

Parents' Online School Reviews Reflect Several Racial and Socioeconomic Disparities in K–12 Education

Nabeel Gillani

Eric Chu

Doug Beeferman

Massachusetts Institute of Technology

Rebecca Eynon 

University of Oxford

Deb Roy

Massachusetts Institute of Technology

Parents often select schools by relying on subjective assessments of quality made by other parents, which are increasingly becoming available through written reviews on school ratings websites. To identify relationships between review content and school quality, we apply recent advances in natural language processing to nearly half a million parent reviews posted for more than 50,000 publicly funded U.S. K–12 schools on a popular ratings website. We find: (1) schools in urban areas and those serving affluent families are more likely to receive reviews, (2) review language correlates with standardized test scores—which generally track race and family income—but not school effectiveness, measured by how much students improve in their test scores over time, and (3) the linguistics of reviews reveal several racial and income-based disparities in K–12 education. These findings suggest that parents who reference school reviews may be accessing, and making decisions based on, biased perspectives that reinforce achievement gaps.

Keywords: *education, inequality, family engagement, achievement gaps, natural language processing, interpretable machine learning, computational social science, regression*

Introduction

Equitable access to quality education continues to be an elusive goal for many nations worldwide (Graetz et al., 2020). In response, choice-based systems have grown in popularity over the past several decades (Musset, 2012): for example, in the form of expanded in-district schooling options, and charter schools. While certain choice-based systems (like some U.S. urban charter school networks) have proven to enhance students' outcomes (Angrist et al., 2011; Deming et al., 2014; Dobbie & Fryer, 2009), many have questioned the extent to which such systems can, and should, be a primary vehicle for increasing equitable access to quality education (André-Béchely, 2005; Lubienski et al., 2009; Musset, 2012).

Whether choice-based systems hold promise for bridging inequalities in education or not, they are increasingly being deployed, and their effectiveness is predicated on families

being well-informed and well-supported to decide which schools are best for their children. To achieve this, parents frequently turn to their social networks to gather feedback on which schools they should send their children to (Ball & Vincent, 1998), often using the characteristics of students who attend a school as signals of quality (Bell, 2009; Schneider & Buckley, 2002). For those subsets of parents enmeshed in privileged networks replete with social capital and access to other “in-the-know” parents (Small, 2009), these behaviors can amplify their tendency to identify and enroll their children in preferred schools, which are in many cases those where their children would be surrounded by others from similar racial and socioeconomic backgrounds. In a macrosense, this practice has fueled a “rich get richer” network effect and exacerbated racial and income segregation in schools (Roda & Stuart Wells, 2013), which further limits the access minority, low income, and other vulnerable



children have to critical educational resources (Reardon & Owens, 2014).

Ironically, this macro-level trend is contrary to the micro-level preferences many parents have expressed for integrated schools (Torres & Weissbourd, 2020), highlighting a disconnect between individual and collective societal incentives that further obfuscates the quest for educational equity.

School ratings sites like GreatSchools.org, Niche.com, and others have gained popularity in recent years as resources to help parents research and identify schools (Lovenheim & Walsh, 2017), and are now an important part of the landscape in facilitating parental choice. They have the potential to offer parents who may be less likely to be tapped into privileged networks the opportunity to identify higher-quality schooling options for their children. A recent study illustrated how showing Section 8 housing voucher recipients information about school quality drawn from such platforms could lead to housing selection near higher quality schools (Bergman et al., 2020). However, there has also been evidence arguing that more information about schools available through ratings websites has exacerbated racial and income segregation (Hasan & Kumar, 2019), especially when disproportionately accessed by more affluent families to decide where to live. One reason for this may be because the ratings on these sites often primarily reflect the schools' test scores, and hence, demographics (Barnum & LeMarr LeMee, 2019)—thus making selection on the basis of “peer performance” more likely (Abdulkadiroglu et al., 2019).

Given this context, it is important to further explore the information presented on these platforms to understand the role they could play in reducing, maintaining, or exacerbating inequalities. An important issue is how quality is defined and communicated on such sites. Yet there is considerable debate about how to define, measure, and enhance “quality” when it comes to K–12 education. While some experts define quality primarily by how students perform on standardized tests—a “snapshot” measure that is highly correlated with race and income (Reardon et al., 2018)—others define it by how well a school helps students make academic progress over time (“student growth,” or “school effectiveness” when the causal effect of the school on student learning can be measured net biases that could emerge due to students' out-of-school experiences and environments). Still others define school quality according to more qualitative, relational measures that collectively describe the “climate” or environment of the school (Cohen et al., 2009).

Several studies have illustrated the relationships between measures of student growth and children's intergenerational outcomes (Chetty et al., 2011; Chetty et al., 2014). These growth measures, which researchers and practitioners often use as a measure of a school's value-add or “effectiveness” (Koretz, 2008; Reardon, Papay, et al., 2019), tend to correlate less with the demographic makeup of students than

snapshot measures like standardized tests (Reardon, 2017). Researchers have also identified several features of effective schools, particularly those that serve minority and low-income youth, which include practices like having high expectations for students; access to tutoring; and regular feedback for teachers (Dobbie & Fryer, 2013; Edmonds, 1979). While effectiveness measures are hardly comprehensive in capturing the quality of education a school offers, they represent a useful metric for understanding the impact schools have on student achievement.

Typically, school ratings websites will present information on measures that seek to describe school effectiveness, test scores, climate, and other indicators. They also present open-ended written reviews posted by parents. Such reviews have been shown to affect parents' perceptions of school quality (Loeb & Valant, 2013), yet remain a relatively underresearched aspect of these platforms. Online school reviews offer a new signal of parents' perceptions about schools that may stem from their desires to inform, or warn, other parents about certain features of schools, publicly raise or diminish the school's profile, or simply have their voices heard. Thus, much like they have for commercial domains like restaurants (Zhang et al., 2010), online school reviews may both shed light on and influence the preferences and values of a critical stakeholder in education: parents. This analogy to the commercial domain is particularly apt in light of arguments about how choice-based education systems often resemble markets stratified by race and social class (Ball, 1993). If school reviews are already influencing how parents make schooling decisions, and are likely to do so even more as reviews abound (like they have for other consumer products), then better understanding the content of these reviews and what they reveal about the quality of products (schools) in education markets is critical.

In this work, we explore the content of parents' written reviews to explore what information they contain about different measures of school quality. The site of our analysis is GreatSchools.org, a popular schools rating website receiving nearly 50 million visits in 2019 (GreatSchools.org, 2020). Our primary goal is to analyze the linguistic content of parents' reviews to identify which school characteristics are correlated with measures of school effectiveness. We build on recent advances in natural language processing and interpreting black-box machine learning models to perform our analyses. Our results reveal a weak-to-nonexistent relationship, on average, between the content of parents' reviews and measures of school effectiveness, instead illuminating strong relationships between reviews and both school-level test scores and demographics. Furthermore, we find that many of the words and phrases that predict test scores and demographics implicitly encode information about the racial and socioeconomic makeup of schools. Finally, we reflect on how these implicit characterizations of demographics

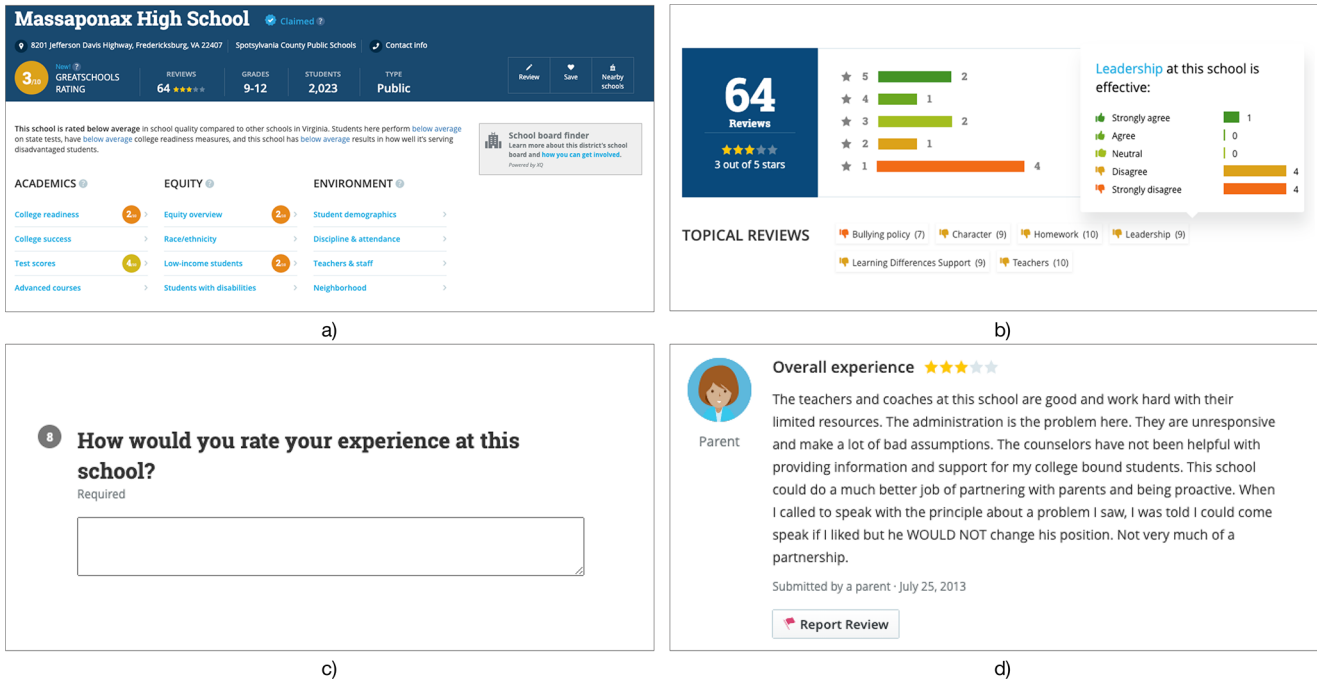


FIGURE 1. Screenshots from a sample GreatSchools.org page: (a) shows the main “above-the-fold” view of a school’s webpage on the platform, including its overall rating and various subscore ratings; (b) shows the average user-provided five-star rating of a school, along with ratings for various topics defined by GreatSchools; (c) illustrates the prompt (as of mid-2020) that parents see when they are asked to submit their written review; and (d) is an example of a written review provided by a parent.

may reinforce existing achievement gaps through “market failures” in choice-based systems, especially when considering the potential power of language to shape human judgment and decision making (Lewis & Lupyan, 2020).

Method

Data Sets

We combine three different data sets for our analyses. First, we use a webscraper to collect approximately 830,000 reviews posted by parents for more than 110,000 school on the popular U.S. K–12 school ratings site GreatSchools.org, with permission from the organization. The reviews total nearly 54 million words. Figure 1 offers screenshots of the GreatSchools reviews interface and explanations of key elements of the site.

We link these schools to the Stanford Educational Data Archive (SEDA; Reardon, Ho, et al., 2019). SEDA reports school-level, nationally normed performance metrics for approximately 80,000 publicly funded elementary and middle schools, averaged from 2008 through 2016. We specifically focus on two performance metrics: (1) average test scores (“test scores”), which provide a snapshot-in-time measure of student performance and (2) student learning rates (“progress scores”), which indicate how much student cohorts improve on standardized tests each year. While the

progress scores do not offer an unbiased measure of a school’s causal effect on student achievement (because they do not account for out-of-school factors that may also affect learning), they reveal more about how much students at certain schools improve relative to others—and are much less correlated with demographics—than test scores (Reardon, Papay, et al., 2019). Hence, we refer to progress scores as proxies for “school effectiveness.” SEDA also includes several school-level characteristics, like the racial and socioeconomic demographics of the school. Filtering down to schools for which SEDA contains our outcome measures of interest and removing those schools for which there are no parent reviews, leaves approximately 54,000 and 40,000 schools with test and progress scores, respectively, corresponding to approximately 454,000 and 361,000 parent reviews.

Finally, for a broader view of the neighborhood context in which a school is situated, we geocode the school addresses available on GreatSchools and link them to tract-level estimates of race, socioeconomic status, and other demographics provided by the 2010 Census and 2015 American Community Survey.

Additional details on data collection and merging can be found in the online Supplemental Appendix. All data collection, storage, and analysis was conducted at the Massachusetts Institute of Technology and approved by its Institutional Review Board protocol No. E-2481.

Identifying Correlations Between Review Text and School-Level Characteristics

Exploratory text analysis is a form of natural language processing (NLP) that is gaining popularity in the social sciences as a method for better understanding human psychology and behavior (Fesler et al., 2019; Gentzkow et al., 2019). Latent Dirichlet Allocation (LDA) is a popular unsupervised method of text analysis (Blei et al., 2003), enabling researchers to identify latent groupings, or “topics,” across documents in textual corpora. LDA has been applied across a wide range of domains, and has also been extended over the past two decades to allow for supervision and other modifications (Blei & McAuliffe, 2008). Structural topic models (Roberts et al., 2013) extend the concepts of LDA by factoring in the possibility that the nature and composition of topics, and how they manifest across documents, might differ according to certain covariates (like demographics, political leaning, etc.).

While such topic models offer a powerful toolset for exploratory text analysis, they have several drawbacks. One is that the number of topics must be specified as a hyperparameter before training the model. There are data-driven methods for identifying these hyperparameters, but they often still produce a coarse or otherwise difficult-to-interpret set of latent topics. Furthermore, there is a high level of researcher discretion required to name the resulting topics, a process that often amounts to inspecting words that have been identified as “belonging to” a given inferred topic and qualitatively determining common underlying semantic relationships between those (sometimes dissimilar) words. Additionally, many of these models treat documents as “bags of words,” discarding potentially valuable information contained in word ordering, multiword phrases, and other structural characteristics that often also have bearing on semantics.

As stated earlier, our primary objective is to identify the words and phrases from parents’ reviews that correlate with different measures of school quality. One straightforward approach involves using supervised machine learning, instead of traditional topic modeling like LDA. In particular, we seek to train a regression model that accepts as an input the reviews from a school in order to (1) predict outcomes describing that school (e.g., test scores, progress scores) and (2) shed light on how important different words and phrases from reviews were in driving that model’s predictions. The subsections below describe the methods we used to complete these two steps.

Fine-Tuning Large-Scale Language Models for Regression.

Advances in deep neural networks over the past several years have fueled several new developments in NLP. Deep neural networks are machine learning models that often automatically identify complex, nonlinear relationships between input variables in order to more accurately make

predictions or perform some other downstream task (LeCun et al., 2015). This has made them particularly adept for different tasks that involve linguistic data, where subtle, context-specific relationships between words and punctuation can have substantial implications for semantics.

One of the most promising developments in machine learning that have been enabled by deep neural networks is “transfer learning.” Often times, a limiting factor in leveraging the power of deep learning models is the nature and amount of available training data. Transfer learning enables researchers to build on top of and continue training (i.e., “fine-tune”) models that have previously been trained (“pre-trained”) on large data sets. This process allows modelers to both borrow and refine the understanding of language acquired by pretrained models in order to improve performance on some other task. This pretrain-and-fine-tune paradigm has been a large driver of recent advances in NLP, enabling models to achieve previously unimaginable (and often ethically concerning) feats like “writing” entirely novel stories after being conditioned on just a few sentences of starter context (Brown et al., 2020).

We augment and fine-tune the popular pretrained, deep neural network language model BERT (Devlin et al., 2019)—which stands for “Bidirectional Encoder Representations from Transformers”—for our regression task. BERT has been pretrained on a large corpus of English books and English Wikipedia (Wolf et al., 2020) and designed to learn complex, nonlinear relationships between words and subword pieces in sentences in order to achieve state of the art performance on a variety of natural language understanding tasks (Devlin et al., 2019). Several researchers have investigated the inner workings of BERT to better understand possible reasons for such strong performance, discovering that it automatically learns and exploits various linguistic features of the input data (like parse trees and word polysemy) that previously required explicit engineering and selection by model developers (Coenen et al., 2019; Manning et al., 2020). In order to identify high-fidelity correlations between parents’ reviews and school outcomes like test and progress scores, our regression model must learn to identify and exploit salient features in the review text to predict these outcomes as accurately as possible. It is for this reason that we fine-tune BERT: to borrow any relevant information about syntax, semantics, and other linguistic features that the model has already learned and use it as a starting point for our bespoke prediction tasks, instead of “starting from scratch.” Our model can then “learn to forget” information distilled by the pretrained model that is not useful for predicting school outcome measures.

The original BERT was trained on snippets of text that are 512 tokens (i.e., “word pieces,” or subword-strings that can be combined to form words in the original input sequence) in length, while a school’s reviews could span thousands of tokens. To address this, we select up to 100

sentences from the most recent reviews posted for each school, and up to the first 30 tokens per sentence, to represent the school. Note that we limit the number of sentences and characters we select due to GPU memory constraints. We train on a random subset of 90% of the data and use 10% for validation; however, given the computational resources required to train each model, we do not perform cross-validation. Our BERT-based model performs better than several baselines, including classical, not-pretrained models (e.g., regularized linear regression using TF-IDF representations of reviews). The online Supplemental Appendix provides additional information on our model architecture, training procedure, and model performance.

Identifying Highly Predictive Phrases. We train the above language models to satisfy our original objective of identifying which features of parents' reviews are most correlated with measures of school quality. Unfortunately, deep neural networks make decisions by exploiting obscure patterns in intermediate representations derived from their original input data (LeCun et al., 2015). These intermediate representations are optimized to increase model accuracy but come at a cost: They make the model a "black-box" whose inner workings are difficult to interpret. In our case, this means that it is difficult to understand precisely which words and phrases are most influential in predicting school performance measures. One popular method for making the decision-making processes of these black box models more transparent is Integrated Gradients (IG; Sundararajan et al., 2017). Intuitively, IG computes an attribution, or "importance" value, per feature (i.e., each word in a review), per instance (i.e., each school). The attribution value represents how much that word, when occurring alongside other words in the instance, changes a neural network's prediction relative to the prediction it would have made in the absence of any linguistic signal (i.e., an empty string). Positive attributions indicate a positive correlation between the word and outcome for that instance; negative attributions indicate a negative correlation. By computing attributions in this way, IG accounts for differences in how the same words and phrases might be used across different reviews and schools.

We use the Captum library (Kokhlikyan et al., 2020) to compute integrated gradients as a measure of how much each word influences our model's predictions, and spaCy's base noun phrase extractor (Honnibal & Montani, 2017) to identify noun phrases as proxies for the primary foci of a given review sentence. After summing the attributions of words comprising each detected noun phrase, we sum and normalize the attributions assigned to noun phrases across all reviews and schools. Positive attributions suggest that the phrase is positively correlated with the outcome that the model was trained to predict; negative values suggest a negative correlation with the outcome. Intuitively, our summation and normalization procedure assigns phrases that occur

frequently and consistently (i.e., generally receive positive or negative attributions in the context of a given school's reviews) high absolute value attributions; conversely, phrases that do not occur frequently, or occur inconsistently (i.e., that sometimes are assigned positive attributions and other times negative) will have attribution values close to zero. Additional details on the summation and normalization procedure can be found in the online Supplemental Appendix.

Phrase Clustering

Given that many noun phrases in reviews are similar (e.g., "school" and "the school"; "excellent school" and "wonderful school"; etc.), we seek to cluster those noun phrases with similar meanings. To do so, we use a separate pretrained sentence "encoder" presented in (Reimers & Gurevych, 2019), which is also built by adapting a variant of BERT. The encoder is a model responsible for mapping an input sentence to a low-dimensional vector representation that captures important semantic information. This vector is considered "low-dimensional" because it uses a number of dimensions that is much smaller than the vocabulary of our data set to represent a given phrase or sentence. Intuitively, the encoder maps two similar sentences (or phrases) to similarly valued vectors if their meanings are similar.

We use an instance of this pretrained encoder, which we refer to as $enc(\cdot)$, to do the following: (1) produce a vector representation for each noun phrase n , that is, $enc(n)$ and (2) cluster these resultant vectors using the HDBSCAN clustering algorithm (McInnis et al., 2017) with a minimum cluster size of 2 (phrases that were not clustered alongside any other phrases are assigned to their own singleton clusters). In the results section, we use this method for two sets of analyses: first, for an overarching look at the kinds of phrases contained across all reviews posted by parents, and second, to analyze correlations between phrases and different measures of school quality as computed by IG on the validation set used to evaluate our BERT-based models. For this second set of analyses, we compute a weighted average and weighted standard deviation of the attribution values assigned by IG to each phrase n in cluster C_n , where weights are represented by the number of sentences n occurs in across the validation set. In both cases, we only include noun phrases that occur in at least 1% of schools in each respective corpus (i.e., 1,100 and 50 in the first and second sets of analyses, respectively).

Sentence Prototypes

When analyzing phrases, we may be interested in better understanding broader context about the kinds of review sentences they tend to be used in. To identify phrase n 's typical review semantic context, we define its sentence prototype, s_n^* , as a review sentence containing n whose

semantic vector representation (again inferred by the method presented in Reimers & Gurevych, 2019) has the highest cosine similarity with the mean across all semantic vector representations of sentences containing n (represented by S_n), that is:

$$s_n^* = \operatorname{argmax}_{s \in S_n} \operatorname{cosine}(\operatorname{enc}(s), \frac{1}{|S_n|} \sum_{s' \in S_n} \operatorname{enc}(s')), \quad (1)$$

Intuitively, this selects the review sentence that is “semantically closest” to the average semantics of all sentences that contain n .

Phrase Rarity

For additional context about how phrase usage differs according to different types of schools, we define the rarity of a phrase as its inverse-document frequency score. Intuitively, this metric indicates how often a phrase occurs in the reviews of a specific type of school (e.g., a lower performing school) compared with how frequently it *could* occur if, in an extreme case, all such reviews contained it. More formally, let $q \in \{1, 2, 3, 4, 5\}$ denote a set of quintiles with respect to some school-level variable (e.g., test or progress scores) where 1 indicates the bottom quintile (bottom 20%), and R_q the set of reviews posted for schools belonging to quintile q (note we could have chosen quartiles, tertiles, or any other discretization). Furthermore, let P denote a cluster of noun phrases drawn from our data, and R_q^P the set of reviews in quintile q , which also contain at least one of the phrases in P . We define the rarity of the set of phrases P in quintile q as $\log(\frac{|R_q^P|}{|R_q|})$. Larger values of this expression imply that a smaller fraction of reviews out of the total number of reviews in a given quintile contain one of the phrases in P (i.e., the phrases are less frequent, or rarer, in the reviews of that quintile).

Results

Correlations Between School Performance Measures and Demographics

Figure 2 shows the Pearson correlations between school performance measures and demographics drawn from SEDA, weighted by the total number of students enrolled at each school. We include GreatSchools’ measures of school performance as well for comparison purposes, namely: (1) their overall rating of a school, which is computed as a function of subscores assigned to several subcategories (GreatSchools.org, n.d.); (2) their test score subscore for the school; and (3) their progress score subscore for the school.

The correlation matrix reveals strong relationships between both the GreatSchools and SEDA test score measures and the racial and income demographics of a school, represented by the percentage of students who are White

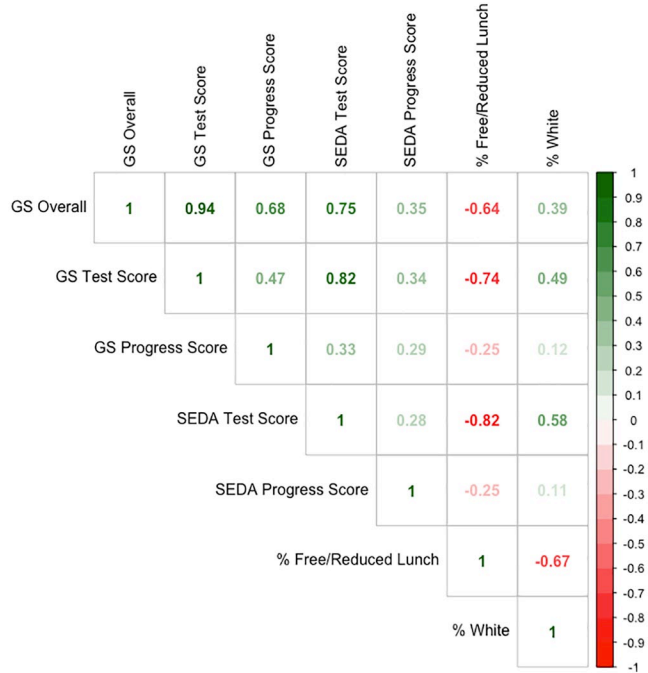


FIGURE 2. Correlations between test and progress scores sourced from GreatSchools and SEDA, along with school-level demographic covariates from SEDA.

Note. Correlations are weighted by the size of the school’s enrollment. Test score measures are highly correlated with racial and income demographics of schools, whereas progress score measures are not. SEDA test and progress scores also appear to have a low correlation with one another, suggesting they are measuring different aspects of school performance. Interestingly, there is a low correlation between the progress scores reported by GreatSchools and SEDA. SEDA = Stanford Educational Data Archive.

(GreatSchools $r = .49$; SEDA $r = .58$) and receive free or reduced lunch (GreatSchools $r = -.74$, SEDA $r = -.82$), respectively. Both the GreatSchools and SEDA progress scores, however, are only weakly correlated with these measures; this reflects prior literature on the relationships between test scores, measures of school effectiveness, and students’ demographics (Reardon et al., 2018). Interestingly, while the GreatSchools and SEDA test score measures are highly correlated with each other ($r = .82$), their progress score analogs are not ($r = .29$). This may be because GreatSchools’ progress score rating is recomputed each year using measures reported by each state’s department of education, and these departments do not follow a single standard for defining and computing student progress (GreatSchools.org, n.d.). We will use the SEDA school performance metrics for the remainder of our analyses because they are nationally normed and capture a longer history of school performance.

Biases in the Availability of Reviews

Given known relationships between parent involvement in schools and demographics (Crozier, 2001; Hornsby &

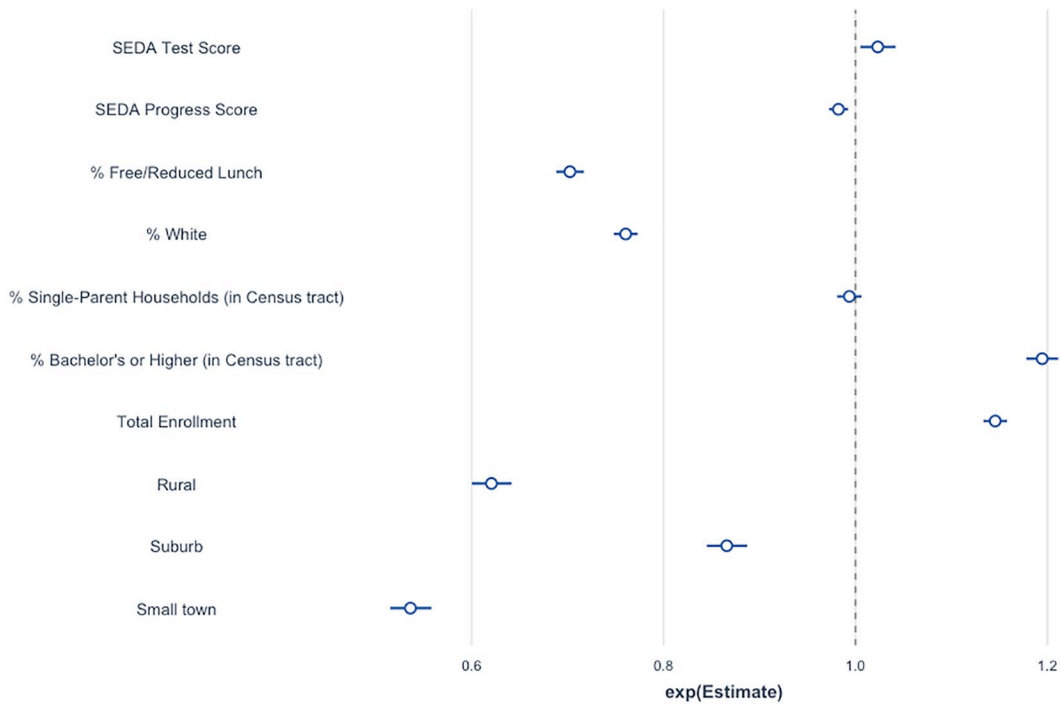


FIGURE 3. *Coefficients inferred from a negative binomial regression of the number of parent reviews schools receive on a basket of performance measures and demographic covariates.*

Note. Predictor variables are standardized before fitting the regression (by subtracting the mean and dividing by the *SD* of the data). Coefficient values indicate odds ratios: for example, a *1SD* increase in the percentage of students receiving free or reduced lunch at a school (approximately 26%) corresponds to the school receiving 30% fewer reviews, holding all other predictors constant. Schools in small towns and rural areas are particularly prone to receiving fewer reviews, while larger schools and schools located in neighborhoods with a higher percentage of bachelors' degree or higher recipients are particularly likely to receive more reviews.

Lafael, 2011), it is likely that there are biases in which schools tend to receive more reviews from parents. To understand these biases prior to analyzing the content of reviews, we use a negative binomial regression to predict the number of parent-provided reviews schools receive as a function of a set of school and neighborhood-level characteristics. The regression coefficients and standard errors are depicted in Figure 3. In general, schools with more students, those located in cities, those in census tracts with a higher percentage of adults who have received at least a bachelor's degree, and those with fewer students receiving free or reduced lunch tend to receive more reviews (after controlling for the other depicted variables). Perhaps counterintuitively, schools that have a higher percentage of White students also tend to receive fewer reviews. The full regression table is included in the online Supplemental Appendix. Conditional on a school receiving at least one review, we see similar trends when analyzing how average word length of reviews differs across different types of schools: schools that are less likely to receive reviews are also more likely to receive shorter reviews, though these differences are much less pronounced (see online Supplemental Appendix for more details).

Exploring the Content of Written Reviews

For an overarching look at the kinds of content contained in parents' reviews, we use the noun phrase extraction and clustering approaches described in the Methods section. Figure 4 shows the outputs of this method when applied to the entire corpus of 830,000 reviews posted by parents across 110,000 schools. The figure shows only those clusters whose most frequent constituent noun phrase occurs in the reviews of at least 5% of schools in the corpus (approximately 5,500 schools—a full list of all noun phrase clusters and their prevalence can be found linked in the online Supplemental Appendix). From the results, we can see that discussions about teachers, staff, principals, and other leadership and administration at the school are highly prevalent. We also see that parents tend to reference other parents, families, and parental/family involvement quite frequently. Parents also discuss the curriculum, homework, and specific academic subjects like math and reading—and in some cases, music and the arts—though less frequently than the aforementioned topics. Also present (though again, less frequent) are discussions about climate-related topics like bullying, fun, community, and care. We note that in some cases, our method highlights differences in language use and/or

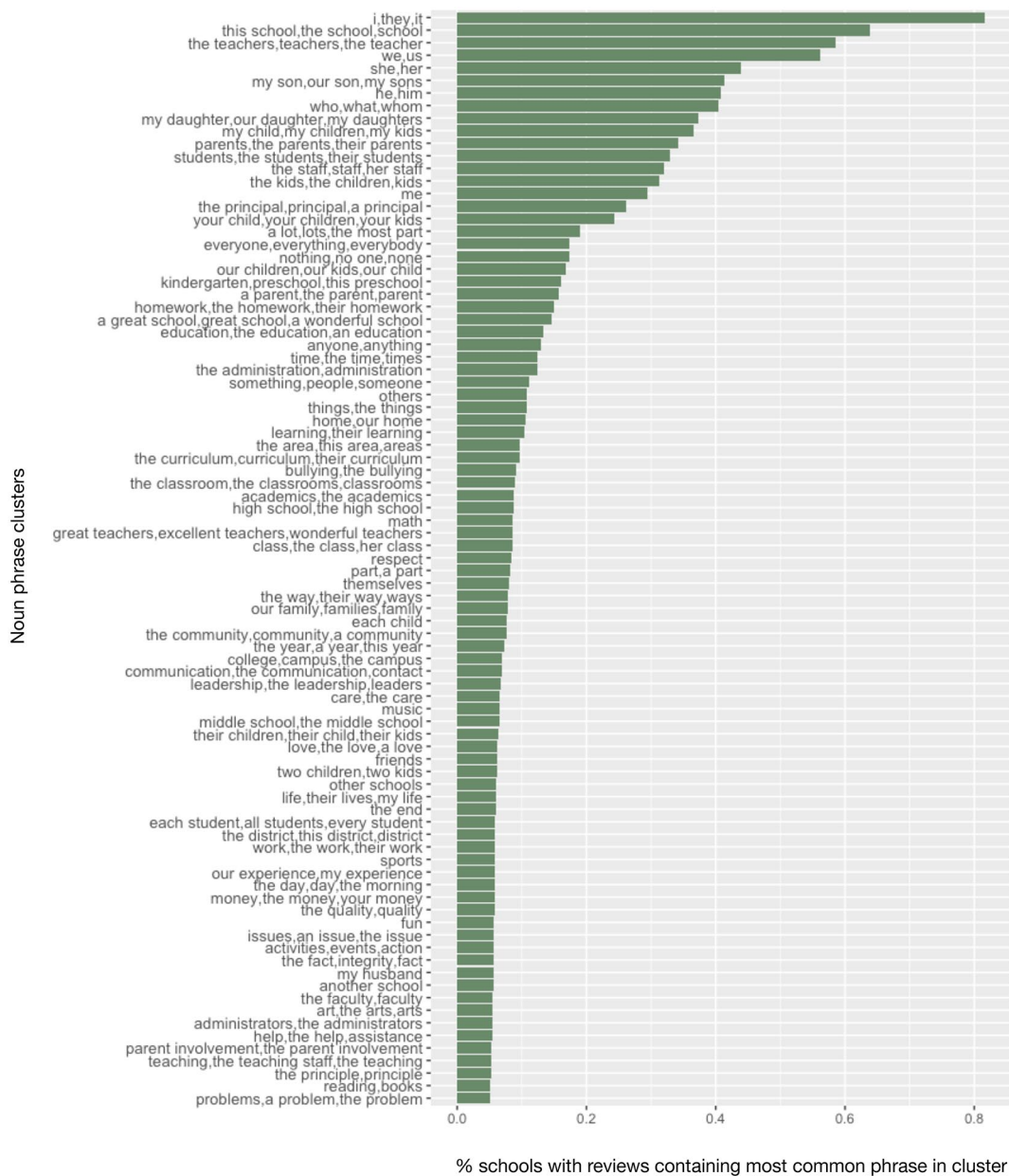


FIGURE 4. Prevalence of noun phrase clusters (topics) across the full corpus of 830,000 reviews posted by parents.

Note. The y-axis shows the top three most frequent phrases comprising each cluster (bar); the x-axis indicates the percentage of schools with at least one review that contains the most prevalent phrase in each cluster. Here, we only show those phrase clusters represented in at least 5% of all schools containing parent-provided reviews. The full list of clusters and corresponding frequency information can be found linked in the online Supplemental Appendix.

intent that might otherwise be ignored or preprocessed away by traditional topic modeling methods like LDA (e.g., “teacher” and “teaching” are categorized into separate clusters: the former indicating a person, the latter a process of instruction).

Linguistic Correlations With Measures of School Quality

With a basic intuition for what content reviews contain, we now turn to the main objective of our study: to better

understand relationships between review content and measures of school quality. There are several possible relationships: Reviews might be correlated with both test and progress scores, correlated with one but not the other, or uncorrelated with both. Figure 5a shows the performance improvement of our model compared with a naive baseline predictor, which predicts the mean outcome value for each instance (this is equivalent, on average, to randomly sampling from the distribution of observed outcome values). Our model can predict a school’s test scores from its review

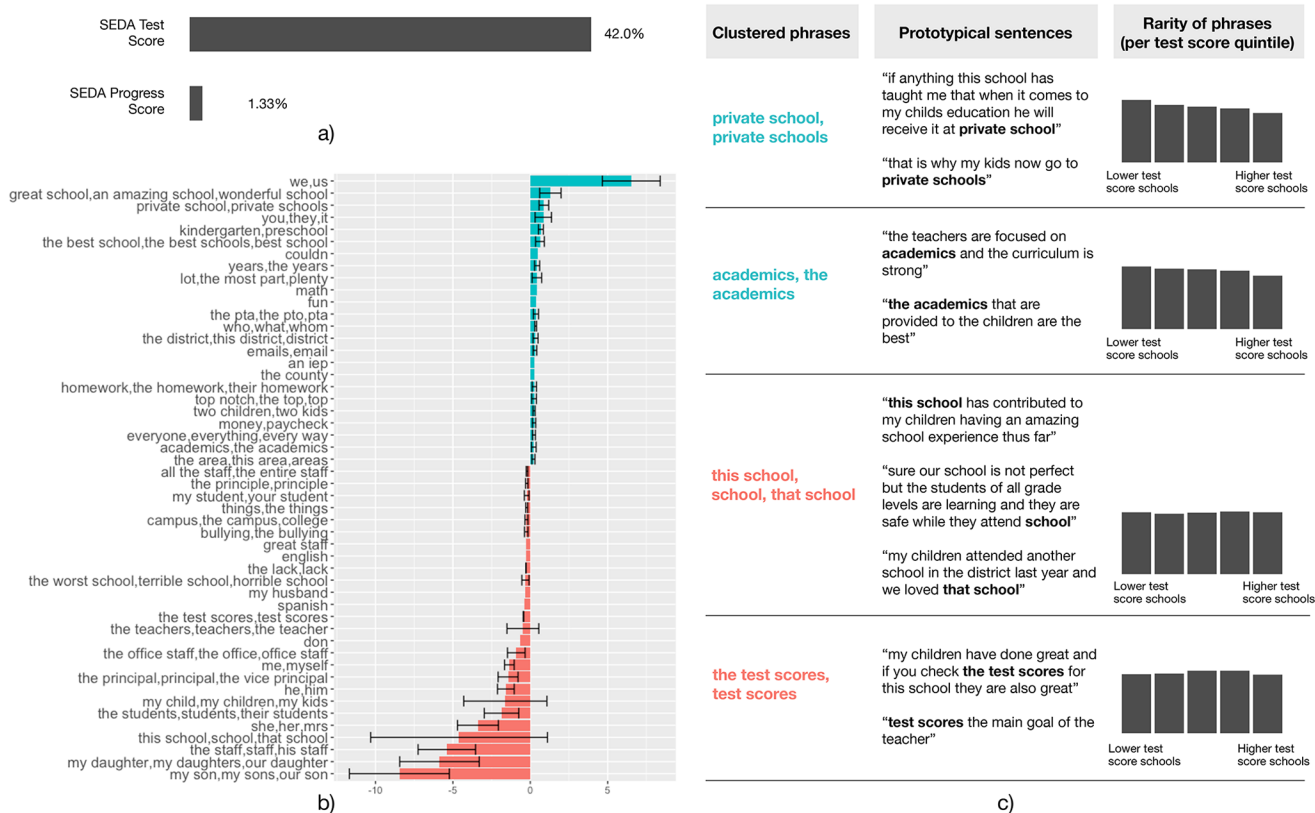


FIGURE 5. Outputs of our regression models when predicting school performance measures: (a) illustrates model performance when predicting test and progress scores, represented as the percentage improvement in mean squared error (MSE) compared with our mean-prediction baseline, and reveals virtually no correlation between review text and measures of student progress; (b) illustrates attributions for the top 50 noun phrase clusters with the highest average absolute value attributions inferred from the model predicting test scores, highlighting several phrases that appear to correlate with racial and income demographics of schools; and (c) offers additional context on how phrases are used by showing prototypical “average” sentences (middle column) and rarity (last column—each bar corresponds to reviews for schools in a given quintile of the test score distribution; taller bars indicate more rarity) for a selected group of high positively and negatively attributed phrase clusters. For (c), we selected clusters before computing their prototypical sentences and phrase rarity scores.

language with a mean squared error that is 42% better than the baseline (or, approximately, within 1 grade level of student achievement); however, its predictions for progress scores hardly improve over the baseline at 1.33%. This suggests that there is little correlation between the language in reviews and student progress scores, and that parents focus more on features of schools that track achievement gaps than school effectiveness when writing reviews. Indeed, this tendency reflects prior research highlighting how parents tend to value measures of peer quality like test scores, and not school effectiveness, when choosing schools for their children (Abdulkadiroglu et al., 2019). If schools respond to what parents’ value, such a preference could drive schools to selectively admit high-performing students and reduce pressures to improve how well they actually help students learn and grow (Rothstein, 2006).

Given the virtually nonexistent average correlation between parents’ reviews and school progress measures, we

turn our attention to better understanding which linguistic features correlate with a school quality measure that our models do have predictive value for: test scores. To do so, we apply IG to our validation set of approximately 5,000 schools in order to produce noun phrase attributions, clustering similar noun phrases using the scheme discussed in the Methods section. Figure 5b illustrates the top 50 phrase clusters (out of approximately 350 total), sorted by the absolute value of their average attribution. The cluster average is computed as a weighted average of the attribution scores for each constituent phrase (where the weights correspond to the number of distinct sentences the phrase occurs in within the validation set; the error bars depict ± 1 weighted SD). In most cases, clusters are defined by 1 to 3 phrases; in cases where there are more phrases constituting the cluster, we display the top three with the highest absolute attribution values.

Through qualitative inspection, we identify several phrases that reflect the racial and income demographics of

schools. For example, there is a positive correlation between phrases related to “private school” and the school’s test scores. Figure 5c shows prototypical, or “average” sentences from reviews that contain a subset of depicted phrases. We see that the prototypical sentences containing phrases about “private school” tend to come from families describing such schools as alternatives to their child’s current (public) school. From the last column in 5c, we can also see that the terms “private school” and “private schools” are generally rarer in the reviews of lower test-score schools compared with their higher test-score counterparts. Phrases relating to “the pta” (Parent Teacher Association) and “emails” are also positively predictive of test scores, perhaps reflecting the tendency for more affluent, nonminority parents to have the time and familiarity with schools to be involved and communicate regularly with teachers (Crozier, 2001; Hornsby & Lafael, 2011), and perhaps even a greater likelihood of being able to afford digital connectivity (Rideout & Katz, 2016). The large, positive attributions assigned to “we, us” in conjunction with the large, negative attributions assigned to phrases like “my son,” “my daughter,” and “my kids” may reflect a correlation between higher test scores and dual parent households. Indeed, dual parent households are known to be more prevalent among Whiter, more affluent groups (Pew Research Center, 2016), and we explore this possibility more later.

When we use adversarial machine learning methods (Pryzant et al., 2018) to remove the influence of racial and income demographics when predicting test scores from reviews, our model’s accuracy drops by approximately 50%. This further highlights the relationships between review language, test scores, and racial/income demographics (see the online Supplemental Appendix for more details on this procedure).

Linguistic Correlations With School Demographics

Fortunately, our analytical framework makes it straightforward to explore relationships between parents’ written reviews and school demographics more directly. To do so, we retrain two instances of our regression model to predict: (1) the percentage of students at schools who are White and (2) the percentage of students receiving free or reduced lunch. In these cases, our model achieves MSEs that are approximately 42% and 47% better, respectively, than what would be expected from a mean-prediction baseline. These MSEs are approximately equivalent to the performance of our model when trained to predict test scores, suggesting a similar magnitude correlation between the information contained in parents’ school reviews and schools’ test scores, racial composition, and students’ socioeconomic status.

Figure 6 depicts the phrase clusters, average attributions, and sentence prototypes for our models predicting race and income. Instead of language that explicitly describes the

racial and income demographics of schools, we once again find phrase clusters that implicitly reflect several well-documented racial and income-based disparities in U.S. K–12 education. For example, Figure 6c illustrates that phrases about “the pta” and “emails” are negatively correlated with the percentage of students receiving free or reduced lunch. This confirms our earlier intuitions that parents who discuss the PTA and emails or other communication with school personnel in their reviews are also likely a part of more affluent school communities. Indeed, prior work has highlighted links between PTAs and affluence, describing how active PTAs often mobilize to untap new funding streams that fill budget gaps and help expand the set of activities and resources available to students and their families (Murray et al., 2019). Parents at Whiter schools are also distinguished by reviews containing the phrase “small school,” perhaps reflecting a concern about class sizes that has been shown, when allayed, to drive parents from private back to public schools (Gilrairie et al., 2018). Once again, phrases like “we, us” positively predict the percentage of White and affluent at schools, whereas phrases containing “my” generally have the opposite correlations. As Figure 6b shows, the sentence prototypes for “we, us” appear to be written by parents who are writing on behalf of themselves and their spouse/partner. Further analysis reveals that phrases describing couples (“we, us”) tend to be rarer in the reviews of schools located in neighborhoods with a higher share of single parents, whereas those using the singular possessive pronoun “my” are more common in those neighborhoods—and rarer where dual parent households are the norm.

Reviews mentioning “special needs” and “an iep” (individualized education plan) are more associated with Whiter and more affluent schools. Interestingly, there is an active debate about the nature of racial and income-based disparities in special education. While some scholars argue that the higher percentage of minority and low-income students enrolled in special education reflects discriminatory practices on the part of educational institutions (O’Connor & Fernandez, 2006; Schifter et al., 2019), others suggest that, relative to their academic needs, not enough minority children in particular are enrolled in special education (Morgan et al., 2015; Morgan et al., 2017). Our results do not contribute to either side of this debate but do suggest that reviews referencing disabilities are more strongly associated with Whiter, more affluent schools. This could reflect barriers to advocating for special education that many low-income, minority parents face (Blanchett et al., 2009), concerns or stigmas about having their children classified as having special needs (Zuckerman et al., 2014), or several other factors. Conversely, special education could be so common a topic of reviews posted for schools with lower income and minority students, that it is more distinctive for parents in Whiter and higher income schools to mention it (and therefore, receive positive attributions by the model). To evaluate the latter hypothesis, we look

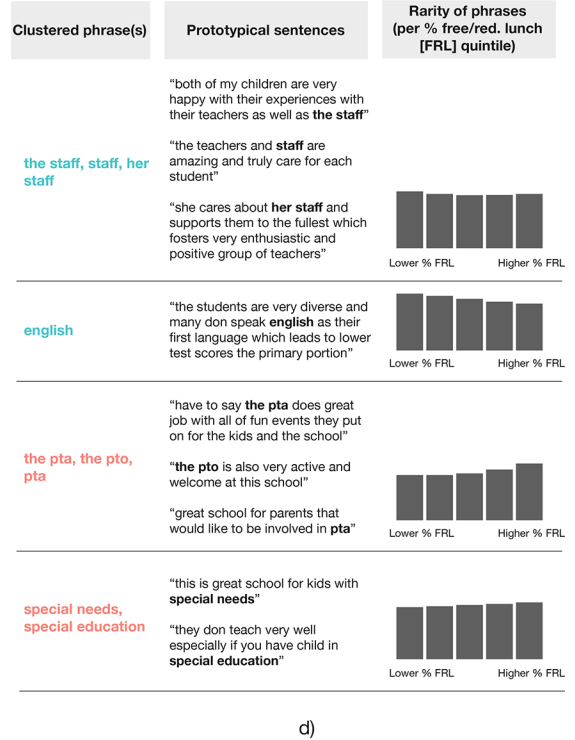
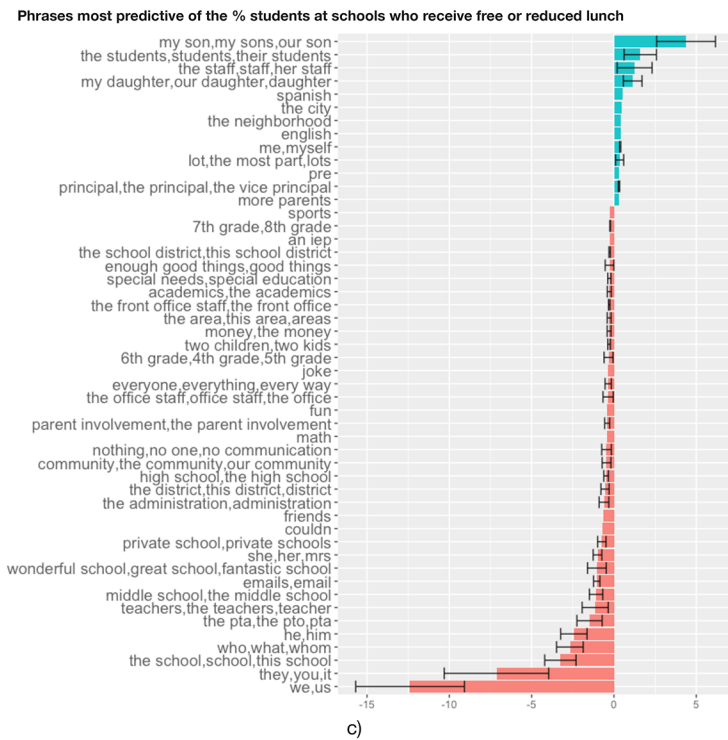
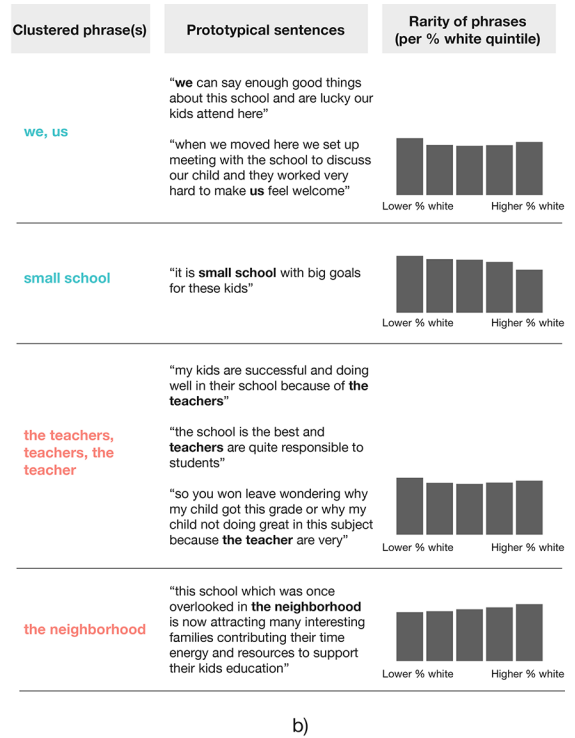
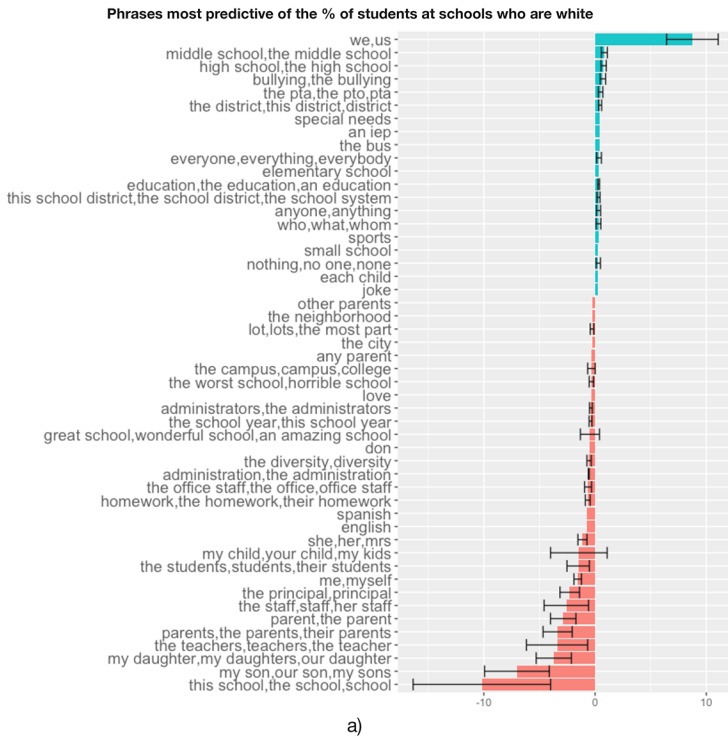


FIGURE 6. Attributions and additional context for phrases identified as being highly predictive of the percentage of students at schools who are White (a and b) or receive free or reduced-price lunch (c and d). Note. As before, attributions are computed by summing the attributions computed per phrase, per example (i.e., school), across the validation set.

at the phrase rarity chart in the last row of Figure 5d. The chart shows that the occurrence of the phrases “special needs, special education” are rarest in the schools serving a large

fraction of low-income students, which suggests that reviews for more affluent schools are actually *more* likely to mention such special education-related phrases.

There are also more speculative findings. For example, phrases about “bullying” and “fun” predict Whiter and more affluent student bodies, respectively. These correlations may reflect concerns about school climate, student comfort, and perhaps even mental health that are often of paramount importance to affluent parents (Jacob & Lefgren, 2005; Pew Research Center, 2015). Furthermore, phrases like “the neighborhood” and “the city” are predictive of less white, higher-poverty student bodies; conversely, “community” predicts lower poverty. One contributing factor might be the generally more positive opinions higher income families have of the places that they choose to live (Pew Research Center, 2015). This, combined with lower social capital that often plagues lower income populations (Small, 2009), might make it difficult for “neighborhoods” to feel like “communities.”

Finally, there are several phrase clusters that do not immediately suggest any relationships with race or income but still appear to be correlated with these demographics: for example, “who, what, whom”; “the school, school, this school”; and “the staff, staff, her staff.” It is possible that these phrases are simply more likely to be used by certain groups of parents than others, in certain semantic settings. The online Supplemental Appendix links to results for a longer list of noun phrases drawn from the data.

Discussion

Our results suggest that the availability of, and language contained within, parents’ school reviews reflect several well-documented racial and income-based disparities in K–12 education. This is to be expected, since the perceptions and actions of subsets of parents now positioned as consumers in education markets have long been linked to issues of school segregation and other inequalities (Reardon & Owens, 2014; Schneider & Buckley, 2002; Weis, 2016) despite them expressing preferences otherwise (Torres & Weissbourd, 2020). What, then, are the possible implications of our findings?

First, biases in who writes school reviews and what these reviews contain threaten to undermine the ability of school ratings websites to support choice-based systems that actually help reduce existing inequalities in education. For example, the fact that parents’ reviews contain little information that correlates with student progress suggests that parents who factor reviews into their decision-making processes are primarily consuming information that recapitulates schools’ test scores, rather than how well teachers and staff at the school may be helping students learn and grow over time. This threatens to perpetuate a tendency for parents to choose schools based on test scores and demographics instead of factors that correlate with student learning.

Relatedly, parents from certain racial or income groups may read disproportionately more about certain topics—like

the PTA, regular email communication with teachers, or even bullying—which in turn could shape what they expect, or even demand, from schools. Recent work from sociology highlights the “privilege dependence” of schools, or the extent to which schools and teachers rely on the volunteer contributions of involved parents, to the point of bending classroom rules for their children when necessary (Calarco, 2020). This suggests that certain parent subgroups often do have demands and that these demands are often met “quid pro quo” in exchange for much-needed classroom support. Furthermore, language use and structure has been shown to reinforce implicit judgements (Cimpian & Markman, 2011; Lewis & Lupyan, 2020); it is possible, then, that reading about schools through racially and socioeconomically coded reviews might reinforce stereotypes about how they operate, or even who attends them. The scenarios we present here are speculative, though arguably not far-fetched, especially considering the strong effects school reviews can have on parents’ perceptions of quality (Loeb & Valant, 2013).

Second, review content might actually influence the overall scores that ratings websites assign to schools. While GreatSchools does not factor parents’ reviews into their overall rating for a school, another popular ratings website (“Niche.com”) does (Niche.com Inc, 2021). Niche, in turn, advertises homes for sale or rent near schools. Incorporating subjective and demographically biased review content into schools’ ratings, then, threatens to fuel housing choices that amplify segregation by race and income. Indeed, some scholars argue that the availability of school ratings that primarily reflect students’ demographics have contributed to precisely this (Hasan & Kumar, 2019). These trends call for greater caution around which parents’ reviews are captured, and how these reviews are ultimately used to inform school choice.

Finally, our linguistic analyses reveal that some of the phrases parents use in their reviews offer a different lens on active debates in education without requiring deliberate, and often expensive, additional data collection. For example the positive correlations between phrases about special needs and the percentage of White, higher income students at schools does not resolve the debate about whether or not low-income, minority students are over or underrepresented in special education, but it does illustrate how the discussion about special education among (a biased group of) parents varies according to these demographics. This is only one example, and much more rigorous analysis and additional research are required to explore the extent to which review language might help resolve, or even anticipate, such educational debates.

There are several limitations in our work. For one, we analyze a limited subset of data from one of several school ratings platforms that parents might reference as they make decisions. Additionally, beyond offering sentence prototypes, we do not deeply investigate the sentiment and other

context describing how the noun phrases identified by our methods. Furthermore, our measure of student progress does not account for out-of-school factors that may affect student growth and achievement, and thus does not necessarily depict the school's causal effect on student learning. We also do not have access to demographic information for individual parents who post reviews, limiting our analyses to exploring the relationship between school-level demographics and review content. Finally, our analyses largely rely on measures of school quality defined according to performance on standardized tests, which offers a limited view of which factors are important to consider when selecting schools.

We believe these and other limitations open the door to several questions to explore in future work, including What are some underlying explanations for *why* reviews are virtually uncorrelated with measures of student progress? How might our results change if we knew the actual racial and socioeconomic demographics for individual parents posting reviews (instead of the current school-level averages)? How consistent are the patterns we've observed across other popular school ratings websites? What are the intentions and feelings behind the reviews parents post? What happens when we expand the set of outcomes we use to define "quality" to include measures grounded less in test scores: for example, school climate and satisfaction? Given the observed biases in which types of schools tend to receive reviews, how might school ratings websites capture and reflect a more inclusive cross-section of parents' perspectives about schools? How might these platforms solicit and highlight aspects of reviews reflecting information about school effectiveness, instead of simply recapitulating demographics? And in the rare circumstances when families do have a choice about the school their child attends, how might access, opportunity, and outcomes change if parents chose schools more on the basis of effectiveness than factors correlated with demographics? We hope this article offers both findings and an analytical toolbox that educational data scientists, social scientists, and practitioners might use to better understand and help reduce racial and socioeconomic disparities in education.

Acknowledgments

We thank GreatSchools for their help in accessing the data used for this study. We also thank Sarah Bana, Peter Bergman, Roger Levy, Aman Panjwani, and Martin Saveski for their valuable comments and ideas to help shape this article.

ORCID iD

Rebecca Eynon  <https://orcid.org/0000-0002-2074-5486>

References

Abdulkadiroglu, A., Pathak, P. A., Schellenberg, J., & Walters, C. R. (2019). *Do parents value school effectiveness?* (NBER

- Working Paper No. 23912). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w23912/w23912.pdf
- André-Béchely, L. (2005). *Could it be otherwise? Parents and the inequalities of public school choice*. Routledge.
- Angrist, J. D., Pathak, P. A., & Walters, C. R. (2011). *Explaining charter school effectiveness* (NBER Working Paper No. 17332). National Bureau of Economic Research. <https://www.nber.org/papers/w17332>
- Ball, S. J. (1993). Education markets, choice and social class: The market as a class strategy in the UK and the USA. *British Journal of Sociology of Education*, 14 (1), 3–19. <https://doi.org/10.1080/0142569930140101>
- Ball, S. J., & Vincent, C. (1998). I heard it on the grapevine: Hot knowledge and school choice. *British Journal of Sociology of Education*, 19(3), 377–400. <https://doi.org/10.1080/0142569980190307>
- Barnum, M., & LeMarr LeMee, G. (2019, December 5). Looking for a home? You've seen GreatSchools ratings. Here's how they nudge families toward schools with fewer black and Hispanic students. *Chalkbeat*. <https://www.chalkbeat.org/2019/12/5/21121858/looking-for-a-home-you-ve-seengreat-schools-ratings-here-s-how-they-nudge-families-toward-schools-wi>
- Bell, C. (2009). Geography in parental choice. *American Journal of Education*, 115(4), 493–521. <https://doi.org/10.1086/599779>
- Bergman, P., Chan, E. W., & Kapor, A. (2020). *Housing search frictions: Evidence from detailed search data and a field experiment*. Teachers College, Columbia University. <http://www.columbia.edu/~psb2101/HousingSearchFrictions.pdf>
- Blanchett, W. J., Klinger, J. K., & Harry, B. (2009). The intersection of race, culture, language, and disability: Implications for special education. *Urban Education*, 44(4), 389–409. <https://doi.org/10.1177/0042085909338686>
- Blei, D. M., & McAuliffe, J. D. (2008). Supervised topic models. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Proceedings of Advances in Neural Information Processing Systems 20. (NIPS2007)*. Neural Information Processing Systems. <https://papers.nips.cc/paper/2007/file/d56b9fc4b0f1be8871f5e1c40c0067e7-Paper.pdf>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M. E. A., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language models are few-shot learners*. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, & H. Lin (Eds.), *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Neural Information Processing Systems. <https://papers.nips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
- Calarco, J. M. (2020). Avoiding us versus them: How schools' dependence on privileged "helicopter" parents influences enforcement of rules. *American Sociological Review*, 85(2), 223–246. <https://doi.org/10.1177/0003122420905793>
- Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR.

- Quarterly Journal of Economics*, 126(4), 1593–1660. <https://doi.org/10.1093/qje/qjr041>
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633–2679. <https://doi.org/10.1257/aer.104.9.2633>
- Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, 82(2), 471–492. <https://doi.org/10.1111/j.1467-8624.2010.01525.x>
- Coenen, A., Reif, E., Yuan, A., Kim, A., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and measuring the geometry of BERT. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Proceedings of Advances in Neural Information Processing Systems 32* (NeurIPS 2019). Neural Information Processing Systems. <https://papers.nips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf>
- Cohen, J., McCabe, L., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, teacher education and practice. *Teachers College Record*, 111(1), 180–213.
- Crozier, G. (2001). Excluded parents: The deracialisation of parental involvement. *Race Ethnicity and Education*, 4(4), 329–341. <https://doi.org/10.1080/13613320120096643>
- Deming, D. J., Hastings, J. S., Kane, T. J., & Staiger, D. O. (2014). School choice, school quality, and postsecondary attainment. *American Economic Review*, 104(3), 991–1013. <https://doi.org/10.1257/aer.104.3.991>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 1, 4171–4186. <https://www.aclweb.org/anthology/N19-1423.pdf>
- Dobbie, W., & Fryer, R. G., Jr. (2009). *Are high quality schools enough to close the achievement gap? Evidence from a social experiment in Harlem* (NBER Working Paper No. 15473). National Bureau of Economic Research. <https://www.nber.org/papers/w15473>
- Dobbie, W., & Fryer, R. G., Jr. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics*, 5(4), 28–60. <https://doi.org/10.1257/app.5.4.28>
- Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership*, 37(1), 15–18.
- Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as data methods for education research. *Journal of Research on Educational*, 12(4), 707–727. <https://doi.org/10.1080/19345747.2019.1634168>
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574. <https://doi.org/10.1257/jel.20181020>
- Gilraine, M., Macartney, H., & McMillan, R. (2018). *Education reform in general equilibrium: Evidence from California's class size reduction* (NBER Working Paper No. 24191). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w24191/w24191.pdf
- Graetz, N., Woyczynski, L., Wilson, K. F., Hall, J. B., Abate, K. H., Abd-Allah, F., Adebayo, O. M., Adekanmbi, V., Afshari, M., Ajumobi, O., Akinyemiju, T., Alahdab, F., Al-Aly, Z., Rabanal, J. E. A., Alijanzadeh, M., Alipour, V., Altirkawi, K., Amiresmaili, M., Anber, N. H., . . . Herteliu, C. (2020). Mapping disparities in education across low-and middle-income countries. *Nature*, 577(7789), 235–238. <https://doi.org/10.1038/s41586-019-1872-1>
- GreatSchools.org. (n.d.). *GreatSchools ratings methodology report*. <https://www.greatschools.org/gk/ratings-methodology/>
- GreatSchools.org. (2020, December). *Greatschools.org analytics*. <https://www.similarweb.com/website/greatschools.org/>
- Hasan, S., & Kumar, A. (2019). *Digitization and divergence: Online school ratings and segregation in America*. SSRN. <http://doi.org/10.2139/ssrn.3265316>
- Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Sentometrics Research. <https://sentometrics-research.com/publication/72/>
- Hornsby, G., & Lafael, R. (2011). Barriers to parental involvement in education: An explanatory model. *Race Ethnicity and Education*, 63(1), 37–52. <https://doi.org/10.1080/00131911.2010.488049>
- Jacob, B. A., & Lefgren, L. (2005). *What do parents value in education? An empirical investigation of parents' revealed preferences for teachers* (NBER Working Paper No. 11494). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w11494/w11494.pdf
- Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for pytorch*. arxiv.org. <https://arxiv.org/pdf/2009.07896.pdf>
- Koretz, D. (2008, Fall). A measured approach: Value-added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator*. <https://www.aft.org/sites/default/files/periodicals/koretz.pdf>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behavior*, 4(10), 1021–1028. <https://doi.org/10.1038/s41562-020-0918-6>
- Loeb, S., & Valant, J. (2013). *Interim report for GreatSchools-Stanford study: Findings from YouGov Surveys on how Americans assess school quality*. Center for Education Policy https://www.greatschools.org/catalog/pdf/How_Americans_Assess_Schools_Quality-Jon_Valant_Stanford_2013.pdf
- Lovenheim, M. F., & Walsh, P. (2017). *Does choice increase information? Evidence from online school search behavior* (NBER Working Paper No. 23445). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w23445/w23445.pdf
- Lubienski, C., Gulosino, C., & Weitzel, P. (2009). School choice and competitive incentives: Mapping the distribution of educational opportunities across local education markets. *American Journal of Education*, 115(4). <https://doi.org/10.1086/599778>
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National*

- Academy of Sciences*, 117(48), 30046–30054. <https://doi.org/10.1073/pnas.1907367117>
- McInnis, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), 205. <https://doi.org/10.21105/joss.00205>
- Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2017). Replicated evidence of racial and ethnic disparities in disability identification in U.S. schools. *Educational Researcher*, 46(6), 305–322. <https://doi.org/10.3102/0013189X17726282>
- Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher*, 44(5), 278–292. <https://doi.org/10.3102/0013189X15591157>
- Murray, B., Domina, T., Renzulli, L., & Boylan, R. (2019). Civil society goes to school: Parent-teacher associations and the equality of educational opportunity. *Russell Sage Foundation Journal of the Social Sciences*, 5(3), 41–63. <https://doi.org/10.7758/rsf.2019.5.3.03>
- Musset, P. (2012). *School choice and equity: Current policies in OECD countries and a literature review* (OECD Education Working Paper No. 66). OECD. <https://doi.org/10.1787/5k9fq23507vc-en>
- Niche.com Inc. (2021). *About niche's k-12 rankings*. <https://www.niche.com/k12/rankings/methodology/>
- O'Connor, C., & Fernandez, S. D. (2006). Race, class, and disproportionality: Reevaluating the relationship between poverty and special education placement. *Educational Researcher*, 35(6), 6–11. <https://doi.org/10.3102/0013189X035006006>
- Pew Research Center. (2015). *Parenting in America*. <https://www.pewsocialtrends.org/2015/12/17/parenting-in-america/>
- Pew Research Center. (2016). *On views of race and inequality, blacks and whites are worlds apart*. <https://www.pewsocialtrends.org/2016/06/27/on-views-of-race-and-inequality-blacks-and-whites-are-worlds-apart/>
- Pryzant, R., Shen, K., Jurafsky, D., & Wagner, S. (2018). Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, 1615–1625). Association for Computational Linguistics. <https://www.aclweb.org/anthology/N18-1146.pdf>
- Reardon, S. F. (2017). *Educational opportunity in early and middle childhood: Variation by place and age* (Working Paper No. 17–12). Stanford Center for Education Policy Analysis. <https://cepa.stanford.edu/sites/default/files/wp17-12-v201803.pdf>
- Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., Chavez, B., Buontempo, J., & Disalvo, R. (2019). *Stanford Education Data Archive* (Version 3.0). Stanford Education Data Archive. <https://stanford.io/2SpttFV>
- Reardon, S. F., Kalogrides, D., & Shores, K. (2018). *The geography of racial/ethnic test score gaps* (Working Paper No. 16–10). Stanford Center for Education Policy Analysis. <https://cepa.stanford.edu/sites/default/files/wp16-10-v201803.pdf>
- Reardon, S. F., & Owens, A. (2014). 60 years after Brown: Trends and consequences of school segregation. *Annual Review of Sociology*, 40(1), 199–218. <https://doi.org/10.1146/annurev-soc-071913-043152>
- Reardon, S. F., Papay, J. P., Kilbride, T., Strunk, K. O., Cowen, J., An, L., & Donohue, K. (2019). *Can repeated aggregate cross-sectional data be used to measure average student learning rates? A validation study of learning rate measures in the Stanford Education Data Archive* (Stanford Center for Education Policy Analysis Working Paper No. 19–08). Stanford Center for Education Policy Analysis. <https://cepa.stanford.edu/sites/default/files/wp19-08-v201911.pdf>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Rideout, V., & Katz, V. S. (2016). *Opportunity for all? Technology and learning in lower-income families*. Joan Ganz Cooney Center at Sesame Workshop. https://www.joanganzcooneycenter.org/wpcontent/uploads/2016/01/jgcc_opportunityforall.pdf
- Roberts, M. E., Stewart, B. M., Tingley, D., & Airoidi, E. M. (2013). *The structural topic model and applied social science*. <https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf>
- Roda, A., & Stuart Wells, A. (2013). School choice policies and racial segregation: Where white parents' good intentions, anxiety, and privilege collide. *American Journal of Education*, 119(2), 261–293. <https://doi.org/10.1086/668753>
- Rothstein, J. M. (2006). Good principals or good peers? Parental valuation of school characteristics, Tiebout equilibrium, and the incentive effects of competition among jurisdictions. *American Economic Review*, 96(4), 1333–1350. <https://doi.org/10.1257/aer.96.4.1333>
- Schifter, L. A., Grindal, T., Schwartz, G., & Hehir, T. (2019). *Students from low-income families and special education*. The Century Foundation. <https://tcf.org/content/report/students-low-income-families-special-education/>
- Schneider, M., & Buckley, J. (2002). What do parents want from schools? Evidence from the internet. *Educational Evaluation and Policy Analysis*, 24(2), 133–144. <https://doi.org/10.3102/01623737024002133>
- Small, M. L. (2009). *Unanticipated gains: The origins of network inequality in everyday life*. Oxford University Press.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic attribution for deep networks* [Conference session]. The 34th International Conference on Machine Learning, Sydney, New South Wales, Australia.
- Torres, E., & Weissbourd, R. (2020). *Do parents really want school integration?* Harvard Graduate School of Education. <https://mcc.gse.harvard.edu/reports/do-parents-really-want-school-integration>
- Weis, L. (2016). Reading and producing research across boundaries that so often divide. In A. R. Sadovnik, & R. W. Couhlan (Eds.), *Leaders in the sociology of education* (pp. 253–275). Brill.

- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Zhang, Z., Ye, Q., Law, R., & Lijun, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, 29(4), 694–700. <https://doi.org/10.1016/j.ijhm.2010.02.002>
- Zuckerman, K. E., Sinche, B., Mejia, A., Cobian, M., Becker, T., & Nicolaidis, C. (2014). Latino parents' perspectives of barriers to autism diagnosis. *Academic Pediatrics*, 14(3), 301–308. <https://doi.org/10.1016/j.acap.2013.12.004>

Authors

NABEEL GILLANI is a PhD candidate at the MIT Media Lab. His research explores computational social science, education, media, and inequality.

ERIC CHU is a PhD candidate at the MIT Media Lab in the Laboratory for Social Machines group. His research focuses on machine learning and computational social science.

DOUG BEEFERMAN is a research scientist at the MIT Media Lab in the Laboratory for Social Machines group. His research focuses on natural language processing and computational social science.

REBECCA EYNON is a professor at Oxford Internet Institute, University of Oxford. Her research explores sociology of education and inequality.

DEB ROY is a professor at the MIT Center for Constructive Communication, with research interests in media analytics and design.