arts literacy scores by Abbott and non-Abbott school districts.

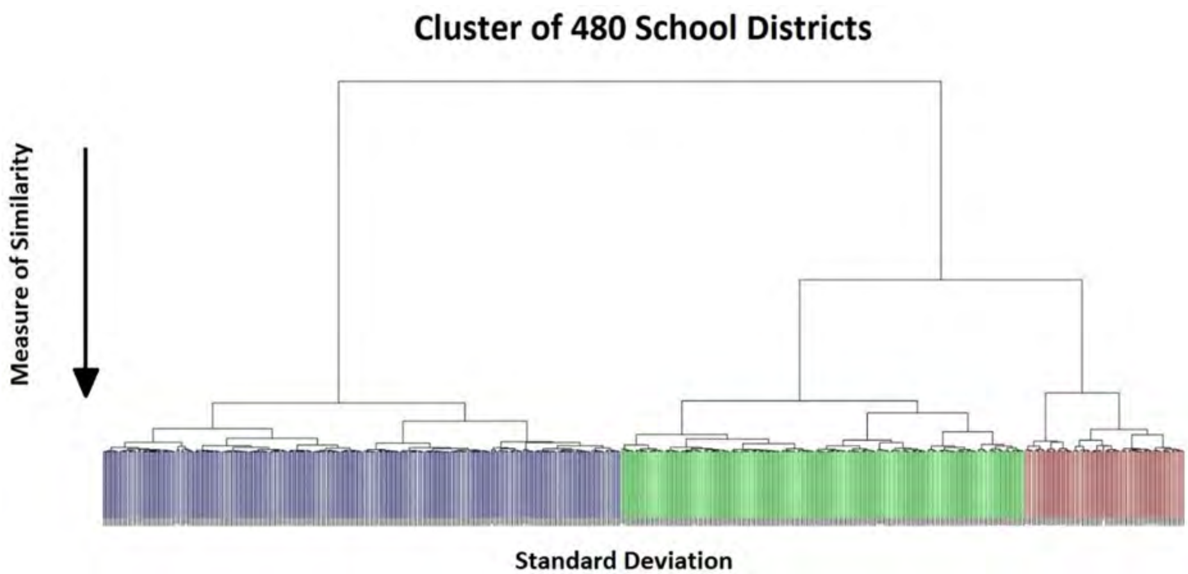

Figure 4 New Jersey school district clustering.

differences due to cluster membership policy strategies. School districts were clustered using hierarchical agglomerative clustering with Ward's grouping rule (Ward, 1963) and Euclidean distance metric over seven extracted variables from the data set. The variables in the cluster analysis included raw LAL scores, raw mathematics scores (even though correlated, their variability might vary by cluster), and racial proportion out of total students with a school district (Note that our exploratory analysis defines racial diversity mathematically as a three-dimensional distribution of White, Black, and Hispanic students and does not account for other student racial populations). We used three modal groups—Black, White, Hispanic (and bias)—to indicate a deviation from equal proportions of the three racial categories.

The coarse-level clustering in Figure 4 shows the major cluster groups that strongly appear to have three clusters (cut the tree at the three-cluster level shown in the figure by a horizontal line) indicated by colors red, green, and purple, showing which school districts are in which cluster. The dendrogram shows cuts that may hierarchically contain other clusters, possibly as many as 10. The number of clusters is determined from a scree plot, which shows the merge history with the cluster similarity, significant jumps in the dendrogram, and cut opportunities. This analysis indicated three larger groups and 10–12 clusters.

The initial analysis confirmed that New Jersey school districts are highly segregated. Analysis of the three cluster groups suggested groups with either (a) high White bias, (b) high Black bias, or (c) districts with higher diversity, where the race dominance was considerably less, although still present at some level.

We assessed the achievement gap between White and Black students within these initial clusters, and results indicated a significantly smaller score gap within the more diverse school districts. Hence, the diversity of school districts appears to be related to the achievement gap. We next tested that hypothesis more directly by explicitly looking for clusters based on diversity and exploring the possibility of a relationship to the achievement gap. Could different clusters have different achievement gaps?

Diversity as a Factor Established

This gap – diversity trend was tested by taking those school districts with at least 100 Black students and 100 White students and doing a statistical test on the difference between the distribution of scores between the school districts as a function of their diversity. We statistically estimated the achievement gap using a Welch's *df* corrected t-test (Welch, 1947), which with larger school district samples, converges to a z-estimate (standard scores; Student, 1908) for all such school districts ($n = 168$ or $\sim 40\%$ of the school districts survived the sample size restriction of at least 100 White and 100 Black students in the same sample). This selection requirement was necessary to have stability in the samples; fewer than 100 students per group produced similar regression values but with higher sampling error as the minimum dropped below 50 students per group. Consequently, 100 was chosen as it still provided a significant number of school districts and schools (greater than 1,200) and had no obvious biasing effects in the analysis.

Diversity was assessed on the modal racial designations (White, Black, Hispanic) using a Shannon entropy diversity measure (Shannon, 1948) based on the proportions of those race biases per school district.

$$-\sum \left\{ \left[\mathrm{p}\,(\mathrm{i})\,\mathrm{Log}\left[1/\mathrm{p}\,(\mathrm{i})\right] \right\} \right.$$

Shannon entropy is a robust measure of the flatness of a distribution; in this case we were interested in the amount of segregation in a school district as measured by the proportion of student types (White, Black, Hispanic) in the school district. These Shannon entropy values varied from 0 to 1, where 0 means no diversity and 1 indicated complete diversity (flat distributions, e.g., .33, .33, .33). The Welch's t-test gap was then regressed against the school diversity to see if a potential causal relation could be detected ("causal" is used in the sense of conditional probability, e.g., $\mathrm{P}(\mathrm{x} \mid \mathrm{y}) > \mathrm{P}(\mathrm{y} \mid \mathrm{x})$).

In Figure 5, we show that there is a very strong relationship ($r = -.51$, $p < .004$) of the third grade Black – White achievement gap on the school district diversity. In fact, for every 10% increase in diversity in this sample, there appears to be a decrease in the achievement gap by 3 points (scaled from 0 to 15, or about 20% change). The effect is stable over the next five grades (up to eighth). We should note that this observation does not imply a causal relationship; rather, it is only consistent with the theoretical framework provided by Fruehwirth (2013), the peer mixing effects models. For example, within the mid-diversity range, there were increasing achievement gap differences that had been predicted by this theory.

In a similar but more simplified analysis, we found that the third grade Black – White achievement gap in LAL scores decreased with increased White, Black, and Hispanic student diversity at the school district level (see Figure 6). School districts were categorized into high, medium, and low diversity based on an analysis of their White, Black, and Hispanic Shannon entropy.

The Clusters Re-Examined: How Many Are There?

Returning to the underlying clusters, we decided it would be instructive to data mine this cluster space a bit more deeply to see if there were common profiles and to determine whether this observed diversity effect might support an alternative approach in characterizing New Jersey school districts and the underlying factors that may be influencing the achievement gap. Again, our analysis suggested very strongly that New Jersey school districts are not homogeneous and that the underlying structure clearly has multiple groups independent of geographic location.

Cluster Profiles

We constructed a divergence difference measure over the cluster merge history (sequential differences of merge scores, used to indicate the most stable cluster number in the search), which we found to be eight. This finding was subsequently

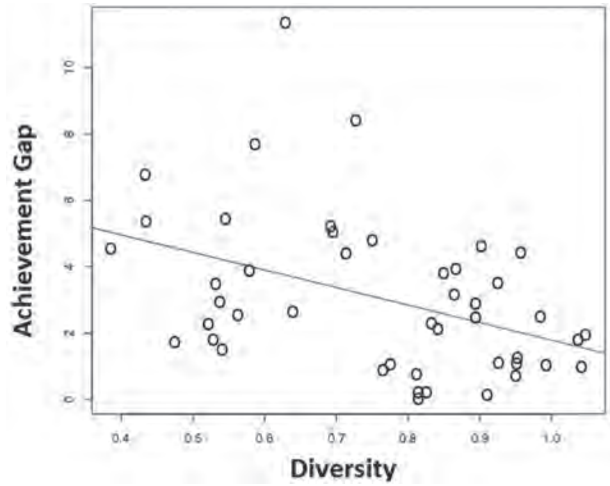

Figure 5 Third grade Black–White achievement gap as related to school district diversity.

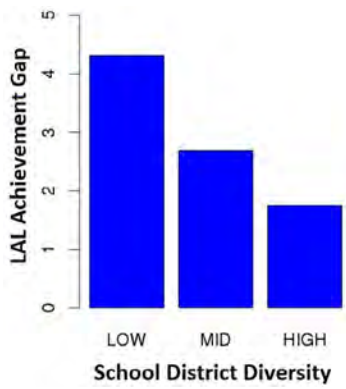

Figure 6 Third grade Black–White language arts literacy score achievement gap by school district diversity.

confirmed with a Fisher's linear discriminant analysis, producing significant F-values ($F = 22$, $p < .001$). Next, we employed a simple and direct way to reveal the cluster properties: We extracted the mean profiles over the seven variables on which the clusters were identified (LAL score, mathematics score, White bias, Black bias, Hispanic bias, SES, and NSD).

In Figure 7, we show the mean profiles for the eight clusters. Each variable (LAL, math, White bias, Black bias, Hispanic bias, SES, NSD) is shown in order on the x-axis, and the mean value is shown in standard score units (z-score, mean centered, and standard deviation scaled), meaning that 0 is at the mean value and + or −1 is one standard deviation above or below the mean for any variable, making them potentially comparable (see Appendix A).

We scaled the data within each cluster to force prototypes, which are the school districts that fall within the central space of the cluster. For a list of prototypical school districts within each cluster, see Appendix B. Prototypical school districts have profiles that approximate the cluster mean values on all the dimensions used to derive them, as seen in Figure 7. One prototypical school district will be used as an example for each cluster description.

The first observation is that the cluster mean profiles are quite diverse. Going from left to right in the top left corner is Cluster 1, which consists of average-sized school districts with above-average achievement test scores, high SES, and a more White and less Black or Hispanic student population. Princeton Public Schools is a prototypical example of a Cluster 1 district, a school district that is among the best in the state and the nation (U.S. News, 2019), where high academic achievement for all students is an expectation (Princeton Public Schools, 2018). A majority of Princeton Public School students are White (55%) or Asian (20%) and not economically disadvantaged (87%). In AY 2016–2017, 76% of the district's students met or exceeded expectations in LAL and 63% did so in mathematics (New Jersey Department of Education, 2018g). Other school district statistics can also be found in the New Jersey Department of Education public documents.

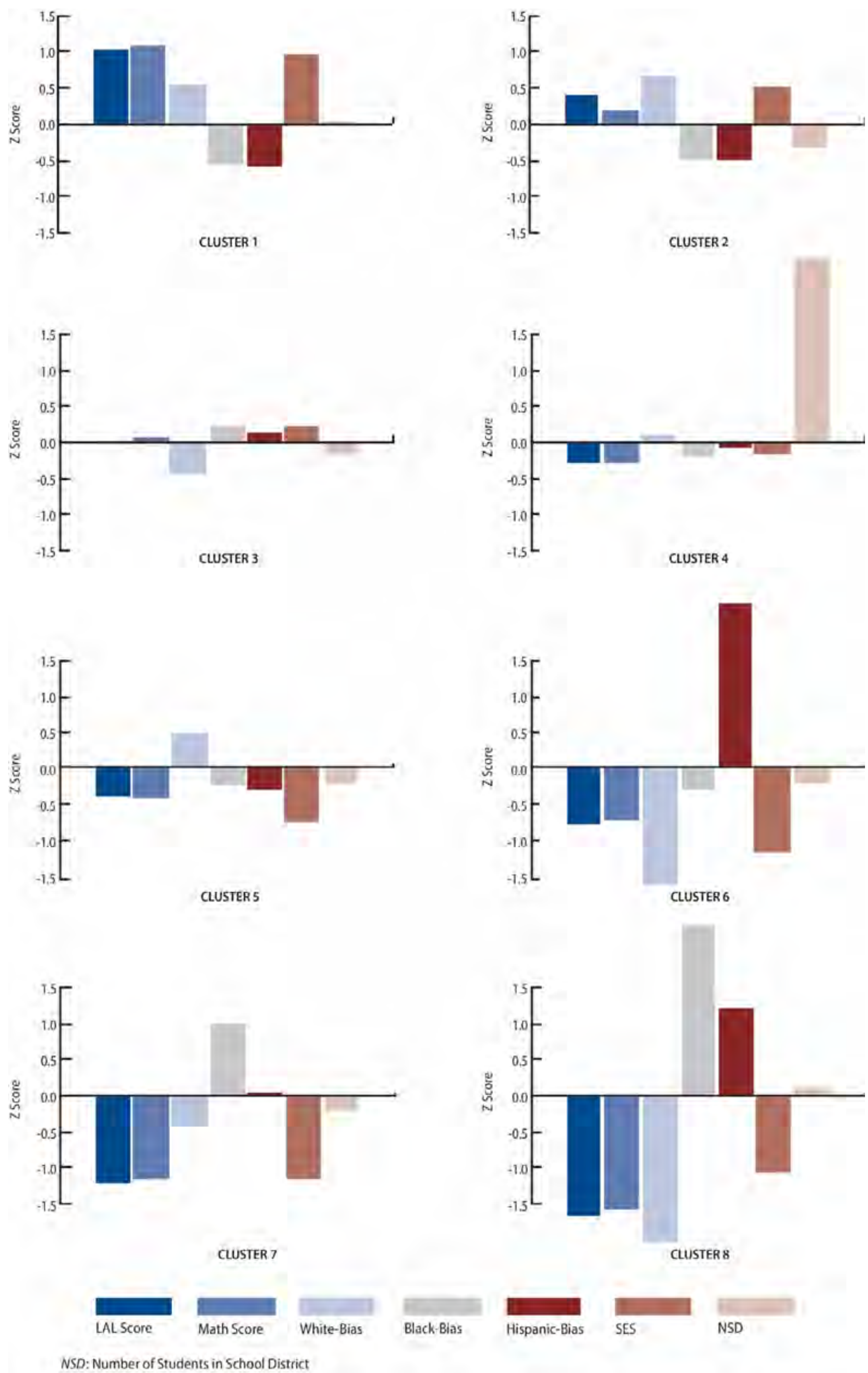

Figure 7 Profiles for the eight New Jersey school district clusters.

The next cluster, Cluster 2, has a strong similarity to Cluster 1; however, the test scores and average SES status of those in the district are slightly less above average. Also, although having a comparatively similar racial/ethnic profile, these school districts tend to be somewhat smaller in size. East Hanover Township School District is a prototypical Cluster 2 district. Seventy-five percent of East Hanover students are White, but only 3% are economically disadvantaged. At the district level, in academic year (AY) 2016–2017, 68% of students met or exceeded expectations in LAL and 54% did so in mathematics (New Jersey Department of Education, 2018a).

Cluster 3 is the only cluster of school districts without a White bias in the student populations and with an average to slightly above average LAL and mathematics scores. The districts in this cluster are on average slightly smaller in size and have more Black and Hispanic students and families with slightly above average SES. Piscataway Township Schools is a prototypical Cluster 4 district, with a diverse and above average SES student population. The Piscataway Township Schools have a diverse student population: 33% Asian, 28% Black, 19% Hispanic, and 16% White; a third of the student population in the district is economically disadvantaged. At the district level, in AY 2016–2017, 59% of students met or exceeded expectations in LAL and 50% did so in mathematics (New Jersey Department of Education, 2018f).

These first three clusters collectively represent the school districts in New Jersey, which serve students with average to above average SES and above average standardized test scores. Cluster 4 is the first of the cluster of districts with below average test scores, though the scores are <.5 standard deviations below average. This cluster consists of very large districts, with mostly diverse student populations, which are slightly Black and Hispanic biased, and slightly below average in SES. The Township of Union Public Schools is a prototypical Cluster 4 district. The district has a diverse student population: 43% Black, 24% Hispanic, 21% White, and 10% Asian; just over a third of the students in the district are economically disadvantaged. At the district level, in AY 2016–2017, 52% of students met expectations in LAL and 39% did so in mathematics (New Jersey Department of Education, 2018h).

Cluster 5 is the only cluster of school districts with White-biased student populations and below-average test scores. These districts tend to be well below average in SES and smaller in size. Mansfield Township School District is a prototypical Cluster 5 district, which serves only the elementary grades (pre–K to 6). A majority of Mansfield Township School District students are White (65%), and just over 30% are economically disadvantaged. A majority of the district's Grade 3 students did not meet academic expectations in AY 2016–2017; only 35% did so in LAL and 46% in mathematics (New Jersey Department of Education, 2018d).

Cluster 6 consists of school districts that are strongly Hispanic biased and lower in SES, with below average test scores. New Brunswick Public Schools is a prototypical Cluster 6 district. New Brunswick Public School students are 89% Hispanic and 9% Black; 86% are economically disadvantaged. At the district level, in AY 2016–2017, only 26% of students met expectations in LAL and 19% did so in mathematics (New Jersey Department of Education, 2018e).

Cluster 7 consists of school districts with Black-biased and low SES student populations and well-below-average LAL and mathematics scores (over one standard deviation below average). Ewing Public Schools is a prototypical Cluster 7 district. Forty-six percent of Ewing Public School students are Black, and 44% are economically disadvantaged. In AY 2016–2017, 42% of students met or exceeded expectations in LAL and 34% of students did so for mathematics (New Jersey Department of Education, 2018c).

Finally, we have Cluster 8, which consists of slightly larger school districts that serve mostly low SES and Black or Hispanic students. Students in this cluster score 1.5 standard deviations below average on the NJ ASK in LAL and mathematics. East Orange Public Schools is a prototypical Cluster 8 district. The district publicly embraces the challenge of providing a quality education to its students and is committed to developing programs that meet the diverse academic and social needs of its student body (East Orange Public Schools, 2018). Over 90% of the East Orange District student body is Black, and nearly 80% is economically disadvantaged. In AY 2016–2017, 36% of Grade 3 students and 32% of Grade 8 students met or exceeded expectations in LAL whereas only 32% and 12%, respectively, did so in mathematics (New Jersey Department of Education, 2018b).

Potential Diversity Confounds

The achievement gap has been shown to be related to school district diversity. However, a number of possible variables might be explanatory of this effect (i.e., both the diversity and segregation correlated with urban settings). In the United States, metropolitan areas are increasingly marked by both racial segregation and racial diversity (Holloway et al., 2012). As we saw in our cluster profiles, diversity is low in some urban districts, whereas it is relatively high in others.

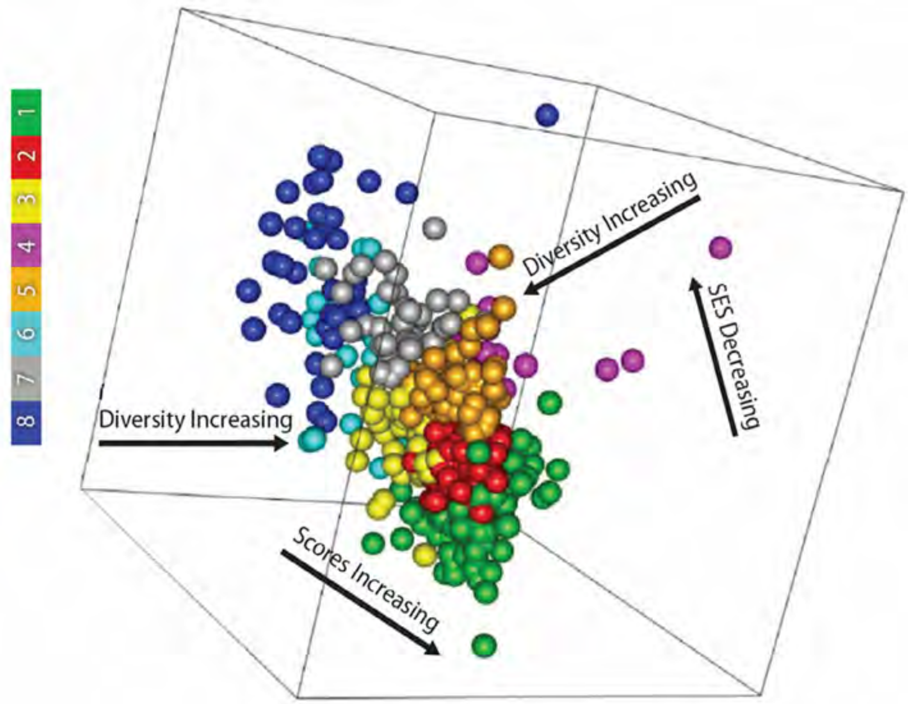

Figure 8 Multidimensional scaling of New Jersey school district clusters.

Socioeconomic variables might also be related to diversity; however, in our analysis, higher socioeconomic variables in the New Jersey samples are correlated with lower diversity. This finding is also true of school districts with the lowest SES, which are highly segregated with some of the lowest diversities. In fact, school districts with midlevel SES tend to have higher diversity and smaller achievement gap outcomes. These are the districts where we see the largest decrease in the achievement gap as diversity is greater. It also should be clear that teacher quality is more likely to be both very high and very low in school districts with very low racial entropy, thus controlling for the potential variation in teacher quality where a modal Black school district may have considerably lower teacher quality than a modal White school district (Peske & Haycock, 2006).

We examined other possible confounds throughout the larger sample by creating critical pairs of districts in order to control for a number of variables, including SES, district size, and geographic locale (e.g., urban, suburban, shore). Consistent with the peer effects models (Fruehwirth, 2013, 2014), the greatest amelioration of the achievement gap was in midsized school districts; however, significantly smaller achievement gaps occurred throughout the spectrum of district size as well as geographic locale but less so in the very high or very low socioeconomic school districts.

Structure of School Districts through New Jersey Dynamic Effects

In order to visualize the complex structure of the school districts, a nonmetric MDS was performed on all variables showing a diversity volume moderated by socioeconomic class (see Figure 8).

The original scores for cluster membership were scaled into three derived (but unknown) variables in order to show the relationship between the cluster members. This illustrated perspective roughly shows the effect of diversity, which is most highly centered at the middle of the cube and decreases as it expands out from the center. Going from the back to the front of the space is a measure of scores increasing and decreasing (green/blue). SES, as noted earlier, has a correlation with scores and clusters, and in this figure we can see that as SES decreases, the diversity of the cluster itself increases (as seen by a larger cluster spread). The figure indicates an increase in cluster variation as school districts perform more poorly, while the other end of the structure shows a tightening of the school district clusters with higher performance. The middle of this structure is where most of the 168 standard deviations show the possible diversity effect described in this paper. This area is where the Fruehwirth theory is most consistent with the predicted mixing effects.

Discussion of Observations and Potential Policy Issues

The achievement gap was examined in New Jersey, one of the most diverse states in the United States, and found to vary as much as 60% with school district diversity. Using data mining techniques, we examined 480 New Jersey school districts with an enormous range of diversity. This large and diverse data set made these analyses possible by improving our ability to detect this diversity effect on test scores due to the increase in signal to noise ratio. Without such a comprehensive data set, it would have been difficult to find any relationship between these test scores and school district properties.

As the present analysis is fundamentally an exploratory, data mining approach, we cannot show a direct causal relationship with the most likely explanation of this surprising effect without more outcome data. For example, what happens to students within each cluster over time? Nonetheless, we can expand on a consistent set of predictions based on peer effect models. The main spillover effect uncovered in our analysis of NJ ASK 2010–2011 data was a significant lower achievement gap with higher diversity per school district, showing a potential relationship between achievement gap and school district diversity.

Note that some school districts did not actually show significant diversity effects or even significant achievement gaps. There appears to be a point at which extreme segregation in those districts with very high overall score performance (ceiling effect) or districts with very low overall performance (floor effect) functionally masks the potential achievement gap. In cases in which there is no smaller achievement gap or no detectable achievement gap at all, median scores are either equally high or low. Although it is possible to find samples of White students (especially males) performing higher than Black students in low performance and predominantly Black schools (Bohrnstedt et al., 2015), our random samples of each race group in these types of districts tended to show similar performance and no achievement gap. This finding likely represents a floor effect rather than an increase in the performance of the minority sample. Alternatively, high-performing predominantly White schools with minority samples also show lower achievement gaps, but in this case, it is likely due to a ceiling effect. In both cases, there may be a tipping point at which the conditions that have created either school districts with very low or very high performance create significant minority peer effects, ameliorating the achievement gap but at the same time either “raising all boats” or dragging them all down.

The potential productive effect of increased diversity shown in this paper occurs in the midrange of school district performance. Finally, this potential effect does not appear to be related to specific interventions or educational policies adopted by the New Jersey government; rather, it seems to be happening organically, possibly due to the rapidly changing demographics and diversity occurring throughout the United States in the last decade and New Jersey in particular (Orfield et al., 2017).

Policy strategies could arise from better understanding this latent cluster structure identified in the New Jersey school districts. Whatever the final implications, at the very least the political, cultural, or socioeconomic dynamics that created these clusters of school districts must be acknowledged and treated differently on the basis of the specific clusters themselves.

There are a number of limitations of the present research: (a) we cannot produce predictive models due to the lack of longitudinal data; (b) the granularity of the data is not homogeneous, in that all variables are not at the student level—for example, the SES variable is at the school district level; and (c) without these data, we cannot create student profiles or generic templates to help understand individual differences between clusters and provide clearer policy strategies. With more data, we could gain a deeper understanding of the structures impacting academic achievement gaps, including racial isolation and poverty. These insights could inform policy issues and improve our strategies for fixing them.

Nonetheless, we can suggest a few obvious policy proposals. The cluster analysis underscores something that most seasoned educators already know—school districts cannot be treated as the same sort of entities independent of their characteristics. Clearly, it would be important to understand the cluster groups and their specific needs as well as unique responses to diversity and its effect on the achievement gap. For example, school districts grouped toward the center of a cluster structure (Figure 8) may benefit from programs that encourage diversity mixing with a higher rate of access. Districts clustered toward the top end (Cluster 8, in blue) are becoming more and more segregated (based on our diversity measures) but are breaking off in various ways from the main cluster. These programs might benefit from policy or interventions that would push them further toward the center of the structure.

School districts across the state might benefit from sharing, reflecting on, and learning from large-scale aggregate patterns at the intersection of student achievement and school district structure. Using this cluster structure as a way to organize, school district management has the potential to create a new level of functional linkage between districts.

Clusters should be reestimated periodically with new information and data, as these landscapes are highly dynamic. Superintendents and district and school administrators could use cluster group identity and classification to share best practices and successful programs with common issues and needs. Savings might result in that across clusters, where there is more similarity and specific remedies can be applied more efficiently, thus releasing resources that at a geographic level may be less effective. So, for example, some of the cluster districts defined before (e.g., low scores – midlevel – diverse) may respond to afterschool programs of a certain type more readily than others. This can be tested and then applied differentially to the various cluster groups with positive response, thus increasing the efficiency of focused experimentation across school district clusters.

In summary, the strategic adoption of using a school district's cluster model could allow for the efficient and effective use of resources and aid in reducing the achievement gap.

Note

1 The focus of the present study is with the constituents of the achievement gap, and other races were not included. However, the Asian group has an upward achievement gap with both White and Black students, which has its own possible thesis. Nonetheless, isolating groups pairwise has no implications for the covariates, as they tend to be independent.

References

American Institutes for Research. (2013, November 26). *Seven AIR experts assess the impact of the landmark "Nation at Risk" report on education in the United States* [Press release]. http://www.air.org/news/press-release/seven-air-experts-assess-impact-landmark-nation-risk-report-education-united

Baker, R. S. J. d. (2010). Data mining. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 112 – 118). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01318-X

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science*, *333*(6045), 975 – 978. https://doi.org/10.1126/science.1204534

Basch, C. E. (2011). Healthier students are better learners: A missing link in school reforms to close the achievement gap. *Journal of School Health*, *81*(10), 593 – 598. https://doi.org/10.1111/j.1746-1561.2011.00632.x

Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., & Chan, D. (2015). *School composition and the Black – White achievement gap* (NCES 2015 – 018). National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/subject/studies/pdf/school_composition_and_the_bw_achievement_gap_2015_methodology.pdf

Card, D., & Rothstein, J. (2007). Racial segregation and the Black – White test score gap. *Journal of Public Economics*, *91*(11 – 12), 2158 – 2184. https://doi.org/10.1016/j.jpubeco.2007.03.006

Cascio, E. U., & Schanzenbach, D. W. (2014). *Proposal 1: Expanding preschool access for disadvantaged children*. The Hamilton Project: Brookings Institute.

Cooley, J. C. (2010). *Desegregation and the achievement gap: Do diverse peers help?* (WCER Working Paper No. 2010-3). Wisconsin Center for Education Research. https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2010_03.pdf

Dearing, E., Kreider, H., Simpkins, S., & Weiss, H. B. (2006). Family involvement in school and low-income children's literacy: Longitudinal associations between and within families. *Journal of Educational Psychology*, *98*(4), 653 – 664. https://doi.org/10.1037/0022-0663.98.4.653

East Orange Public Schools. (2018). *East Orange School District: Mission, vision & beliefs*. http://www.eastorange.k12.nj.us/Mission-Vision-Beliefs

Ferguson, R., Hackman, S., Hanna, R., & Ballantine, A. (2009). *Raising achievement and closing gaps in whole school systems: Recent advances in research and practice* [2008 conference report]. The Achievement Gap Initiative at Harvard University. http://www.agi.harvard.edu/events/2008Conference/GETTING_IT_DONE_02_24_09.pdf

Fruehwirth, J. C. (2013). Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics*, *4*(1), 85 – 124. https://doi.org/10.3982/QE93

Fruehwirth, J. C. (2014). Can achievement peer effect estimates inform policy? A view from inside the black box. *The Review of Economics and Statistics*, *96*(3), 514 – 523. https://doi.org/10.1162/REST_a_00385

Guryan, J. (2004). Desegregation and black dropout rates. *American Economic Review*, *94*(4), 919 – 943. https://doi.org/10.1257/0002828042002679

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2008). *Standards-based reform in the United States: History, research, and future directions*. RAND. https://www.rand.org/pubs/reprints/RP1384.html

Hanson, S. J., Gagliardi, A. D., & Hanson, C. (2009). Solving the brain synchrony eigenvalue problem: Conservation of temporal dynamics (fMRI) over subjects doing the same task. *Journal of Computational Neuroscience*, *27*(1), 103–114. https://doi.org/10.1007/s10827-008-0129-z

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Holloway, S. R., Wright, R., & Ellis, M. (2012). The racially fragmented city? Neighborhood racial segregation and diversity jointly considered. *The Professional Geographer*, *64*(1), 63–82. https://doi.org/10.1080/00330124.2011.585080

Johnson, R. C. (2011). *Long-run impacts of school desegregation & school quality on adult attainments* (NBER Working Paper No. 16664). National Bureau of Economic Research. https://doi.org/10.3386/w16664

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, *35*(7), 3–12. https://doi.org/10.3102/0013189X035007003

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J. S., & Bowen, N. K. (2006). Parent involvement, cultural capital, and the achievement gap among elementary school children. *American Educational Research Journal*, *43*(2), 193–218. https://doi.org/10.3102/00028312043002193

McKinsey & Company. (2009). *The economic impact of the achievement gap in America's schools*. http://dropoutprevention.org/wp-content/uploads/2015/07/ACHIEVEMENT_GAP_REPORT_20090512.pdf

New Jersey Department of Education. (2004). *Grades 3 and 4 New Jersey Assessment of Skills and Knowledge*. https://www.state.nj.us/education/assessment/es/njask_tech_report04.pdf

New Jersey Department of Education. (2011). *New Jersey Assessment of Skills and Knowledge grades 3–8: 2010 technical report*. http://www.nj.gov/education/assessment/es/NJASK_tech_report10.pdf

New Jersey Department of Education. (2016). *Administrative code and statute, chapter 8, standards and assessment*. http://www.state.nj.us/education/code/current/title6a/chap8.pdf

New Jersey Department of Education. (2018a). *New Jersey school performance reports: East Hanover Twp 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=1190&distName=EAST%20HANOVER%20TWP&year=2016-2017

New Jersey Department of Education. (2018b). *New Jersey school performance reports: East Orange 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=1210&distName=EAST%20ORANGE&year=2016-2017

New Jersey Department of Education. (2018c). *New Jersey school performance reports: Ewing Twp 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=1430&distName=EWING%20TWP&year=2016-2017

New Jersey Department of Education. (2018d). *New Jersey school performance reports: Mansfield Twp 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=2970&distName=MANSFIELD%20TWP&year=2016-2017

New Jersey Department of Education. (2018e). *New Jersey school performance reports: New Brunswick City 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=3530&distName=NEW%20BRUNSWICK%20CITY&year=2016-2017

New Jersey Department of Education. (2018f). *New Jersey school performance reports: Piscataway Twp 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=4130&distName=PISCATAWAY%20TWP&year=2016-2017

New Jersey Department of Education. (2018g). *New Jersey school performance reports: Princeton Regional 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=4255&distName=PRINCETON%20REGIONAL&year=2016-2017

New Jersey Department of Education. (2018h). *New Jersey school performance reports: Union Twp 2016–2017*. https://rc.doe.state.nj.us/schoollist.aspx?district=5290&distName=UNION%20TWP&year=2016-2017

Noguera, P., Darling-Hammond, L., & Friedlaender, D. (2015). *Equal opportunity for deeper learning*. Jobs for the Future. https://jfforg-prod-prime.s3.amazonaws.com/media/documents/Equal-Opportunity-for-Deeper-Learning-100115a.pdf

Orfield, G., Ee, J., & Coughlin, R. (2017). *New Jersey's segregated schools: Trends and paths forward*. UCLA Civil Rights Project. https://www.civilrightsproject.ucla.edu/research/k-12-education/integration-and-diversity/new-jerseys-segregated-schools-trends-and-paths-forward/New-Jersey-report-final-110917.pdf

Patton, D. U., Woolley, M. E., & Hong, J. S. (2012). Exposure to violence, student fear, and low academic achievement: African American males in the critical transition to high school. *Children and Youth Services Review*, *34*(2), 388–395. https://doi.org/10.1016/j.childyouth.2011.11.009

Peske, H. G., & Haycock, K. (2006). *Teaching inequality: How poor and minority students are shortchanged on teacher quality*. The Education Trust. https://edtrust.org/wp-content/uploads/2013/10/TQReportJune2006.pdf

Princeton Public Schools. (2018). *Who we are*. https://www.princetonk12.org/district/about-us

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A., & Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study: Final report* (OPRE Report 2012–45). Administration for Children & Families. https://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf

Reardon, S. F. (2016). School segregation and racial academic achievement gaps. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *2*(5), 34–57. https://doi.org/10.7758/rsf.2016.2.5.03

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. *3*, pp. 249–277). Elsevier. https://doi.org/10.1016/B978-0-444-53429-3.00004-1

Sciarra, D. G., & Hunter, M. A. (2015). Resource accountability: Enforcing state responsibilities for sufficient and equitable resources used effectively to provide all students a quality education. *Education Policy Analysis Archives*, *23*(21), 1–31. https://doi.org/10.14507/epaa.v23.2032

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Silver, M. (2004). Trends in school reform. *New Horizons for Learning*, *5*(2005).

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25. https://doi.org/10.2307/2331554

Tractenberg, P., Orfield, G., & Flaxman, G. (2013). *New Jersey's apartheid and intensely segregated urban schools: Powerful evidence of an inefficient and unconstitutional state education system*. Rutgers University Institute on Education Law and Policy. https://clime.newark.rutgers.edu/sites/CLiME/files/IELP%20final%20report%20on%20apartheid%20schools%20101013.pdf

Tumin, M. M. (1959). *Desegregation: Resistance and readiness*. Princeton University Press. https://doi.org/10.1515/9781400876341

U.S. News. (2019). *Search best high schools*. https://www.usnews.com/education/best-high-schools/search?ranked=true&state-urlname=new-jersey.

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). *Achievement gaps: How Black and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (Statistical Analysis Report No. NCES 2009-455). National Center for Education Statistics.

Vigdor, J., & Nechyba, T. J. (2007). Peer effects in North Carolina public schools. In P. Peterson & L. Woessmann, (Eds.), *Schools and the equal opportunity problem* (pp. 73–102). MIT Press.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.

Wu, S. (2011). *Growing New Jersey minority population reaches majority in some municipalities*. Labor Market & Demographic Research. https://www.nj.gov/labor/lpa/pub/lmv/LMV_7.pdf

Appendix A

Derivation of Variables

LAL scores (raw scale)—taken directly from each student's NJ ASK record.

Mathematics scores (raw scale)—taken directly from each student's NJ ASK record.

Diversity scores—constructed per school district from the proportion of students in that district that were indicated to be White, W; Black, B; and Hispanic, H. A relative entropy or information-based calculation was done with the three proportions, which would be 1.0 with <.33,.33,.33> distributions or 0 with any of the cases <1,0,0>, <0,1,0>, or <0,0,1>.

School district socioeconomics (SocEco)—Although student-level SES data were not available, there were 10 levels within the school district indicating an average weighting per student, A to J, which indicated whether the school district was low to high in 10 steps. The codes A and J did not appear in the survey; hence, we used eight levels and coded them as integers for each of the school districts. We further used the proportion of students who had reduced or free lunch (RFL) in that school district to weight the categorical codes, biasing them toward continuous and providing a more valid socioeconomic variable. Principal component analysis (PCA) was used to construct the composite SocEco measure between DFG codes (as integers) and the RFL proportions, creating a new variable with more than 70% of variance of the original variables.

School district size—the actual number of students in that school district, which could range from 10s to 1000s

Appendix B

New Jersey Municipal School Districts Cluster Prototypes

Cluster 1

- Princeton
- Madison
- Clinton Township
- Manasquan
- Livingston Township
- Marlboro Township
- Metuchen

Cluster 2

- East Hanover Township
- Florham Park Borough
- Freehold Township
- Stanhope Borough
- Vernon Township

Cluster 3

- Edgewater Park Township
- Fort Lee
- Lawrence Township
- Montclair
- Parsippany Troy Hills
- Piscataway Township
- Ridgefield
- South Orange Maplewood
- South Plainfield

Cluster 4

- Bellmawr Borough
- Union Township
- Brigantine City
- Jamesburg Borough
- Totowa Borough
- West Cape May Borough

Cluster 5

- Hawthorne
- Highlands Borough
- Hopewell Township
- Ocean Township
- Stow Creek Township
- Mansfield Township

Cluster 6

- New Brunswick
- Bogota
- Kearny
- West New York

- Bound Brook Borough

Cluster 7

- Berlin Township
- Beverly City
- Elmwood Park
- Elsinboro Township
- Ewing Township
- Mount Holly Township
- Neptune City
- Seaside Heights Borough

Cluster 8

- Franklin Township
- Jersey City
- Pleasantville
- East Orange
- Newark
- Oselle Borough

Suggested citation:

Hanson, S. J., Saunders, S., Aponte, A., Copeland, R., Nettles, M.. (2020). *Diversity associated with reductions in the achievement gap: Data mining the 2010-2011 New Jersey Assessment of Skills and Knowledge (NJ ASK)* (Research Report No. RR-20-12). Educational Testing Service. https://doi.org/10.1002/ets2.12292

Action Editor: James Carlson

Reviewers: Qiwei He and Anita Sands