REACH

National Center for
Research on Education
Access and Choice

Is a Picture Worth 51 Million Words?

A Text Analysis of Public User Reviews of Schools

Douglas Harris, Jamie Carroll, Debbie Kim,
Nicholas Mattei, & Olivia Carr

Technical Report
Published November 8th, 2024

Is a Picture Worth 51 Million Words?
A Text Analysis of Public User Reviews of Schools

Douglas N. Harris
Jamie M. Carroll
Debbie Kim
Nicholas Mattei
Olivia G. Carr

Tulane University
November 8, 2024

Abstract: Massive online user review platforms, with their star ratings and text reviews, are reshaping the information available for consumer and public service decisions. We study the leading K-12 schooling platform, GreatSchools, applying machine learning (Natural Language Processing, NLP) to 600,000 reviews that encompass the vast majority of the nation's traditional public, charter, and private schools (84,009 schools in total), supplemented with qualitative analysis of a subsample of reviews. Encompassing more than fifty million words of text, our initial analysis pre-specified eight broad topics and 27 sub-topics and coded review words into these categories. We find that parents write the vast majority of reviews and tend to write more about School Staff and School Culture than students. More generally, text reviews vary in important ways across user types (parents, students, teachers, principals), school sector (traditional public, charter, and private schools), grade level, and demographics. The partial correlation between topics and star ratings also differs across user types and sectors. Taken together, these results suggest that user reviews are less useful than they appear and less useful than with other kinds of products. Our analysis points to design features that might improve their usefulness. The variation in content and value of the reviews also has methodological implications as it shows how NLP can complement qualitative research methods with such large volumes of text.

Acknowledgements: This research was carried out under the auspices of the National Center for Research on Education Access and Choice (REACH) based at Tulane University, which is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C180025 to The Administrators of the Tulane Educational Fund. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, GreatSchools. or any other organization. We would like to thank Ethan Macy-Cushman and Molly Shields who provided valuable research assistance, as well as anonymous reviewers.

Author Information: Douglas Harris is Professor and Chair of the Department of Economics at Tulane and director of REACH (dharri5@tulane.edu). Jamie M. Carroll is the Associate Director of Research for REACH at Tulane University. Debbie Kim is a Senior Research Scientist in the Education and Child Development group at NORC at the University of Chicago. Nicolas Mattei is Associate Professor of Computer Science at Tulane University. Olivia Carr is freelance educational consultant.

I. Introduction

The effects of government services depend not only on government policy, or even implementation, but also consumer choices. One key example is families choosing among publicly funded schools. Tiebout-style choice, where housing is bundled with schooling, has been a topic of longstanding interest (Domina et al., 2021; Hoxby, 2000; Owens & Candipan, 2019), but, increasingly, more direct "school choice" programs, including intra- and inter-district choice, charter schools, and vouchers are proliferating. At least one-third of school-age children attend schools other than their assigned public schools (Harris et al., 2017) and this was before COVID-19 and expansion of universal voucher programs took hold (Cordes et al., 2023).

Information is one key factor shaping schooling decisions. Families rely heavily on information from their social networks, especially family and friends, when making schooling decisions (Altenhofen et al., 2016; Goldring & Phillips, 2008). This "word of mouth" source is especially useful for learning about more subjective and hard-to-measure elements of schooling. For example, families value availability of after-school care in elementary schools and extracurricular activities in high school (Harris & Larsen, 2023) and may also have strong preferences over school culture (e.g., approach to student discipline), whether schools adopt more progressive versus conservative approaches to education, and how schools prepare children for adulthood. It may be difficult to measure these factors in a standardized way. On the other hand, word-of mouth depends on personal social networks that are stratified by family income (Corcoran & Jennings, 2019; Teske et al., 2007).

Another common information source is websites of individual schools and those created by state departments of education and non-profit organizations. These sites provide information about more standardized metrics, such as student test scores and high school graduation rates, as well student demographics, all of which parents seem to think are important (Glazerman & Dotter, 2017; Harris & Larsen, 2023; Schneider & Buckley, 2002). They also have the advantage of being freely accessible to all.

Online user review platforms represent a more recent development. These allow consumers (and others) to describe and evaluate their experiences in schools in more subtle and flexible terms. Consumers can rate products overall (e.g., 1-5 stars) and enter open-ended text reviews that explain the reasons behind their overall ratings. With regard to schooling, one of the most widely known online user review platforms is GreatSchools. When typing in a school name into any major search engine, the GreatSchools webpage for that school is often one of the first links listed. One reason for its popularity is that, when families view homes for sale or rent on sites such as realtor.com and Zillow, they also see information on the nearest schools--data provided by GreatSchools.

Platforms such as GreatSchools, as well as a newer version called Niche[1], combine elements of these standardized and non-standardized sources into a "one stop shop." Like other online tools, they also avoid the problem of unequal social networks since the platforms are freely accessible to everyone, free of charge and regardless of social capital and personal connections. In addition to star ratings and text, the GreatSchools site also includes, and increasingly emphasizes, standardized data on student demographics and student outcomes from federal and other data sources–much like the information available on school/state websites.

The utilization of online user platforms is likely to continue growing in schooling, not only because of the increased use of the internet but because of expanding school choice policies that give families more options. At least one-third of school-age children no longer attended their neighborhood schools before COVID (Harris et al. 2017) and the "pandemic pod" phenomenon has gone mainstream in the form of microschools (McShane, 2024). Fourteen states have now adopted universal school vouchers/ESAs (Tarnowski, 2024) and many of these are being designed with their own user platforms so that families can shop not only for schools but also tutors and standalone academic and extracurricular activities. Information shapes schooling choices (Corcoran & Jennings, 2019; Valant & Weixler, 2022). Therefore, to choose wisely among these myriad options, families are going to need more and better information.

Unlike the literature from medical user platforms (Hong et al., 2019; J. Liu et al., 2020) and higher education platforms (Gregory, 2011), however, the study of such K-12 school user platforms is nascent. We therefore focus on several core questions that can guide the design of user platforms, inform the likely successes and failures of school choice policies, and point toward a rich future research agenda. First, what aspects of schooling are most commonly raised in the GreatSchools reviews of schools? Using a combination of Natural Language Processing (NLP) and traditional qualitative analysis, we find that the most commonly discussed topics are Instruction/Learning and Overall Quality, followed by School Culture, and School Staff. The other, less discussed topics are Resources and School-level Features. If we believe that people write about what they care about, these results are informative about what parents, students, and other school stakeholders believe to be important, complementing evidence from survey (Burgess et al., 2015), interviews (Bell, 2009; Kleitz et al., 2000), internet searches (Schneider & Buckley, 2002) and revealed preferences (Glazerman & Dotter, 2017; Harris & Larsen, 2016, 2023; Beuermann et al., 2023). One key observation from our analysis is that parents value a wider variety of school characteristics than is typically available for research purposes.

Second, how do the reviews vary by user type (parents, students, educators, and others), school-level demographics (race/ethnicity and income), and sector (traditional public, charter,

---

[1] Unlike GreatSchools, Niche relies more on students to review schools (and colleges) and uses incentives to encourage a large number of reviews. As we explain, GreatSchools is driven more by parent reviews and adult needs.

and private schools), and overall school rating level? [2] Since parents write the vast majority of reviews, the overall results mostly reflect the topics that parents talk about. However, students write proportionally more than parents about school Overall Quality, School Safety, and Extracurricular Activities. Given the ongoing expansion of school choice policies, it is also important to consider differences across schooling sectors. Compared with traditional public schools, reviews of both charter and private schools include more about School-level Features and less about School Staff. Charter reviews include more about Instruction and Learning and private school reviews include less about Physical Environment than reviews of TPS. We offer possible reasons for these differences and what this might mean for the usefulness of online reviews for improving decisions.

Third, what do users value most in schooling?[3] To answer this, we regress the star rating on a vector of indicators describing the topics discussed. Again, we find considerable differences by user types and school sectors. For example, positive ratings of TPS are associated with text discussion of School Staff and School-level Features, but the opposite is true in private schools. We are not aware of prior research that has compared how different educational stakeholders view and value education in different ways.

These findings, taken together, suggest that text reviews, while of great interest to readers (Valant & Newark, 2020), are not as informative as they might seem. The fact that reviewers write about different topics, and seem to value different elements of schooling across user types and sectors, makes the information difficult for users to process, interpret, and use. This is partly rooted in the fact that schooling is an unusual context with "exceptionally ambiguous output" (Hess, 2004). In theory, open-ended data collection might allow users to capture that ambiguity in a way that is useful to other consumers. But in practice, user platforms seem to assume that consumers are sophisticated qualitative researchers who have a great deal of time on their hands to think not what each reviewer is saying and take into account the background of the reviewer (parent, student, etc.) and the sector they have chosen. Users also have to take into account the length of reviews, which the data also suggest varies across user and school types.

Perhaps the most obvious challenge in interpreting user reviews is selection bias. While the number of reviews is very large (N>600,000), the *share* of people that provide user reviews (i.e., the equivalent of the "response rate") is very low and likely self-selected. In online

---

[2] GreatSchools actually creates two separate ratings: the "review rating" is the average of the individual 1-5 star ratings from the user reviews, while the "GreatSchools rating" is based on student test scores, high school graduation, advanced coursework offerings, college entrance exams, and "equity" of outcomes. In the present study, we focus only on the *individual* review rating, which allows us to focus on the relationship between the individual star rating and the same individual's review text.

[3] As discussed in more detail later, we did not take into account the particular linguistic structure of the sentences to attempt sentiment analysis. For instance, "The teachers are not good" and "The teachers are good" may be interpreted the same way by some of our methods. But sentiment analysis is prone to error (Wankhade et al., 2022). Instead, we use the rating each reviewer assigned to the school to infer sentiment, e.g., when teachers are mentioned in a 5-star we implicitly assume the reference to teachers is positive.

communities, studies have generally found that around 90% of users are "lurkers" who read or consume but do not post, 9% use the service a little by posting one or two things, and 1% account for the vast majority of online content (Nielsen, 2006).[4] While the overall distribution of GreatSchools ratings mirrors the distribution of letter grades given by surveys of random samples of parents (West, 2022), this is likely not the case for individual schools from which parents are choosing.

In the Conclusion, we discuss possible solutions that might make user review platforms more informative, including combining surveys with NLP and using NLP to create customized, school-level syntheses of the text reviews. We also discuss some implications for school choice policies, especially in light of the recent and fast expansion of universal school vouchers and education savings accounts (ESAs) that are being accompanied by their own, new user review platforms.

Our final contribution of this work pertains to research methods. We show how such a large volume of unstructured text can be analyzed using NLP, a form of machine learning.[5] We also show how NLP methods compare with qualitative methods. We took random samples of reviews and team members coded them without knowledge about how NLP coded the reviews. We find that our NLP strategy tends to "over-code" reviews into topics because of the mechanical way in which it assigns words to topics. Qualitative coding also has the advantage of being more accurate and nuanced. For example, in the qualitative analysis of what different reviewers write about in the second research question, we find that students wrote more about their direct experiences with teachers in classrooms and were less likely to compare their current schools with other schools, while parents write more about how well schools prepare students for life after high school and make more cross-school comparisons. As another example, in the analysis of the third question, the qualitative analysis suggests that positive reviews (4-5 stars) included more verifiable information and discussed the schools' academic excellence and instruction. In contrast, negative reviews relied more on vague adjectives and described tastes and feelings. These conclusions would not have arisen from the NLP analysis alone. Moreover, NLP seems to require a similar amount of time to implement as qualitative analysis of random samples of text. Therefore, in the process of explaining how NLP can be used for research purposes, we also caution researchers to consider the more nuanced questions that can be answered with qualitative methods. We are more optimistic about the use of NLP for summarizing information provided by user review platforms than about their use as a research method.

---

[4] This trend has been found several times though recent studies on Twitter users find that closer to 25% are active users while 75% are passive (Antelmi et al., 2019; Nielsen, 2006).

[5] Others have used this technique to analyze text data of other kinds of user reviews (Dave et al., 2003; B. Liu, 2012).

The next section provides more background about GreatSchools, how user reviews are entered, and the data themselves, including the question of representativeness. This is followed, in Section III by discussion of our analytic methods (NLP and regression analysis with a supplemental qualitative analysis). Section IV discusses our results and Section V concludes and considers, for example, what these findings imply about the design of school user review platforms and the strengths and weaknesses of school choice policies that depend on good information.

II. Background: User Reviews and GreatSchools

In the early years of the internet, content was mainly generated by those designing and managing websites. More recently, in web 2.0, the internet has come to be dominated more by user-generated content. User reviews, in particular, have become an important and popular way for individuals to learn about consumer goods, travel destinations, contractor services, and more. User review platforms are also increasingly integrated within purchasing systems so that consumers can shop, find information, and make purchases all at once. The rise of Amazon is perhaps the best and most widely used example of this revolution in consumer purchasing.

Mining online user reviews for information has been a popular subject in computer science since the earliest days of the internet and, with the growth from star ratings to full text reviews, this task has become more complicated yet. As in the current study, prior research has centered on understanding and summarizing the content of these reviews for purposes of product recommendation and marketing (Dave et al., 2003; B. Liu, 2012).

While these reviews can be the source of important and relevant information, there are a number of limitations with online ratings, including pay-for-play reviews and social influence bias in which reviews may be written by those who have a personal stake in the outcome or be influenced by prior reviews by others (Aral & Walker, 2014). Online reviews tend to be written by individuals who either have a strong positive or strong negative opinion of the subject, leading many websites including YouTube to abandon a 5-star system for simple thumbs-up/thumbs-down (Rajaraman, 2011). Another issue is fake or spam reviews, either positive or negative, which can come from malicious actors or even positive "ratings drives" on behalf of a particular product or service (Lim et al., 2010). Within the computer science communities that focus on predicting and interpreting user reviews, this "non-missing-at-random" assumption is widely documented (Marlin & Zemel, 2009; Schnabel et al., 2016).

These issues with online information also apply to user reviews of schools. GreatSchools' stated mission is to help all parents make more informed decisions about where to send their children to school. Approximately 43 million people visited GreatSchools in 2018. Some people connect to the site through internet searches of specific school names. Others connect indirectly through housing search websites, such as Zillow and realtor.com, which show the names of the public schools that are zoned for the given neighborhood as well as nearby private and charter schools.

Prior research on GreatSchools ratings and reviews highlights the importance of websites like this to provide information to parents about schools, but also the limitations. As policy changes occurred related to school choice, searches on GreatSchools increased as parents tried to find more information about their new school options (Lovenheim & Walsh, 2018). The availability and content of reviews are correlated with the racial/ethnic and socioeconomic composition of schools, more so than measures of student academic growth (Gillani et al., 2021). On the other hand, there is some evidence that the content of reviews predicts future changes in schools' composition and test scores (Li et al., 2023). However, interpreting any of these studies requires a deeper understanding what different reviewers write about and what this implies about what they value.

Figure 1 shows the information people see on the GreatSchools site when they click on a specific school, including a breakdown of the academic rating,[6] and demographic information about students who attend the schools. Toward the bottom of the site is the average 5-star rating and specific text-based user reviews–and the opportunity for users to add additional reviews. In this sense, information begets information. People might go to the site to learn about some other school and, in the process, be prompted to input information about other schools with which they might be familiar.

These user reviews are the data that are the basis for the present project. Table 1 provides a small sampling of the user text reviews that consumers see when they go to the GreatSchools site. While these examples were selected at random, some patterns are evident that are reinforced by our later data analysis. In the sampled 1-star reviews, families talk about how their children are bullied and treated harshly by students and teachers. These low reviews are also longer and more specific. This may suggest that families' negative impressions are driven by a handful of specific individuals (e.g., bullies) or specific events (e.g., a dangerous event at the bus stop).

The reviews with high user ratings, in contrast, are more general and refer more to the environment and extracurriculars (e.g., growing vegetables and dance), though one of the higher reviews also refers to the quality of teachers and the science lab. Overall, in both the high and low-rated excerpts, reviewers discuss topics of instructional quality, peer relationships, school safety, school leadership, and facilities. This reinforces that educational stakeholders care about a wide range of school factors.

Our dataset consists of all of GreatSchools' text review data from 2002-2019. Table 2 describes our dataset by reviewer and school type. The number of reviews and number of schools

[6] The current version of the academic rating includes four components: the Student Progress Rating or Academic Progress Rating (growth on test scores), College Readiness Rating, Equity Rating (achievement gaps), and Test Score Rating.

with reviews are roughly consistent with the shares of all schools that exist in each category.[7] The top row provides information for the entire data set, i.e, the corpus. Figure 2 displays the overall distribution of the number of words per text review.

Parents write the overwhelming majority of reviews (75%), followed by students (8%), teachers (3%), and principals (0.2%). Schools average only 8 reviews in total (among those who have any reviews). On average, public schools with more students receiving free or reduced-price lunch and those with lower achievement scores have fewer reviews than those with more students from high income families and with higher achievement levels. Note that these totals include all types of users (parents, students, educators, and others) and are summed across all years. In contrast, a school with 400 students would also have roughly 800 parents and 30 School Staff each year, and probably 5-10 times as many over a 17-year period, which is the length of the present analysis. This implies a "response rate" of less than 0.1%

For our analysis, we are interested in a set of specific subgroups or aspects of data.[8] These include: user (parent, student, teacher, principal), school sector (charter, traditional public, private), and user rating (number of stars given by reviewer, 1-5).

III. Methods

III.A. General Methods

Text reviews found online, such as GreatSchools' reviews, are often referred to as *unstructured data* (Jurafsky & Martin, 2024) in that there is only freely written text that is not coded for particular features or items of interest. In this project, we apply techniques from natural language processing (NLP) to infer useful aspects of user feedback. Within computer science, NLP is a broad term that encompasses many different goals and techniques, each of which can be used for a variety of problems, from extracting facts from text to generating completely new text e.g., ChatGPT (Achiam et al., 2024). One of the most popular NLP applications involves analyzing a large corpus (set) of documents (e.g., user reviews) and *automatically* finding groups of relevant words that are indicative of various categories of review and/or finding sentiments associated with these categories. These tasks are typically called aspect based sentiment classification (ASBC) (Brauwers & Frasincar, 2022), and we use some of these techniques in our work. The key advantage is that, because NLP is automated, these techniques can perform these tasks quickly and accurately in massive amounts of text data.

In all of our analysis we follow the same standard steps in preparing the text data for analysis. Within the NLP world there are many ways to identify what constitutes a *keyword* that is worthy of analysis or indicative of a topic or category. We will elaborate on the specific

[7] The total share of charter, private, and TPS schools in the CCD are roughly 7, 10, and 83 percent, respectively.
[8] In the computer science literature, these subgroups are called "domains."

techniques for generating these keywords in the subsequent sections, but we start with the analysis steps that are common to both research questions. The following steps are fairly standard across most NLP-based analyses (Bird et al., 2009; Manning et al., 2008).

Step 1: Clean and Process the Data. We processed all the text data with an automated spell corrector, removing non-English words as well as words containing non-alphabetic characters and removing stop words (commonly used words such as "the" or "and" that provide little to no information about content or meaning). We then normalized the text by converting it to all lower case and removing all punctuation, as the structure of the sentences themselves is not necessary for our analysis. Using a list of common place and person names we removed proper names and replaced them with <PERSON> and <PLACE> because we are most interested in the nature of the discussion contained in the text, not individuals. More details on all these steps including programming packages used are in Appendix D.

Step 2: Identify Keywords. We then identified a set of keywords. For the *frequent keywords* used in Research Question 1 we take the 1,000 most commonly used words from the corpus. We chose to focus on single words rather than n-grams because it reduced the error rate and made the labeling task more straightforward. For Research Questions 2 and 3, we also build a set of *log-likelihood keywords* using a log-likelihood ratio test that measures the relative frequency with which a word is used in one dataset (i.e., set of reviews) as compared to another dataset. For this analysis, we compared the entire set of user reviews to a subset of user reviews that differed on one of several aspects (e.g., private schools versus all schools). In this way we are able to identify words that are used "unusually frequently" in reviews with one aspect as compared to the whole dataset. With private schools, for example, we expect to see words such as "tuition" or "prayer" occur more frequently in private school reviews and the log-likelihood estimate allows us to test that and find these words automatically.

Step 3. Apply Topic Codes to Keywords. We developed and applied a set of topic codes that represents the range of subjects and concerns reviewers discussed or expressed about schools. We based the initial set of codes on what we know matter to parents from research and existing parent surveys (e.g., CCSR 5Essentials Parent Survey, Panorama Family-School Relationships Survey). As we worked through the keywords, coding with this initial set, we flagged the keywords that did not fall into the topic codes. Then, we inductively created additional topic codes (e.g., Physical Environment) and subtopic codes (e.g., Location as a subtopic of Physical Environment) to categorize the subjects or concerns that remained to be coded.

This is one part of the process where NLP becomes useful. We could have just stopped above, relied entirely on the manual/human processes that are common in qualitative research, and proceeded to identify the most common words. Instead, we took the additional step of building a model of how words are used specifically in the GreatSchools dataset, i.e., a *language* *model.* Language models use machine learning techniques to identify how specific words are used in a specific dataset, e.g., how your text prediction on your phone adapts to the way *you* use language. For example, the word "work" could take on many meanings such as homework, teachers' jobs, or coordination across people or groups. In this case, the language model learns how users of the GreatSchools site use language.

We created a language model using the popular Word2Vec method, which is the best method for word associations and language models. This method involves creating a prediction model that given a context, i.e., the words around a target word, predicts a target word. The fundamental idea is that "you know a word by the company it keeps," and if we can train a model to be very good at this "fill in the blank" type problem, then we can use that model to understand what words are similar to each other.

We can break the process of creating a language model into steps. First, we create a vector representation of each word. If there are 10,000 words (after the above cleaning steps), then this vector has 10,000 dimensions. When a given word is observed, it is coded as "1" and all the others are coded as "0." This input vector is the first layer of the model.

Each of the 10,000 words is also a target word whose use we are trying to predict, but it is too cumbersome and uninformative to use only one of each of the 10,000 words to predict each of the others. Instead, we choose a window, usually about five words before and after a word, of nearby words that establish the *context*. The goal of our model is to predict a word to "fill in the blanks". Given the five words before and after some target word, we encode that set of words (context) as above where there is a 1 in the vector for every word that is present. We want the output of our model to predict the target word, i.e., the one we held out that fills in the blank. Note that in this task, the target word is not included as input, but rather we use the target word to evaluate the model and improve it. The word "work" in the above example might often be close to the word "study," so when we see "study," there's a reasonable chance "work" will be next. This step creates a number of word pairs we used in the next step, described below.

The above step is the first layer in a 3-layer neural network for the prediction task, the first layer is the ~10,000 dimension vector mentioned above. The second layer is smaller, causing the neural network to compress the representation into a smaller space, and we use the fairly standard 100 dimensional real valued vector (similar to 100 principal components) for this layer. The third layer (also called the output layer) is again a 10,000 dimension vector that corresponds to the input layer. This middle compression step is similar in spirit to principal components analysis (PCA) or factor analysis, and also shares a similar purpose: to reduce the dimensionality where there are a large number of variables, to both ease computation and have the network learn words that go together. In this case, our neural network takes an input of 10,000 dimension vectors and encodes them in 100 dimensions. The idea is that this smaller vector space forces the model to learn, and hence to represent in this 100 dimension vector, how words are used similarly.

The first step in building the language model (called "training" in machine learning) is to take all our context, target word pairs as described above and split the data into 80% as the training set, and the remaining 20% to be used as a test (validation) set. We also select the number of dimensions for our second layer, which we set to 100 as discussed above. Using the training set and starting with a random set of parameters for the model, we adjust those parameters using a loss function to adjust those parameters when the model does not fill in the blanks correctly. We continue this predict / adjust cycle over the training data until the model converges, i.e., the size of the updates to the parameters is suitably small and the overall accuracy of the prediction task is suitably high. We then use the test set to ensure that the model is able to handle out of sample predictions, i.e., prevent overfitting. We can then use the second, 100-dimension layer, to understand what words are used similarly.

The main purpose of all of this, in our case, is to help us manually code the ~1,000 keywords in the GreatSchools data into our topic list and, if necessary, to adjust the topics themselves. The language model allows us to identify the words that are most similar to each keyword. Continuing the above example, the language model tells us that the semantic neighbors of "work" are: 'cooperate', 'collaborate', 'communicate', 'connect', 'nosed', 'collaborating', 'tasks', 'banded'. These words have more to do with how the students or staff are working, so this word is categorized into School Culture. In contrast, the closest semantic neighbors of "homework" we get: ''hw', 'worksheets', 'assignments', 'busywork', 'classwork', 'worksheet', 'quizzes', 'schoolwork', 'papers', 'assignment'. This is clearly being used to discuss the tasks students are assigned so we categorize this word into Instruction & Learning. So, even though work and homework might seem similar, NLP distinguishes between them based on how they are actually used in context.

The fact that we used NLP in this manner means that our method is semi-supervised. It is supervised in the sense that we started by manually coding the keywords into categories. But it is only semi-supervised because we used the NLP to assist us in the coding. We could have, instead, let the model identify the topics entirely on its own, i.e., an unsupervised model.[9]

Step 4: Annotate Reviews with Codes. After we manually coded the keywords into mutually exclusive topics, we then used NLP to code each review as containing a particular topic if the review contained one or more of the keywords associated with a given topic. There are some subtle issues with how we applied these codes, e.g., the difference between teach and teaches, which we expand on in Appendix D. Note that our coding of reviews does not distinguish the number of times a topic is discussed in a particular review, only which set of

---

[9] In a related paper, we used data unsupervised models to predict future changes in test scores and school demographics (Li et al., 2023).

topics were discussed. At the end of this step we are left with a large spreadsheet where each unit of observation (row) is a review and the columns are indicator variables for whether various topics and subtopics appear in the reviews (based on the above process) and various additional data on school types and so on. a. We then use this spreadsheet for all our analysis below.

To be clear, throughout our analysis involves simple frequencies of the number of reviews that contain a given topic, not the number of times the topic arises (which could be multiple times in a given review). Also, note that the unit of analysis in all cases is the review. This means that schools with more reviews are given more weight. It also means that each review can have any combination of topics; it can be coded as including all the topics/subtopics, none of the topics/subtopics,[10] and everything in between. Later, we provide descriptive statistics regarding the frequency of these topics and various patterns.

III.B. Method for Research Question 1: What topics do reviewers write about?

For identifying the *frequent keywords* we manually applied topic codes to the 1,000 most frequently used words. Using the semantic neighbors technique described above to obtain a set of words used "most like" a particular keyword across all the reviews we were left with 457 keywords that were sorted into one of eight Level-1 topics: School-level Features, Instruction and Learning, Unclassified, School Culture, School Staff, Resources, Overall Quality, and Physical Environment.[11] Each word was additionally sorted into one of 25 Level-2 labels that are more specific versions of the Level-1 labels as shown in Table 3.

III.C. Method for Research Questions 2 and 3: How do the topics differ across user types, ratings, school demographics, and sectors?

For identifying the *log-likelihood keywords* we used the relative frequency of each keyword for reviews from each aspect group (school type, school quality, and review quality) compared with the entire dataset. Specifically, we measure relative frequency using the log-likelihood (LL) method. We split our text into seven sub-corpora based on the domains of school type (three types), school quality (two types), and review quality (two types). We then identified the unusually frequent words in each subcorpus. Using a p-value of 0.01, statistically distinctive keywords were selected for each subdomain. Additional details on this calculation can be found in Appendix D.

---

[10] This is quite rare. Only 0.5% of the reviews code all the topics as zero.

[11] The reason for the reduction in the number of keywords from 1,000 to 457 requires some explanation. In the "work" example above, we showed how the NLP helped us to categorize that particular word. But we also decided in this step whether the initial keywords deserved to be in any category at all. In some cases, the NLP helped us make this determination while, in other cases, we made this decision ourselves. For example, words like "combination", "else", and "light" were taken off the list since their nearest neighbors indicated these words did not add any meaning to the review.

Since the method is different from that of the simple frequency method we used in the first research question, the number of keywords also differs. For the purposes of our preliminary analysis, we examined the top ~1,800 keywords from which we selected 773. Many keywords were irrelevant to our research questions and were excluded. For example, though we tried to filter out school names, some made it through because the words are not only used as proper names (e.g., "green," "hill," "forest," and "grove"). In any event, formal names of schools do not address our research questions, so we excluded them from the analysis.[12] We then manually placed each of the 773 keywords into a single Level-1 and single Level-2 topic as discussed above. Table 3 displays topics, subtopics, and example keywords from each subtopic. In summary, we apply NLP using a simple frequency approach to answer the first question about what topics and sub-topics reviewers discuss; and we use the LL method to answer the second research question about what differentiates the various sub-groups of reviews.

Finally, we use these data to regress each category and subcategory on reviewer and school characteristics using linear probability models. We merge the GreatSchools coded reviews with the National Longitudinal Schools Database (NLSD), which includes characteristics about public and private schools, including the school level (elementary/middle school versus high school) and demographic characteristics (diversity index, proportion of students receiving free/reduced-priced lunch, and urbanicity). The main analyses examine whether categories were more or less likely to appear in reviews written by different users and for different kinds of schools. We also examine the relationship between the number of stars reviewers assigned to schools and the categories to understand whether patterns differ between positive and negative reviews.

Given that reviews vary in their length and the number of topics that appear, we also performed robustness checks that consider these elements of reviews. Specifically, we ran the same models with the outcome as the *proportion* of topics discussed in the review that were a specified topic (the category indicator divided by the total number of categories discussed) or controlling on the number of words in a review. For analyses focusing on differences by reviewer type, we also performed school fixed effects models to isolate differences in types of reviewers from differences in the schools themselves. We mostly share results that were robust to all specifications, but also note where the results differ by method in some circumstances.

III.C. Qualitative Analysis

Analyses of unstructured text are typically carried out using qualitative methods that manually search the text for words and phrases, or codes, developed deductively from prior research and inductively from themes found in the data (Auerbach & Silverstein, 2003). Previous studies of user review text reviews have often involved manual, qualitative coding of

[12] We noted earlier that we replaced school names with <SCHOOL> to minimize this problem, but some school names made it through this filter.

small samples of reviews (Hovy et al., 2015). Such analyses provide great insight into the data, but can be time- and labor-intensive, thereby limiting the scope of analyzable data.

To supplement the NLP analysis performed above, we performed additional qualitative analyses on random subsamples of reviews for two purposes. First, to validate the coding done by the NLP process described above we manually coded reviews into the Level-1 categories to check for consistency and examined the meaning of frequent words in the corpus we were unable to classify. Second, to gain a deeper understanding of patterns we observed among student and parent reviews, we developed additional themes to investigate that did not or could not emerge from the NLP-based analysis.

In all of the subsamples, we started with a random sample of 500-575 reviews that were in the 25th-75th percentile on length (i.e., 30-107 words). Limiting on word length helped to ensure that there would be enough words to carry out a qualitative analysis, while also filtering out reviews that might be unnecessarily long and difficult to code.

For each research question, two coders reviewed the same 20 reviews using the initial coding scheme. Any discrepancies in coding were discussed until a consensus was reached on how each review should be coded. The coders then refined the coding scheme to ensure reliability The reviews were randomly divided up between the coders to complete the coding process. Coders continued to meet on a regular basis to discuss any new questions about the coding scheme. The details of these analyses vary by research question and are discussed in more detail in the findings section and in Appendix B.

IV. Findings

IV.A. Findings: What topics do reviewers write about?

We begin by summarizing the percentage of reviews that mention a particular topic, along with the most common keywords pertaining to that topic, using the simple frequency or count method in Table 3. This is followed by a deeper analysis of the most commonly discussed topics and how we can be confident our methods are capturing GreatSchools review topics accurately.

IV.A.1. Main Topics and Validity of Analysis

The most commonly discussed topics, shown in Table 3, were Overall Quality, which includes words (mostly adjectives) related to positive or negative judgment of the school, (87.07%) and Unclassified (87.52%), which includes words used often in reviews that are mostly vague nouns that didn’t fit into the other categories (e.g., students and families) with almost identical frequency. School Staff, which includes words that describe teachers and other adults working in the school, (76.05%), School Culture, which includes words that describe feelings about the school environment and relationships, (71.72%), and Instruction and Learning, which includes words about the ways students are taught and what they are taught, (63.64%) were also discussed often, but less so than the first set of topics. School-level Features, which includes words related to the focus of the school or type of school, (42.74%), Physical Environment, which includes descriptions of the school building and location, (36.08%), and Resources, which includes descriptions of the extracurricular activities and school offerings, (33.03%) were discussed least often.

The fact that GreatSchools users discussed all of the topics we identified fairly often provides some face validity to our method. The topics that we organized each key word into represent important areas of education and appear to match the items reviewers of schools tend to focus on. As an additional validity check, we applied qualitative analysis to a small sample random of 500 reviews. For this analysis, we provided coders with only the review text and asked them to decide which of the topic areas the review covered (see Appendix B for more details). The simple correlation between the NLP and qualitative coding ranges from +0.24 to +0.75 across topics. Overall, The NLP codes a higher percentage of reviews as containing each topic. For example, among all the cases where the qualitative method coded a review as mentioning instruction, 89 percent were also coded for instruction under the NLP. However, 56 percent of the reviews coded as not mentioning instruction by the qualitative coding did have this code in the NLP. This mismatch occurs when words being tagged by NLP are used in reviews in different ways than the NLP categorization assumes. For example, a review including the phrase "one of a kind" was tagged as being about School Culture by the NLP because it included the word "kind." Another review was coded as being about Instruction and Learning by the NLP for including the word "learning," when in fact that word is part of the name of the school included in the review: "Kids Learning Center."[13]

This disjoint between the NLP and qualitative analyses is not ideal from a validity and reliability standpoint and could create at least two problems. First, it means the frequencies of topics reported in Table 3 are inflated. Second, our analyses for the subsequent research questions might lead to attenuated or otherwise-biased relationships with school and user types. Even if the over-coding in NLP is effectively random, as appears to be the case, this measurement error will attenuate the coefficients in the regression estimates (discussed later). A worse possible scenario is that the coding errors are correlated with the covariates in ways that bias them in unknown directions. We return to this issue again later.

IV.A.2. Additional Analysis of "Unclassified" Words

[13] As noted earlier, we filtered out school names and replaced them with <SCHOOL>. This is an example where that strategy does not work because the name does not contain the word "school." In other cases, a lack of capitalization and/or misspelling could allow some proper names to slip through.

Many of the words used frequently in reviews did not fit into any of the pre-specified lists and we sought to learn more about this. We applied traditional qualitative analysis to a small random sample of 575 reviews that had coded an Unclassified word (regardless of what other topics were coded). We highlight the words that were labeled as Unclassified in the review and coders determined what these words meant in context. Two-thirds of these sampled reviews contained words coded as Unclassified that actually pertained to pre-specified categories discussed above (40 percent of these were in School Culture). Although NLP did not code these particular words within the pre-specified categories, in most cases, these reviews were still coded by NLP as containing the pre-specified category using other words in the review. School selection, or the process by which students are admitted to schools, was the commonly discussed new category (14 percent of sampled reviews). In only 4 of 575 Unclassified reviews, did the qualitative analysis code a review as covering a topic that the NLP did not code as such.

The fact that the qualitative analysis of Unclassified reviews did not turn up more entirely new topics might not seem surprising given that we started with topics that seemed important to families in prior research, and that some of the topics were selected based on what we observed in a small sample of reviews at the beginning of the project. Still, the fact that the qualitative analysis did not turn up more or different topics was not preordained. In the early steps, we identified the most common words and coded them into topics, but it could have been that even the relatively common words were so infrequent, as a percentage of the total, that most reviews did not contain any of them. This highlights the different approach taken in the qualitative analysis. We could have discovered new topics in the qualitative analysis, e.g., because some topics were discussed using different words, each of which was individually infrequent. This is not what we find. Rather, we show that our NLP process can reliably identify meaningful elements of user reviews.

This section therefore presents both substantive findings on the first research question and helps set the stage for the rest of the analysis. We show that some topics predominate in GreatSchools reviews, while others are less common. Moreover, we provide justification for examining patterns in the NLP-based review codes to answer the last two research questions.

Another key observation from this section is that parents and students, as the main reviewers, care about many different elements of schooling. This reinforces what others have pointed out about revealed preference type studies that focus just on test scores and student demographics are missing much of what is important. While Overall Quality is the most common topic, and test scores are an important component of that, only 25% of respondents mention Evaluation and most of these are not explicitly about test scores. Extracurriculars, for example, also seem to be important (Harris and Larsen, 2023), to name just one.[14]

IV.B. Findings: How Do the Topics Raised Vary Across Review Types?

IV.B.1. A Basic Framework for Interpreting Variation in Responses

In this subsection and the next one, we study the variation in reviews across groups. To help interpret these results, we created a basic theoretical framework for interpreting the results. We leave it to future researchers to create a more fully fleshed out theory, but the basic framework below proved useful in the discussion that follows.

We argue that there are three main reasons why user reviewers might vary across users. Broadly, this framework is based on what economists would call "heterogeneous preferences."[15] This has three dimensions that are relevant here. First, some reviewers might view some characteristics positively and others might view the same characteristics negatively. For example, teachers might view strict discipline more positively while students/parents might view strict discipline negatively. In this case, we would expect all of these groups to write about discipline in their reviews at similar rates, but we will see different signs on the regression coefficients across the user types when regressing the star ratings on the topics. We call this the "taste" hypothesis.

Different users also might view the topics the same way (same tastes), but weigh them differently. For example, all user types might want strict discipline, but teachers might not consider this as important as students and parents do. In this case, we might expect teachers to write more often about discipline than students and parents and to give higher rating to schools when discussing discipline. This is a variation on the taste theory but it yields a different prediction: that the relationship between ratings and topics goes in the same direction for all school and user types, but to varying degrees. We call this the "emphasis" hypothesis.

Different users also believe in different purposes for education and vary in what they think schools should offer and what constitutes a "good" school performance on any given dimension. Users may discuss topics more or less often depending on what they expect schools provide students. For example, students may care more about the extracurricular activities and courses available in a school than their parents, who may believe the purpose of school is preparing their children for the future. In this case, we would expect students and parents to both

[14] Beuermann et al. (2023) find that parents also value schools that reduce criminality and teen pregnancy. However, it is difficult to connect these outcomes to the words that families use to describe their schools. They may be related, for example, to school discipline and the values they instill.

[15] We also considered incorporating what Harris and Larsen (2023) talk about as the role of household constraints in driving how much families might weigh factors like after-school care. However, our analysis cannot address this because GreatSchools, and user review platforms generally, do not capture information about respondents' demographics or locations.

discuss different topics more often and rate schools differently based on how easily they are impressed by what schools do in a given dimension. We call this the "expectations" hypothesis.

These hypotheses lead to somewhat different hypotheses in the analyses that follows. In the frequency analyses (IV.B.2), when one group writes more about a given topic, this likely suggests varying levels of emphasis or expectations. The analyses that link the topic frequency to the star ratings is more informative because it provides a clearer sense of direction about tastes. In particular, when we see a topic come up more often in 5-star reviews for one group and see the opposite for another group, this suggests that the groups want opposing school characteristics. In both analyses, it is difficult to distinguish between expectations and emphasis, however.

It is even more difficult to distinguish between the various theories when we examine results by school type. This is because users have made active choices to associate with particular schools, and those schools may be more or less effective on the various dimensions that users care about. For example, suppose that families select charter schools because they are stricter or more academically focused. In that case, if charter user star ratings are positively with Instruction and Learning, then this could be either because families have different preferences or because the schools themselves are more effective. We cannot distinguish these interpretations with these data.

The above issue when comparing across school types, not user types, but another issue arises with user types. In particular, different groups might have the same preferences have the same schooling circumstances, but write about them in different ways. For example users with higher levels of education might be able to more clearly articulate their views and be able to write more, in total, than others in the same amount of time.

Again, we do not see this as a full-fledged theoretical framework, nor have we exhausted the range of issues that arise when applying it, but we think this discussion does help interpret some of the results by user and school type. These theories capture an element of truth and we discuss the potential implications of this for school choice and how we view the different sectors in our conclusion.

IV.B.2. Differences in the Frequency of Topics

In this section, we examine the topics discussed by sub-groups, Research Question 2. Recall that these results are based on the *relative* frequency of topics, which means that some of the topics may not have been discussed very often overall.

Since we have multiple dimensions of subgroups, all of which are correlated, we carry out this analysis of the keywords created from the LL method using a simple regression framework. Table 4 provides estimates from linear probability regressions where the dependent variable is whether a topic is mentioned at least once in a review (that is, if multiple keywords from the same topic show up in a single review, then we still count this as only one reference). These variables are regressed on a vector of school/user subgroups: user type, school type, school level, a racial diversity index[16], percent free-lunch-eligible, and urbanicity. Each column is a separate regression. The unit of analysis is the individual review. Robust standard errors are clustered at the school level and reported in parentheses.

We also include the regression results of when the length of the review (number of words) is the dependent variable (see Table 4, column 1). Who writes the most? Principals write an average of five and a half more words than parents (excluding stop words), while teachers write the least. Reviews of charter schools are much longer than reviews of traditional public schools. For example, the coefficient on charter in the first column of Table 4 indicates that charter school reviews average 13 more words per review than TPS reviews. Racially homogeneous, elementary/middle, and non-urban school reviews are shorter than racially diverse, high school, and urban school reviews. The latter might reflect the more complex social dynamics that arise in diverse schools.

Next, we turn to the topics discussed. Each coefficient indicates the probability that a topic comes up for that school/user type, relative to the omitted category. It is possible for a given subgroup to discuss all the topics more/less than the omitted category and we generally see both positive and negative coefficients across topics for each subgroup.

We view the covariates as being meaningful if they are both statistically significant and have magnitudes greater than 0.010, so that the given subgroup is at least one percentage point more likely to mention the topic than the omitted group. In addition, we only discuss coefficients that are robust to alternative specifications. Below, we discuss the results for subgroups of interest separately, recognizing the results all come from a single set of regressions where we are also controlling for the other factors at the same time (each column is a separate regression). We also mention a few patterns that emerged from our analyses of subtopics (Tables 5A-5D), but we describe those patterns and analyses in more depth in Appendix C.

- *Results by User Types*. For user groups, parents are the omitted/reference category.
    - Principals write more than parents about most topics. For example, the coefficient in the top row and third column in Table 4 means that principal reviews are 3.9 percentage points more likely than parents to mention Physical Environment. However, it is noteworthy that principals write very little about School Staff, which can include both the staff they manage and themselves.

[16] The racial diversity index is in the range of [0,1]. It is defined as one minus the sum of squared racial enrollment shares. A school that is 100% of a single racial group has a diversity index of 0, while a school that is equally split among the groups has the highest possible diversity. (This converges to 1.0 as the number of racial groups goes to infinity.) We also estimate a version using conventional, separate racial subgroup categories.

- Students write less than parents about almost everything, but write the most about Overall Quality, which is dominated by general, adjective evaluations. This could reflect either that students have more limited ability to articulate their views or that they have so many experiences and dimensions to consider that it would be difficult for anyone to express everything they have to say. Student experiences with school are much more extensive than parents. Students write especially little about School Culture and Physical Environment.[17]

- Teachers write less than parents on Resources and Unclassified topics, but more than parents on School Culture and Instruction and Learning. In the analyses of subtopics within Resources (Table 5D), we find that parents wrote the least about extracurricular activities and the most (other than principals) about course offerings.

- *Results by School Sector*. Here, TPS is the omitted/reference category.
    - Charter reviews focus especially on Instruction and Learning and School-level Features compared with TPS, which may reflect that they often have specific themes (e.g., the arts). Analyses of subtopics (Table 5B) reveal that these patterns are driven by charter school reviewers writing more about Instruction and Curriculum than TPS reviewers. [18] In the analyses of School Culture subtopics (Table 5C) we find that charter school reviews discuss Student Discipline and Interpersonal Relationships more than reviews of TPS.

    - Both charter and private school reviews are less likely to discuss School Staff than reviews of TPS. Private school reviews also include much more about School-level Features and less about the Physical Environment than reviews of TPS. As shown in Table 5C, the School-level Feature that seems to dominate private school reviews is whether the school is religious. The subtopic analyses in Table 5A suggest that, within Physical Environment, differences are driven by more of a focus on the location and facilities in TPS reviews than reviews of private schools.

    - The analysis of subtopics within Overall Quality (Table 5B) shows that charter school reviews discuss Evaluation more than reviews of TPS and reviews of both

[17] In the main models, students write less than parents about Instruction and Learning, but we view the conclusion as indeterminate in this case because the result is not robust to considering the number of topics included in the review (see appendix table A.4.1).
[18] Although charter school reviews appear to discuss School Culture more in the main models, this relationship is not statistically significant when we condition on the number of categories discussed and the number of words in the review (see appendix tables A.4.1 and A.5.1), potentially because charter school reviews are longer in general.

private and charter schools discuss how these schools prepare students for the future more than for TPS.[19]

- *Results by School Level*. Elementary/Middle schools are the reference/omitted category. Reviews about high schools are less likely to discuss School Staff and more likely to discuss Resources and School-level Features. This could reflect the increasing specialization of school activities as students reach higher grade levels. For example, extracurricular activities are School-level Features that become more important as students get older (Harris & Larsen, 2023). In fact, analyses of subtopics within Resources (Table 5D) finds high school reviews are much more likely to include discussion of Extracurriculars/Electives, consistent with (Harris & Larsen, 2023).

Returning to the heterogeneous preferences hypotheses, recall that this type of analysis can really only tell us whether there seems to be heterogeneity in emphasis of expectations across these measurable groups. The above discussion provides ample evidence of this.

The results by school sector also shed light on the different tastes, emphases and expectations of reviewers for TPS compared with charter and private schools. While many areas of the country allow choice for TPS, it is often the parents and students that select charter or private schools that are opting out of their residentially assigned school. This choice may be linked to the School-level Features and Instruction and Learning in charter and private schools, as reviewers of these schools write much more often about these topics. We return to this discussion in the following results section to examine how topics are associated with the ratings given by different kinds of reviewers.

IV.B.3. Additional Model Specifications for NLP Analyses

We performed a number of robustness checks on the findings discussed above. In particular, given the differences in review length shown in Table 4, we re-estimated the models controlling for review length (see Appendix Table A.5.1). In addition, given that some reviews may discuss many topics, we also ran models that predicted the *proportion* of topics in the review for each review category (See Appendix Table A.5.1). While many of our results are robust to the alternative specifications, some of the coefficients for the indicators of racial and economic composition of the schools switched signs (with statistical significance in both cases). Thus, the results by demographics are not robust to alternative regression specifications.

It is difficult to say which specification is preferred because this depends on the reasons why some reviews mention more topics than others, which are difficult to ascertain. For the shorter reviews, it might be that the reviewers: (a) only care about the small number of things

[19] Although private school reviews appear to discuss Overall Quality less than TPS in the main models displayed here, this relationship is not statistically significant when we consider the number of categories discussed (see appendix table A.4.1).

they mention; (b) care about many things fairly equally, but ran out of time or just picked one or two; or (c) had a hard time articulating some topics more than others. If (a) is the reason for the shorter reviews, then the results in Table 4, which do not control for review length, are arguably preferred. But if the reason is (b) or (c), controlling for review length seems important. We have no evidence on the reasons why review length varies and leave this for future research.

We also included models with alternative covariates to test robustness. First, we replace the racial diversity index with separate variables for the percent of students in each racial group (Table A1). We find that reviews about schools with a higher proportion of Black students include more about Instruction and Learning and less about the Physical Environment. Reviews about schools with a higher proportion of Hispanic students include less about the Physical Environment and School-Level Features. This was true for the models described above that consider review length and proportion of topics discussed.

Another appendix table (Table A2) shows the relationship between the topics and academic achievement levels. (This is not in the main tables because all private schools and many TPS and charter schools have missing achievement. In the appendix, we impute the missing values for TPS and charter schools, but leave private schools out of the analysis.) Schools with higher test scores have more text about all the topics, with the exception of the Physical Environment.

In addition, we re-estimated the models adding school fixed effects, as shown in appendix table A.7. Again, in the prior section, we only discussed results that were robust across all of these various specifications

IV.B.4. Qualitative Analysis as a Robustness Check on NLP

In section IV.A.1, we noted some differences between the qualitative and NLP coding. To test whether this might explain any of the above results for the second research question, we re-estimated the regression models using only the qualitative analyses and then compared these to Table 4.

The regression analyses with the qualitatively coded data show magnitudes and directions very similar to what we have already reported. Of the 128 coefficients reported, only four show any meaningful differences. For example, with qualitative coding, charter school reviews appear to discuss Overall Quality more than public school reviews, whereas there is no relationship with the NLP coding. In no cases, do the coefficients reverse with statistical significance across the two methods. These results can be found in Appendix Table A.6.

IV.B.5. Deeper Qualitative Analysis of User and School Types

To gain a better understanding of how parents and students discuss schools differently, we performed a qualitative analysis of a subset of schools, examining the subjectivity of reviews, the proximity to the classroom experience, the focus of academics and caring, and comparisons to other schools. Specifically, , we identified a sample of 14 charter schools, 17 private schools, and 17 TPS, each of which had at least 5 parent reviews and 5 student reviews (if the numbers were larger we took a random sample to achieve exactly those numbers of parent and student reviews). This allowed us to compare student and parent reviews of the same schools, and to do so across different school types.

We find that students are more likely than parents to discuss experiences with teachers in classrooms, while parents are more likely to emphasize the importance of preparing students for their lives after school. Parents are also more likely than students to compare the current school to other schools. Table 6 displays examples of reviews by parents and students from the same school, but emphasizing different aspects of the school.

The qualitative analysis also reinforced and deepened our understanding of the prior differences by school type. We find that reviews of charter schools are much more likely to give personal examples but also much less likely to talk about their schools as caring. Some research has found that charter school students report their teachers care less about them in surveys (Carroll et al., 2023). Reviews of traditional public schools were less likely than charter and private school reviews to make comparisons with other schools, likely because students and parents in charter and private schools have made more active decisions to leave traditional public schools.

IV.C. Which Topics are Associated with Higher User Ratings?

If the goal is to understand what users think is important about schools, then knowing the topics they write about is a good place to start. We would expect people to write about what matters most to them. But we can go further by comparing the topics users write about with how they rate schools. The topics that come up in 5-star reviews might not be the same as in 1-star reviews and this could tell us more about what users think is most important.

To address this question, we regressed the star rating on a vector of indicators for whether a topic was raised in the associated review. We again estimate via Ordinary Least Squares (OLS) for ease of interpretation. This analysis also restricts the sample because only 85% of reviews have a star rating.[20]

[20] Principals, in particular, rarely attach star ratings to their reviews; only 5.54% of principal reviews include ratings, compared to approximately 85% for all other reviewer types. In addition, there are differences by sector; about 16% of reviews of TPS are missing a star rating, compared with 14% of reviews of charter schools and 11% of reviews of private schools. Less than one percent of reviews also have a rating but no text. These are also dropped from all aspects of the analysis.

The first column of Table 7 shows that higher rated reviews are much more likely to include an Overall Quality indicator, and also are more likely to mention school Resources and Instruction/Learning. To be clear, this analysis does not distinguish positive from negative Overall Quality indicators (e.g., "wonderful" from "terrible"). This suggests that, when users like their schools, they tend to think about that in broad, vague terms without reference to specifics. Reviews with low star ratings are much more likely to mention the Physical Environment.

The subsequent columns in Table 7 provide the same results but limit the sample to specific user types and sectors. It could be that what makes people prefer charter schools, for example, is not the same as other types of schools. Likewise, it would be that the factors that make parents like schools are not the same as their children.

Results by User Type. Some of the results are consistent across parents, students, and teachers.[21] All three groups give worse reviews when they raise the Physical Environment and better reviews when they mention Overall Quality and Resources. Teachers give better reviews when discussing School Culture. Parents give better reviews when discussing Instruction and Learning. When students bring up an Unclassified topic, their reviews tend to be much worse (the opposite is true for parents and teachers).

Based on our theoretical framework of heterogeneous preferences, we see evidence for both the taste and emphasis hypotheses. Parents and students seem to have a different valence/taste than teachers on School Culture. How stakeholders view schools therefore seems to be a case of "where you stand depends on where you sit." On the other hand, with Physical Environment and Overall Quality, we see no clear difference across user types—both the signs and magnitudes are indistinguishable across user types.

Results by School Sector. Across traditional public, charter, and private schools, we see starker differences. The relationship between star ratings and School Staff mentions are all positive in TPS but negative in private and charter schools. Also, the negative relationship between Physical Environment and rating is almost three times larger in public schools than private schools (charter schools are in between), consistent with the emphasis hypothesis. The positive relationship between Instruction/Learning and rating is almost twice as large in private schools than charter schools. In only a few areas do we see similarity across sectors: reviews mentioning Resources and Overall Quality have higher quality ratings while those mentioning the Physical Environment are associated with negative ratings. The results across all the topics are largely unchanged when we add in controls for user type, school sector, and other variables shown earlier in Table 4.

Returning to the theory, we note that the interpretation of patterns by sector is more complicated than for user types. Another theory is that people are motivated to both write a

[21] We exclude principals from this analysis because very few principals give star ratings, as described above.

review and mention certain factors when schools perform below/above expectations on those factors. In that case, the above relationships can be viewed as reflecting differing initial expectations. For example, it could be that students or parents have higher expectations—or set a higher bar-–on some topic than teachers do. This complicates matters because the predictions from this "expectations" theory cannot be clearly distinguished from the taste or emphasis theories.

Recall that it is more difficult to isolate the reasons behind differences by different results across school types. With this in mind, consider again the disjoint between the Physical Environment/quality relationship, which is strongly negative for TPS, but only moderately so with private schools.[22] This could be because: (a) TPS parents have higher expectations for the Physical Environment; (b) TPS parents place greater emphasis on the Physical Environment; or (c) TPS actually have worse Physical Environments. Distinguishing these explanations, especially between (c) and the others, would require much more data on schools' actual Physical Environment, which includes the sub-topics of Location, Facilities, and Building Quality.

IV.C.2. Qualitative Analysis of Positive and Negative Reviews

We carried out additional qualitative analysis to better understand the differences between positive and negative ratings. Positive reviews were more likely to include objective facts and to discuss academics and life after school (e.g., college). These reviews tend to include the vague adjectives that are in the Overall Quality category ("This school is great!"), but then describe specific elements related to their positive rating and language. Negative ratings focused more on emotional language and opinions. We saw no statistically significant differences between positive/negative ratings by whether reviewers told stories involving classroom experiences, gave personal examples, or talked about how caring adults were in the school. Table 8 displays some example reviews highlighting these patterns.

V. Discussion and Conclusion

One purpose of this analysis is to better understand what people talk about in user reviews, which provides a window into what they see as most important. In this respect, our work builds on many prior studies of what families prioritize when choosing schools (e.g., Glazerman & Dotter, 2017; Harris & Larsen, 2023). The most common topics in GreatSchools reviews are Unclassified topics that we could not categorize, Overall Quality, and School Staff. Since Overall Quality often entails broad descriptions and adjectives, this reinforces that GreatSchools reviews tend to be vague. Reviews that discuss Overall Quality are also especially likely to have higher star ratings, along with reviews that discuss Resources and Instruction and Learning. In contrast, reviews discussing Physical Environment are closely associated with

[22] In this discussion, we focus on parents since they represent the vast majority of reviews in every user type.

negative ratings, along with reviews that discuss School Culture and School Staff, but to a lesser extent.

Our analyses by user and school types contributes to our understanding of how parents, students, teachers, and principals view their schools differently and how this may also vary in traditional public, charter and private schools. Not surprisingly, users tend to write about what they know best. Principals and teachers write more about modes of instruction because they are the ones who both determine and implement these aspects of the schooling process. Parents, in contrast, write little about Instruction and Learning, probably because they do not observe it. But parents write more than students about how the reviewed school compares with other schools because they hear more about other schools through their own "shopping" and through interactions with other parents at work and in their communities. Most students, in contrast, only interact with students within their schools and, even when they do talk with, are not likely to compare notes about their textbooks and teachers.[23]

The patterns we observe with students are of particular interest given that they are the ones being immediately served and the rise of other platforms, such as Niche, that heavily recruit students to share their views. If the topics discussed were driven mainly about familiarity, students would write a lot more than parents, especially about instruction. But this is not what we observe. We saw no clear difference between students and parents on Instruction and Learning. Instead, we suspect that students' increased familiarity is offset by reduced emphasis on discussing Instruction and Learning in their reviews, relative to parents. This aligns with the fact that parents write less than students about extracurricular activities, which may be especially important to students (Harris & Larsen, 2023).

The largest differences are not by user types, however, but by school type. Some of these are obvious (e.g., users from private schools write more often about religion), but other patterns are less obvious. Charter and private school reviews are longer, which might reflect that these families–more often the active choosers–have put more thought into schooling than those attending their residentially assigned traditional public schools, so that they have more to say.

More generally, we saw almost no commonality at all in the relationship between topics and ratings across sectors. Charter reviews focus more on school discipline practices. Charter and private school reviews also focus more on preparing students as adults.[24] This could reflect differences in both preferences and school characteristics and effectiveness on the various schooling dimensions. Still, a broader theme, consistent with prior research, is that users seem interested in a wide variety of school attributes and programs and that their tastes vary (Harris & Larsen, 2023; Beuermann et al., 2023).

[23] An exception is that users who are providers (teachers and school leaders) do not write about themselves, which they know a great deal about, perhaps because it seems self-serving.
[24] The latter conclusion is in Appendix C.

A second purpose of our work is to understand whether reviews provide information for readers to make school decisions. While parents may find the information helpful in making their schooling decisions, our results suggest that the information may actually be more limited First, prior research suggests that reviewers are not representative of all school stakeholders. The GreatSchools reviews are written overwhelmingly by parents and those who have a strong positive or negative opinion about their experiences in the school. Only a tiny fraction of people who have had experience with the school or who visit the GreatSchools site post a review, so, at the very least, there is considerable measurement error. On other review platforms, there is evidence that reviews are not truthful, as some positive reviewers have a personal stake in the outcome and some negative reviews are written by malicious actors. It's unclear whether these problems arise with school review platforms. Schools also may engage in "ratings drives" to encourage their community to add reviews to the site, artificially inflating the number of reviews they receive.

Even if reviews were representative and reliable, a second limitation is the vagueness and unstructured nature of the information included in reviews makes them challenging to interpret. The overwhelming majority of reviews include vague adjectives and do not describe the qualities, activities or orientations of schools in ways specific enough to make them useful to others. Even when they do provide more in-depth information, the unstructured nature of the data makes it difficult to compare similar attributes within or between schools. Education has "exceptionally ambiguous output" (Hess, 2004) and is responsible for a wide range of outcomes (Brighouse et al., 2018), which makes it more difficult to describe than other goods and services. No single reviewer could cover all topics that people could value in a given school and users are rarely clear about their own tastes and values. A parent looking for a school that aligns with their values would have a hard time determining that from the available reviews for a school.

Increasing the number of reviews is not necessarily going to solve these issues. It is unlikely that the share of reviewers for a given school will ever be high enough that self-selection is not a problem. A larger volume of reviews would increase the number of reviews that contain greater depth, but they do not solve the complexity, taste variation, and unstructured data problems. Moreover, when the number of reviews becomes large, the time, effort, and difficulty of synthesizing the information grows considerably.

This challenge of interpretating the results leads us to suggest three possible directions for user platforms to become more useful. First, they could build in more structured data collection (e.g., surveys with closed ended questions) that would provide more standardized and comparable information across the schools from which families may be choosing. GreatSchools has implemented some more structured data collection since the start of our analysis. One problem with that approach is that it tends to reduce user engagement, making the first problem of "response rate" even worse.

Second, as the number of text reviews increases, the user platforms could use NLP to digest and summarize the information not only in the aggregate, as we have done here, but also for particular schools.[25] Our analysis suggests that such an approach would potentially overestimate categories discussed in reviews, but also reduce the cognitive complexity of the unstructured data already being collected, owing to the fact that different users write about different topics. AI-based summaries could help reduce that complexity.

A problem with the above approach is that AI algorithms might conflate the value systems of different users and the different decisions they have to make. Parents and students have to decide where to attend and principals and teachers have to decide where to work. To address this, another option would be to separate reviews and review summaries by user types. Having different kinds of users writing in these platforms is helpful because different users are having to make decisions about schools. This would allow users to seek out information that is most relevant to their decisions and educational values, while also allowing users to get a more well-rounded picture of each school if they wish to look at the school from various perspectives. For example, when parents seek information, the platforms could allow them to ask the site to summarize reviews about teaching and learning written by other parents, but they could also separately examine the reviews of students and teachers.

With the expansion of school choice across the country, both in where choice is and in the kinds of choices available, it is important to consider the implications of the information on sites like GreatSchools on school choice more generally. We discussed the three main reasons that make reviews vary by user--differing tastes, emphases, and expectations–and found additional differences by type of school. Some families choose to exit TPS for charter and private schools, and we might expect that those users who exit have different tastes, emphases, and/or expectations–especially with families that choose private schools, given their substantial tuition payments. This adds an additional layer of complexity for families considering schools that are in different sectors. Many users talk about school selection processes, which reinforces prior evidence that the choice is broader than families choosing schools, but also includes schools choosing families. While our results are focused on comparing TPS, charter, and private schools, the recent and rapid expansion of education savings accounts (ESAs) has been accompanied by user platforms aimed at providing information to families for narrower services such as tutoring and extracurriculars. Our analysis suggests that user review platforms are more promising in these contexts because the services are much more specific and easier to evaluate objectively. For example, the "culture" of a tutoring organization is unlikely to be relevant. On

[25] GreatSchools has started using keywords from reviews to organize them and collecting more structured data since this analysis began. Another user platform that includes schools, Niche, asks reviewers to grade schools on six dimensions: Academics, Teachers, College Prep, Diversity, Clubs and Activities, and Administration.

the other hand, the other kinds of information on sites such as GreatSchools (e.g., test scores) will not be available because schools serving students with ESAs are not required to report any standardized outcome information. Parents will be quite dependent on the text and star ratings in those platforms.

Our analysis certainly cannot be viewed as an evaluation of school choice. Still, as information is a key element of the choice process, it does provide an important insight into how the schooling market is limited by the information available and those providing the information. The study of user review platforms in education is relatively new. For this reason, we designed our study to begin understanding the content of reviews and patterns in the data, as a precursor to a broader analysis of the topic and potential policy directions. These platforms clearly appear to be part of the future of consumer information and they are likely to grow in influence in schooling in this new era of expanded school choice when information is ever important. Future research should consider how consumers use these sites to make decisions, and how that may differ by the type of consumer.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Altenhofen, S., Berends, M., & White, T. G. (2016). School Choice Decision Making Among Suburban, High-Income Parents. *AERA Open*, *2*(1). https://journals.sagepub.com/doi/full/10.1177/2332858415624098

Antelmi, A., Malandrino, D., & Scarano, V. (2019). Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule. *Companion Proceedings of The 2019 World Wide Web Conference*, 1035–1038. https://doi.org/10.1145/3308560.3316705

Aral, S., & Walker, D. (2014). Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment. *Management Science*, *60*(6), 1352–1370. https://doi.org/10.1287/mnsc.2014.1936

Auerbach, C., & Silverstein, L. B. (2003). *Qualitative Data: An Introduction to Coding and Analysis*. NYU Press.

Bell, C. A. (2009). All Choices Created Equal? The Role of Choice Sets in the Selection of Schools. *Peabody Journal of Education*, *84*(2), 191–208. https://doi.org/10.1080/01619560902810146

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Brauwers, G., & Frasincar, F. (2022). A Survey on Aspect-Based Sentiment Classification. *ACM Computing Surveys*, *55*(4), 65:1-65:37. https://doi.org/10.1145/3503044

Burgess, S., Greaves, E., Vignoles, A., & Wilson, D. (2015). What Parents Want: School Preferences and School Choice. *The Economic Journal*, *125*(587), 1262–1289. https://doi.org/10.1111/ecoj.12153

Carroll, J. M., Harris, D. N., Gerry, A., & Weixler, L. (2023). *Voices of New Orleans Youth 2022: How are our city's children doing after three unusual years?* Education Research Alliance for New Orleans. https://educationresearchalliancenola.org/publications/voices-of-new-orleans-youth-2022-how-are-our-children-doing-after-three-unusual-years

Corcoran, S. P., & Jennings, J. L. (2019). Information and School Choice. In *Handbook of Research on School Choice* (2nd ed.). Routledge.

Cordes, S. A., Lenhoff, S. W., Schwartz, A. E., Singer, J., & Samantha, T. (2023). *Choice in a Time of COVID: Immediate Enrollment Decisions in New York City and Detroit*. National Center for Research on Education Access and Choice. https://reachcentered.org/publications/choice-in-a-time-of-covid

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th International Conference on World Wide Web*, 519–528. https://doi.org/10.1145/775152.775226

Domina, T., Carlson, D., Carter III, J., Lenard, M., McEachin, A., & Perera, R. (2021). The Kids on the Bus: The Academic Consequences of Diversity-Driven School Reassignments. *Journal of Policy Analysis and Management*, *40*(4), 1197–1229. https://doi.org/10.1002/pam.22326

Glazerman, S., & Dotter, D. (2017). Market Signals: Evidence on the Determinants and Consequences of School Choice From a Citywide Lottery. *Educational Evaluation and Policy Analysis*, *39*(4), 593–619.

Goldring, E. B., & Phillips, K. J. R. (2008). Parent preferences and parent choices: The public–private decision about school choice. *Journal of Education Policy*, *23*(3), 209–230. https://doi.org/10.1080/02680930801987844

Gregory, K. M. (2011). How Undergraduates Perceive Their Professors: A Corpus Analysis of Rate My Professor. *Journal of Educational Technology Systems*, *40*(2), 169–193.

Harris, D. N., & Larsen, M. F. (2023). What Schools Do Families Want (and Why)? Evidence on Revealed Preferences From New Orleans. *Educational Evaluation and Policy Analysis*, *45*(3), 496–519. https://doi.org/10.3102/01623737221134528

Harris, D. N., Witte, J. F., & Valant, J. (2017). The Market for Schooling. In *Shaping Education Policy* (2nd ed.). Routledge.

Hess, F. M. (2004). *Revolution at the Margins: The Impact of Competition on Urban School Systems*. Rowman & Littlefield.

Hong, Y. A., Liang, C., Radcliff, T. A., Wigfall, L. T., & Street, R. L. (2019). What Do Patients Say About Doctors Online? A Systematic Review of Studies on Patient Online Reviews. *Journal of Medical Internet Research*, *21*(4), e12521. https://doi.org/10.2196/12521

Hovy, D., Johannsen, A., & Søgaard, A. (2015). User Review Sites as a Resource for Large-Scale Sociolinguistic Studies. *Proceedings of the 24th International Conference on World Wide Web*, 452–461. https://doi.org/10.1145/2736277.2741141

Hoxby, C. M. (2000). Does Competition among Public Schools Benefit Students and Taxpayers? *American Economic Review*, *90*(5), 1209–1238. https://doi.org/10.1257/aer.90.5.1209

Kleitz, B., Weiher, G. R., Tedin, K., & Matland, R. (2000). Choice, Charter Schools, and Household Preferences. *Social Science Quarterly (University of Texas Press)*, *81*(3), 846–854.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 939–948. https://doi.org/10.1145/1871437.1871557

Liu, B. (2012). Sentiment Analysis: A Fascinating Problem. In B. Liu (Ed.), *Sentiment Analysis and Opinion Mining* (pp. 1–8). Springer International Publishing. https://doi.org/10.1007/978-3-031-02145-9_1

Liu, Y., Ren, C., Shi, D., Li, K., & Zhang, X. (2020). Evaluating the social value of online health information for third-party patients: Is uncertainty always bad? *Information Processing & Management*, *57*(5), 102259. https://doi.org/10.1016/j.ipm.2020.102259

Lovenheim, M. F., & Walsh, P. (2018). (Re)searching for a school: how choice drives parents to become more informed. *Education Next*, *18*(1), 72–78.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. https://www.amazon.com/Introduction-Information-Retrieval-Christopher-Manning/dp/0521865719

Marlin, B. M., & Zemel, R. S. (2009). Collaborative prediction and ranking with non-random missing data. *Proceedings of the Third ACM Conference on Recommender Systems*, 5–12. https://doi.org/10.1145/1639714.1639717

McShane, M. Q. (2024). A New Crop of School Models Expands Choice. *Education Next*, *24*(2), 1.

Nielsen, J. (2006). *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities*. Nielsen Norman Group. https://www.nngroup.com/articles/participation-inequality/

Owens, A., & Candipan, J. (2019). Social and spatial inequalities of educational opportunity: A portrait of schools serving high- and low-income neighbourhoods in US metropolitan areas. *Urban Studies*, *56*(15), 3178–3197. https://doi.org/10.1177/0042098018815049

Rajaraman, V. (2011). *Analysis and Design of Information Systems*. PHI Learning Pvt. Ltd.

Rayson, P., & Garside, R. (2000). Comparing Corpora using Frequency Profiling. *The Workshop on Comparing Corpora*.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). Recommendations as Treatments: Debiasing Learning and Evaluation. *Proceedings of The 33rd International Conference on Machine Learning*, 1670–1679. https://proceedings.mlr.press/v48/schnabel16.html

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

Tarnowski, E. (2024, March 29). *School Choice in the States: March 2024*. EdChoice. https://www.edchoice.org/engage/school-choice-in-the-states-march-2024/

Teske, P., Fitzpatrick, J., & Kaplan, G. (2007). Opening Doors: How Low-Income Parents Search for the Right School. In *Online Submission*. https://eric.ed.gov/?id=ED495279

Valant, J., & Newark, D. A. (2020). The Word on the Street or the Number from the State? Government-Provided Information and Americans' Opinions of Schools. *Journal of Public Administration Research and Theory*, *30*(4), 674–692. https://doi.org/10.1093/jopart/muaa010

West, D. M. H., Paul E. Peterson, Martin R. (2022, August 16). Partisan Rifts Widen, Perceptions of School Quality Decline. *Education Next*. https://www.educationnext.org/partisan-rifts-widen-perceptions-school-quality-decline-results-2022-education-next-survey-public-opinion/

Table 1. Examples of Great Schools Text Reviews

One-Star Reviews	Five-Star Reviews
My daughter just started here and she hates it she says kids pick on her friends and she's afraid they're going to keep picking on her so I'm seriously thinking of putting her in another school.	I love this school I went here all my life and loved it the best 12 years of my life.
Lots of yelling. One administrator uses a bullhorn to yell at kids daily and is very rude to both students and parents. Very unpleasant atmosphere. My kids hate going to school there. We are pulling both of our kids and going to private. I would not recommend this school to anyone. We came from another state where my kids actually enjoyed going to school. Can't wait to see them actually excited about going to school again.	Excellent school great learning environment for students. Excellent teachers who care for children. Safe and clean facility. The curriculum is challenging the extracurricular offerings are enormous and the school has incredible computer science labs as well as a new state of the art playground.
Heed all these warnings. This is the worst school. My son had been major bullied in 7th grade and 8th grade. The old principal in 7th grade had taken care of the student that was bullying him in 7th grade. Since the new principal started her reign in his 8th grade year, the bully situation didn't stop. Instead it intensified on the bus and at school. I tried to get him to handle the situation himself but the teachers and principal just did not care. It seems that they don't want to be bothered and I as a parent take this seriously. My son never bothered anyone. He stayed to himself with the exception of a few friends that he had hung out with. He was in all honors classes but he also was not into sports and they are all about the sports. They just don't care or take time out of their schedule unless your child is a child of the school board then you're in the clique. If you care about your kid's education that is not affected by bullies then stay. We had to move out of state to get him out of that bad situation.	We moved to go to [school name]. I have four kids and two are still there. I have been so impressed with the teachers and the school has a science lab that compares with my college science lab that the students have used numerous times. My second grader grew vegetables in the school garden and made vegetable soup for the staff. The school offers clubs and extra curricular activities such as dance, track, academic teams, etc. My husband and I would recommend this school to anyone wanting their child to be challenged and be surrounded by a multicultural atmosphere.

Notes to Table 1: Example reviews showing the difference between one star reviews and reviews rated five stars (out of five).

Table 2. Great Schools' Reviews, 2002-2019

Sub-domain	Number of Reviews (%)	Number of Schools with Reviews (%)	Mean Number of Reviews per School ± Std. Deviation	Median Number of Reviews per School	Max Number of Reviews per School	Number of Ratings (% of rated reviews)	Mean Individual Rating ± Std. Deviation
Full Corpus	677,116 (100%)	83,789 (100%)	8.08 ± 14.76	5	1,685	578,578 (100%)	3.88 ± 1.57
Parent	504,241 (74.47%)	76,963 (91.85%)	6.55 ± 10.93	4	860	427,262 (73.85%)	3.91 ± 1.56
Principal	1,425 (0.21%)	1,336 (1.59%)	1.07 ± 0.28	1	4	79 (0.01%)	4.94 ± 0.25
Student	53,894 (7.96%)	22,451 (26.79%)	2.40 ± 6.24	1	541	45,681 (7.90%)	3.83 ± 1.52
Teacher	21,604 (3.19%)	13,313 (15.89%)	1.62 ± 1.64	1	48	18,457 (3.19%)	4.25 ± 1.38
Trad. Public	450,800 (66.58%)	62,927 (75.10%)	7.16 ± 13.99	4	1,685	379,365 (65.57%)	3.79 ± 1.58
Charter	64,046 (9.46%)	3,986 (4.76%)	16.07 ± 25.75	8	658	54,971 (9.50%)	3.84 ± 1.60
Private	162,270 (23.96%)	17,090 (20.40%)	9.50 ± 12.93	6	545	144,242 (24.93%)	4.12 ± 1.49
Elementary/ Middle	566,485 (83.66%)	69,171 (82.55%)	8.19 ± 12.48	5	658	486,457 (84.07%)	3.91 ± 1.56
High	110,478 (16.32%)	14,795 (17.66%)	7.47 ± 22.45	4	1,685	91,986 (15.89%)	3.70 ± 1.60
Urban	501,937 (74.13%)	63,924 (76.29%)	7.85 ± 15.29	5	1,685	425,903 (73.61%)	3.86 ± 1.57

Not Urban	89,737 (13.25%)	20,843 (24.87%)	4.30 ± 7.13	3	333	74,407 (12.86%)	3.79 ± 1.60
Good Review	404,417 (59.73%)	76,074 (90.79%)	5.32 ± 7.39	3	562	404,417 (69.90%)	4.85 ± 0.36
Bad Review	174,161 (25.72%)	55,132 (65.80%)	3.16 ± 3.26	2	78	174,161 (30.10%)	1.63 ± 0.79

Notes to Table 2: Showing the distribution of total number of reviews over various domains. Different school types show the distribution over reviews given to public, private, and charter schools. Different review qualities show the difference in frequency between good reviews (a four or five star rating) and bad reviews (a three or less star rating).

Figure 1: Information on the Great Schools site for a specific school

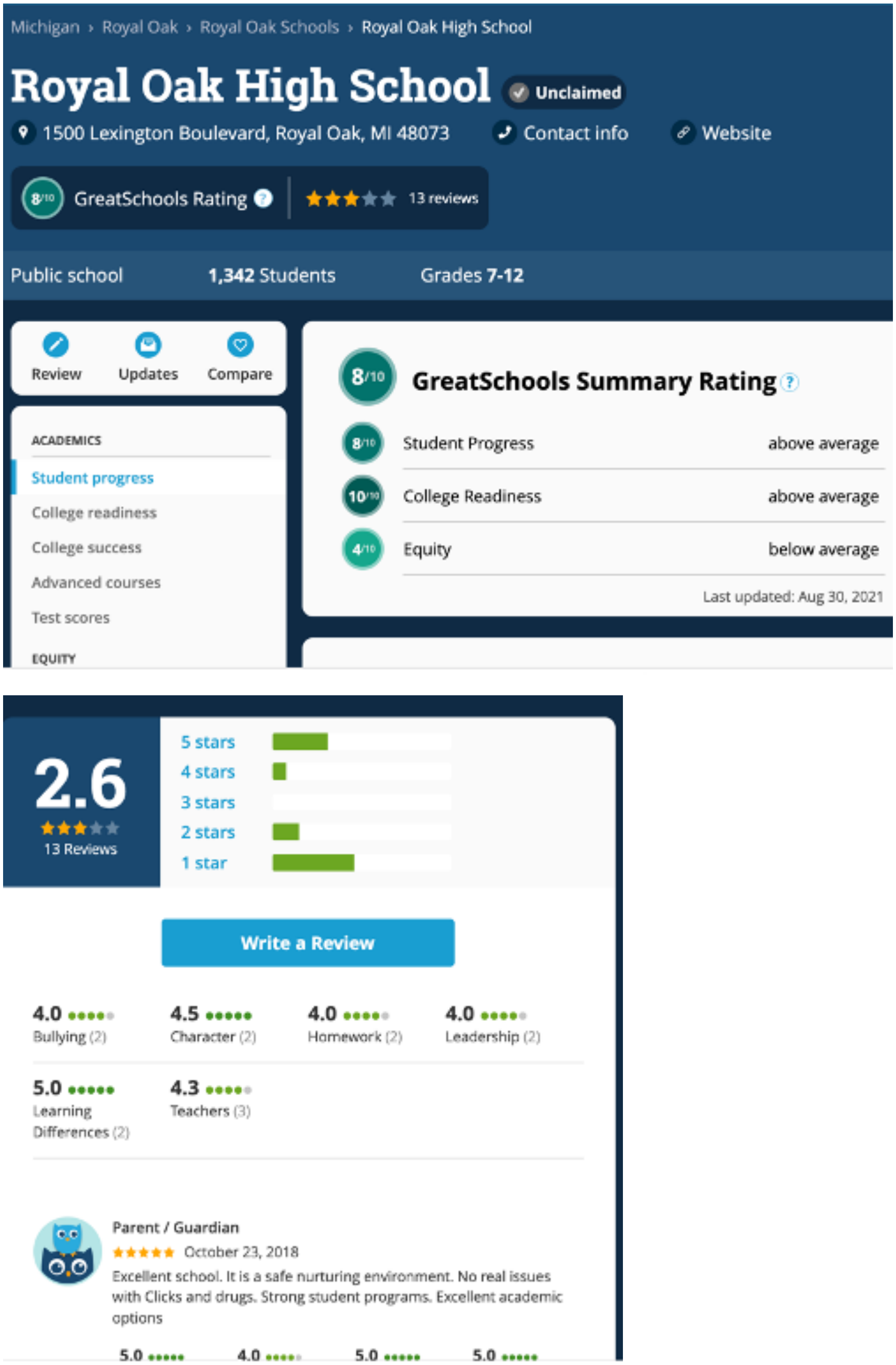

Figure 2: Partial Kernel Density Plot of the Number of Words in Each Review

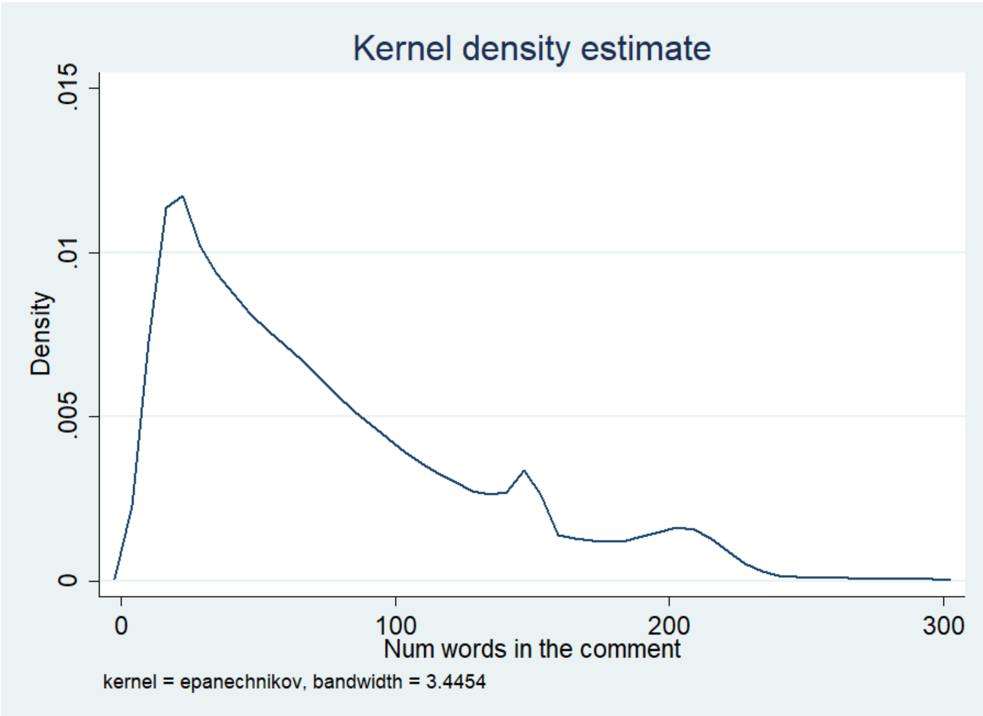

Table 3. Review Keywords by Topic and Sub-topic, 2002-2019

Topic	Number of Reviews (%)	Sub-topic	Number of Reviews (%)	Example Keywords
Instruction and Learning	430,909 (63.64%)	Instruction	131,359 (19.40%)	Constructivist, Hybrid, Individualized, Remedial
		Curriculum/Curricular Materials	167,331 (24.71%)	Literature, Geography, Homework, Assignment
		Learning Experience	353,267 (52.17%)	Understand, Taught, Pass, Explore
Overall Quality	589,593 (87.07%)	Evaluation	173,345 (25.60%)	Accountability, Rate, Star, Tested
		Postsecondary/Graduation	68,995 (10.19%)	Graduate, University, College, Future
		Preparation	23,159 (3.42%)	Prep, Preparation
		Quality Indicator	563,275 (83.19%)	Struggling, Awesome, Horrible, Excellence
Physical Environment	244,279 (36.08%)	Location	117,840 (17.40%)	District, Area, Zone, Town
		Building Quality	13,257 (1.96%)	Clean, Disgusting, Outdated, Immaculate
		Facilities	161,303 (23.82%)	Classroom, Cafeteria, Bathroom, Trash
Resources	223,625 (33.03%)	Extracurriculars/Electives	89,543 (13.22%)	Club, Football, Piano, Chess
		Offerings	180,179 (26.61%)	Magnet, Gifted, Preschool, Disabled
School Culture	485,628 (71.72%)	Student Discipline	44,263 (6.54%)	Suspended, Discipline, Detention, Punitive
		School Safety	52,115 (7.70%)	Threaten, Fight, Bully, Retaliation
		Interpersonal Relationships	178,019 (26.29%)	Communication, Approachable, Fellow, Clique
		School Environment	387,240	Warm, Welcoming, Unfriendly, Racist

			(57.19%)	
School-Level Features	289,396 (42.74%)	Religious	42,274 (6.24%)	Spiritual, Trinity, Holy, Parish
		School Finances	57,411 (8.48%)	Payment, Tax, Expense, Afford
		School Type	191,939 (28.35%)	Elementary, Catholic, Private, Secondary
School Staff	514,971 (76.05%)	Teacher Quality	37,941 (5.6%)	Inexperienced, Qualified, Trained, Unprepared
		Other School Staff	510,451 (75.39%)	Fired, Teacher, Principal, Turnover
Unclassified	592,592 (87.52%)	Students	561,907 (82.99%)	Child, Scholar, Enrollment, Alumni
		Family	230,935 (31.44%)	Parent, Husband, Sister, Family
		School Choice/Enrollment	198,323 (29.29%)	Lottery, Transferring, Touring, Enroll

Notes to Table 3: A review is considered to mention that topic if a keyword labeled with that topic appears.

Table 4: Regression Results Comparing the Likelihood of Review Types

	(1) # of Words in the Review	Average Topics per Review (Std. Dev.)	Topics (2) Physical Environment	(3) Resources	(4) School Staff	(5) Instruction and Learning	(6) Overall Quality	(7) School Culture	(8) School-Level Features	(9) Unclassified
Reviewer Type:										
Principal	5.580*	5.203	0.039**	0.045***	-0.125***	0.060***	-0.026**	0.055***	0.054***	-0.032***
	(2.449)	(1.840)	(0.013)	(0.013)	(0.013)	(0.012)	(0.010)	(0.011)	(0.013)	(0.009)
Student	-0.872*	4.593	-0.070***	0.002	-0.062***	-0.019***	0.014***	-0.076***	-0.032***	-0.199***
	(0.393)	(1.971)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Teacher	-7.286***	5.042	-0.011**	-0.035***	0.002	0.031***	-0.013***	0.069***	0.004	-0.039***
	(0.466)	(1.747)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
School Type:										
Charter	13.152***	5.052	-0.004	-0.008	-0.057***	0.122***	0.005	0.025***	0.066***	0.023***
	(1.145)	(1.798)	(0.005)	(0.005)	(0.006)	(0.004)	(0.003)	(0.004)	(0.005)	(0.003)
Private	2.053	5.342	-0.088***	-0.041***	-0.121***	0.048***	-0.027***	-0.007	0.125***	0.007
	(1.817)	(1.730)	(0.010)	(0.011)	(0.010)	(0.011)	(0.008)	(0.010)	(0.011)	(0.007)
School Level:										
High School	4.099***	4.897	0.035***	0.064***	-0.086***	0.032***	0.010***	-0.018***	0.090***	-0.003
	(0.887)	(1.882)	(0.005)	(0.003)	(0.005)	(0.004)	(0.002)	(0.005)	(0.005)	(0.004)
Demographics:										
Herfindahl Diversity Index	12.867***	-	0.004	0.055***	0.044***	0.015**	0.019***	0.052***	0.004	0.027***
	(1.148)		(0.006)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)	(0.006)	(0.005)
Prop. FRPL Eligible	0.726	-	-0.030***	-0.059***	-0.016***	-0.061***	-0.041***	-0.026***	-0.088***	0.012***
	(0.911)		(0.005)	(0.004)	(0.005)	(0.005)	(0.003)	(0.004)	(0.005)	(0.003)
Not Urban	-3.197***	4.859	0.018***	-0.023***	-0.001	-0.023***	-0.005**	-0.006*	-0.003	-0.003
	(0.591)	(1.780)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)

Constant	67.141***	-	0.380***	0.331***	0.806***	0.609***	0.878***	0.709***	0.401***	0.885***
	(0.719)		(0.004)	(0.003)	(0.003)	(0.004)	(0.002)	(0.003)	(0.004)	(0.003)
% Reviews with the Topic	-		36.08%	33.03%	76.05%	63.64%	87.07%	71.72%	42.74%	87.52%
Observations	677,116		677,116	677,116	677,116	677,116	677,116	677,116	677116	677116
R^2	0.023		0.005	0.006	0.022	0.033	0.003	0.013	0.042	0.049
Adjusted R^2	0.023		0.005	0.006	0.022	0.033	0.003	0.013	0.042	0.049

Notes to Table 4: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and unknown categories. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5A: Regression Results Comparing the Likelihood of Review Types: Physical Environment and School Staff

	Physical Environment			School Staff	
	(1)	(2)	(3)	(4)	(5)
	Location	Building Quality	Facilities	School Staff	Teacher Quality
Reviewer Type:					
Principal	0.040***	0.006	0.028*	-0.135***	0.024**
	(0.011)	(0.004)	(0.012)	(0.013)	(0.008)
Student	-0.071***	-0.000	-0.022***	-0.063***	-0.013***
	(0.002)	(0.001)	(0.002)	(0.003)	(0.001)
Teacher	0.004	0.001	-0.016***	0.003	0.022***
	(0.003)	(0.001)	(0.003)	(0.003)	(0.002)
School Type:					
Charter	-0.038***	-0.006***	0.033***	-0.059***	0.014***
	(0.003)	(0.001)	(0.004)	(0.006)	(0.001)
Private	-0.082***	-0.004	-0.035***	-0.132***	0.033***
	(0.007)	(0.003)	(0.009)	(0.011)	(0.006)
School Level:					
High	0.060***	-0.001*	-0.012***	-0.096***	0.023***
	(0.003)	(0.001)	(0.003)	(0.005)	(0.001)
Demographics:					
Herfindahl Diversity Index	-0.045***	0.004***	0.049***	0.046***	-0.004*
	(0.004)	(0.001)	(0.005)	(0.006)	(0.002)
Prop free lunch eligible	-0.044***	0.001	-0.005	-0.012**	-0.011***
	(0.003)	(0.001)	(0.004)	(0.005)	(0.001)
Not Urban	0.034***	0.003***	-0.010***	-0.000	-0.006***
	(0.002)	(0.001)	(0.003)	(0.003)	(0.001)
Constant	0.223***	0.017***	0.220***	0.802***	0.049***
	(0.003)	(0.001)	(0.003)	(0.003)	(0.001)
% Reviews with the Subtopic	17.40%	1.96%	23.82%	75.39%	5.60%
Observations	677,116	677,116	677,116	677,116	677,116
R^2	0.011	0.000	0.004	0.025	0.007
Adjusted R^2	0.011	0.000	0.004	0.024	0.007

Notes to Table 5A: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and unknown categories. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 5B: Regression Results Comparing the Likelihood of Review Types: Instruction and Learning and Overall Quality

	Instruction and Learning			Overall Quality			
	(1) Instruction	(2) Curriculum/Curricular Materials	(3) Learning Experience	(4) Evaluation	(5) Postsecondary/Graduation	(6) Preparation	(7) Quality Indicator
Reviewer Type:							
Principal	0.078***	0.050***	0.083***	0.027*	0.068***	0.045***	-0.054***
	(0.012)	(0.013)	(0.013)	(0.012)	(0.010)	(0.007)	(0.011)
Student	-0.013***	-0.003	-0.023***	-0.021***	0.017***	0.004***	0.019***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)
Teacher	0.086***	-0.025***	0.021***	-0.022***	0.011***	0.008***	-0.011***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)	(0.003)
School Type:							
Charter	0.074***	0.095***	0.116***	0.034***	0.044***	0.030***	-0.005~
	(0.003)	(0.005)	(0.004)	(0.005)	(0.002)	(0.002)	(0.003)
Private	0.004	0.000	0.089***	-0.038***	0.058***	0.021***	-0.032***
	(0.008)	(0.010)	(0.011)	(0.009)	(0.007)	(0.004)	(0.009)
School Level:							
High	-0.021***	0.044***	0.027***	-0.065***	0.195***	0.037***	-0.002
	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)	(0.001)	(0.002)
Demographics:							
Herfindahl Diversity Index	0.022***	0.054***	-0.005	0.047***	-0.018***	-0.001	0.022***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.003)	(0.002)	(0.004)
Prop free lunch eligible	-0.052***	-0.062***	-0.046***	-0.005	0.010***	0.001	-0.055***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.002)	(0.001)	(0.003)
Not Urban	-0.010***	-0.037***	-0.014***	-0.015***	-0.003*	-0.003***	-0.006**
	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)
Constant	0.190***	0.222***	0.490***	0.252***	0.041***	0.018***	0.849***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)

% Reviews with the Subtopic	19.40%	24.71%	52.17%	25.60%	10.19%	3.42%	83.19%
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.013	0.016	0.036	0.007	0.071	0.010	0.004
Adjusted R^2	0.013	0.016	0.036	0.007	0.071	0.010	0.004

Notes to Table 5B: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary, and Urban, and models include flags for each imputed data point and unknown categories. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5C: Regression Results Comparing the Likelihood of Review Types: School Culture and School-level Features

	School Culture				School-level Features		
	(1) Student Discipline	(2) School Safety	(3) Interpersonal Relationships	(4) School Environment	(5) Religious	(6) School Finances	(7) School Type
Reviewer Type:							
Principal	-0.015*	0.001	0.041***	0.077***	0.065***	-0.050***	0.086***
	(0.006)	(0.007)	(0.012)	(0.013)	(0.009)	(0.006)	(0.013)
Student	-0.010***	0.013***	-0.018***	-0.087***	-0.013***	-0.006***	-0.001
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.001)	(0.003)
Teacher	-0.002	-0.013***	0.046***	0.088***	0.014***	-0.013***	0.017***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)	(0.004)
School Type:							
Charter	0.018***	-0.009***	0.031***	0.002	0.007***	0.021***	0.035***
	(0.002)	(0.002)	(0.005)	(0.004)	(0.001)	(0.002)	(0.004)
Private	0.002	0.006	-0.033***	0.014	0.200***	0.052***	0.006
	(0.005)	(0.006)	(0.010)	(0.011)	(0.010)	(0.007)	(0.010)
School Level:							
High	0.000	-0.006***	-0.022***	-0.025***	-0.010***	0.020***	0.096***
	(0.001)	(0.001)	(0.003)	(0.004)	(0.001)	(0.002)	(0.003)
Demographics:							
Herfindahl Diversity Index	0.011***	0.020***	0.057***	0.036***	-0.015***	0.010***	0.009$^{\sim}$
	(0.002)	(0.002)	(0.005)	(0.005)	(0.002)	(0.003)	(0.005)
Prop free lunch eligible	0.038***	0.045***	-0.011**	-0.061***	-0.001	-0.009***	-0.086***
	(0.002)	(0.002)	(0.004)	(0.004)	(0.001)	(0.002)	(0.004)
Not Urban	-0.000	-0.004**	-0.008**	-0.013***	0.005***	0.000	-0.003
	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)
Constant	0.046***	0.053***	0.252***	0.587***	0.022***	0.064***	0.286***
	(0.001)	(0.002)	(0.003)	(0.003)	(0.001)	(0.002)	(0.003)
% Reviews with the Subtopic	6.54%	7.70%	26.29%	57.19%	6.24%	8.48%	28.35%
Observations	677116	677116	677116	677116	677116	677116	677116
R^2	0.002	0.003	0.004	0.016	0.136	0.013	0.013
Adjusted R^2	0.002	0.003	0.004	0.016	0.136	0.013	0.013

Notes to Table 5C: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary, and Urban, and models include flags for each imputed data point and unknown categories. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5D: Regression Results Comparing the Likelihood of Review Types: School Resources and Unclassified

	Resources		Unclassified			
	(1) Extracurriculars/ Electives	(2) Offerings	(1) Students	(2) Family	(3) School Choice/ Enrollment	(4) Other
Reviewer Type:						
Principal	0.051***	0.038**	-0.060***	0.016	-0.097***	0.095***
	(0.010)	(0.013)	(0.010)	(0.013)	(0.011)	(0.013)
Student	0.049***	-0.035***	-0.240***	-0.217***	-0.024***	-0.077***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
Teacher	0.018***	-0.049***	-0.050***	-0.007~	-0.124***	0.025***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
School Type:						
Charter	-0.003	-0.007~	0.028***	0.036***	0.048***	0.100***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Private	-0.007	-0.037***	0.005	0.040***	0.042***	0.043***
	(0.007)	(0.010)	(0.008)	(0.011)	(0.010)	(0.010)
School Level:						
High	0.115***	-0.004	0.011*	-0.084***	0.029***	0.028***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)
Demographics:						
Herfindahl Diversity Index	0.012***	0.058***	0.027***	0.046***	0.025***	0.030***
	(0.003)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
Prop free lunch eligible	-0.056***	-0.036***	0.036***	-0.070***	0.069***	-0.054***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Not Urban	0.001	-0.029***	0.001	-0.027***	-0.001	-0.010***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	0.120***	0.276***	0.833***	0.390***	0.234***	0.277***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
% Reviews with the Subtopic	13.22%	26.61%	82.99%	34.11%	29.29%	29.43%
Observations	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.024	0.005	0.054	0.047	0.011	0.022
Adjusted R^2	0.024	0.005	0.054	0.047	0.011	0.022

Notes to Table 5D: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary, and Urban, and models include flags for each imputed data point and unknown categories. Standard errors are clustered at the school level and are shown in parentheses. ~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6: Reviews from parents and students from the same school to highlight differences in the aspects of schools parents and students emphasize

	Parent	Student
Reviews from the same Traditional Public School	**Does not discuss classroom experiences** [School Name] is a great school. My son graduated in 2010 and is an UCSD sophomore. Most of his friends got in UC. Actually, no matter which school you attend, if you study hard dream will come true.	**Discusses classroom experiences** My school is awesome because when you don't understand or misunderstand something you could always ask teachers to assistance you to became understanding the problems, and you will know that once you get it would be a piece of a cake of any subjects
Reviews from the same Private School	**Focus on preparing for life after high school** [School Name] has lost its purpose! They are so concerned about feminism that there is no focus on love of self, love of others or loving God....it's all about winning. Winning sports titles, getting scholarships and going to elite colleges. College preparatory should also be about LIFE Preparatory! These young women are only learning about how to get into college -- I want my daughter to learn about how to live in college and beyond! The stress [School Name] puts on the girls is unreal! Busywork, rules, ridiculous and petty punishments....the joy is gone form the campus.	**Does not discuss life after high school** The teachers are some of the smartest individuals you will ever meet, classes are small, [School Name] is right next door, tons of dances, mass once a month. Very tolerant of different religions. I'm Mormon going to a Catholic school. [School Name] is awesome!
Reviews from the same Charter School	**Compares to other schools in the area** Highest score possible!!! Saved my kids' lives from the suburban, cookie-cutter-kid public schools that thought they couldn't accomplish anything and gave	**Does not compare to other schools in the area** I am a graduating senior this year, and it has been an absolutely wonderful experience. I wouldn't trade it for anything. The classes are challenging, but the teachers

	them tremendous confidence and opportunity. I’m forever grateful for [School Name] !!!	are always willing to work with you if you have any problems. The faculty truly cares about students and is willing to do whatever they can so they succeed.

Notes to Table 6: Reviews have been edited to exclude the school name.

Table 7: Regression Results Predicting Star Ratings, by User Type and School Sector

		Respondent Type			School Sector		
		(1)	(2)	(3)	(4)	(5)	(6)
	All	Parent	Student	Teacher	TPS	Charter	Private
Physical Environment	-0.362***	-0.379***	-0.281***	-0.285***	-0.412***	-0.327***	-0.174***
	(0.005)	(0.005)	(0.017)	(0.023)	(0.006)	(0.017)	(0.009)
Resources	0.183***	0.192***	0.165***	0.212***	0.224***	0.118***	0.128***
	(0.005)	(0.005)	(0.016)	(0.022)	(0.006)	(0.017)	(0.009)
School Staff	-0.026***	-0.029***	-0.048***	-0.313***	0.107***	-0.204***	-0.164***
	(0.006)	(0.007)	(0.018)	(0.025)	(0.008)	(0.018)	(0.010)
Instruction and Learning	0.100***	0.105***	-0.020	0.110***	-0.032***	0.204***	0.351***
	(0.005)	(0.006)	(0.016)	(0.025)	(0.006)	(0.018)	(0.012)
Overall Quality	0.495***	0.505***	0.449***	0.275***	0.522***	0.359***	0.417***
	(0.008)	(0.009)	(0.031)	(0.034)	(0.009)	(0.026)	(0.016)
School Culture	-0.081***	-0.072***	-0.210***	0.233***	-0.212***	0.052***	0.203***
	(0.005)	(0.006)	(0.016)	(0.030)	(0.006)	(0.018)	(0.011)
School-level Features	0.002	0.015***	-0.039**	-0.015	0.032***	-0.139***	-0.223***
	(0.005)	(0.005)	(0.016)	(0.022)	(0.006)	(0.017)	(0.009)
Unclassified	-0.030***	0.086***	-0.375***	0.280***	-0.115***	0.052*	0.225***
	(0.008)	(0.010)	(0.016)	(0.036)	(0.009)	(0.027)	(0.018)
Constant	3.561***	3.465***	3.929***	3.800***	3.593***	3.598***	3.390***
	(0.014)	(0.014)	(0.045)	(0.046)	(0.017)	(0.042)	(0.027)
Observations	578,578	427,262	45,681	18,457	379,365	54,971	144,242
R^2	0.024	0.026	0.040	0.034	0.034	0.020	0.032
Adjusted R^2	0.024	0.026	0.040	0.033	0.034	0.020	0.032

Notes to Table 7: Coefficients from OLS models estimating the star ratings given by reviewers. Analyses exclude reviews without star ratings. Standard errors are clustered at the school level and are shown in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 8: Examples of Positive and Negative Reviews by Parents and Students.

Review Type	Example
Positive student review for a traditional public school with more **objective** information **emphasizing life after high school**.	[School Name] graduated 651 in the class of 2013. Graduating students will be attending schools such as Dartmouth College (to which a student received a full scholarship), Yale University (one of two students attending received a full scholarship), Stanford University, University of Chicago, Oberlin College & Conservatory, Tulane University, Vassar College, University of Alabama, Arizona State University, Colorado State University, Florida A&M University and Miami University. Almost all Virginia schools are represented, as well as all of the armed forces. The class includes two National Merit Finalists, four National Achievement Scholarship Program Semi-Finalists and one National Achievement Scholarship Recipient.
Negative student review for a traditional public school that is more **subjective** and **does not discuss life after high school**	I don't like this school one bit. There is no way the school can improve. Some tips for others is not to go to this school period. Your life will be miserable if you go to this school. I can't explain in words how bad I hate this school.
Positive student review for the same traditional public school discussing **both academics and caring**.	I am currently a sophomore at [School Name] and I am thrilled to be a Cardinal. I am inside this school every day, and I can most definitely say that this school provides a sense of security as well as family. We are now under new administration and they will not put up with some of thestuff our old principal did. I gave [School Name] a 5/5 because of the academics, school pride, and athletics. GO CARDINALS!
Negative parent review for the same traditional public school discussing **caring, but not academics.**	Better hope your child does not have a problem with bullies because [School Name] does NOT care!!! You can't get any help from teachers or administrators or school board to deal with the problem either.

Notes to Table 8: Reviews have been edited to exclude the school name.

Appendices

A. Supplemental Tables
B. Methods for Traditional Qualitative Analyses
C. Analysis of Subtopics
D. Technical Appendix

Table A1: Regression Results Comparing the Likelihood of Review Types, Alternative Specification of Race

	(1) Physical Environment	(2) Resources	(3) School Staff	(4) Instruction and Learning	(5) Overall Quality	(6) School Culture	(7) School-level Features	(8) Unclassified
Reviewer Type:								
Principal	0.036**	0.042**	-0.127***	0.058***	-0.027**	0.052***	0.053***	-0.032***
	(0.013)	(0.013)	(0.013)	(0.012)	(0.010)	(0.011)	(0.013)	(0.009)
Student	-0.069***	0.002	-0.062***	-0.020***	0.014***	-0.076***	-0.032***	-0.199***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Teacher	-0.010**	-0.035***	0.002	0.030***	-0.013***	0.069***	0.004	-0.039***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
School Type:								
Charter	-0.001	-0.007	-0.057***	0.121***	0.004	0.027***	0.066***	0.023***
	(0.005)	(0.005)	(0.006)	(0.004)	(0.003)	(0.004)	(0.005)	(0.003)
Private	-0.086***	-0.043***	-0.123***	0.049***	-0.028***	-0.010	0.123***	0.005
	(0.011)	(0.011)	(0.010)	(0.011)	(0.008)	(0.010)	(0.011)	(0.007)
School Level:								
High	0.035***	0.063***	-0.086***	0.031***	0.009***	-0.019***	0.089***	-0.004
	(0.005)	(0.004)	(0.005)	(0.004)	(0.002)	(0.005)	(0.005)	(0.004)
Demographics:								
Students: Prop Black	-0.033***	-0.003	-0.008	0.033***	0.009*	-0.015**	0.017**	0.015**
	(0.006)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)	(0.006)	(0.004)
Students: Prop Hispanic	-0.058***	-0.018***	0.008	-0.012*	0.010**	-0.033***	-0.054***	-0.018***
	(0.006)	(0.005)	(0.005)	(0.006)	(0.003)	(0.005)	(0.006)	(0.004)
Students: Prop other race	0.029*	0.069***	0.046***	0.044***	0.028***	0.053***	-0.017	0.008

	(0.012)	(0.009)	(0.012)	(0.012)	(0.006)	(0.011)	(0.011)	(0.010)
Prop free lunch eligible	0.002	-0.050***	-0.015*	-0.067***	-0.046***	-0.008	-0.076***	0.013**
	(0.006)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)	(0.006)	(0.005)
Not Urban	0.014***	-0.025***	-0.002	-0.020***	-0.004*	-0.011***	-0.008*	-0.005*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
Constant	0.380***	0.345***	0.818***	0.608***	0.882***	0.726***	0.409***	0.897***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.005	0.007	0.022	0.034	0.004	0.014	0.042	0.050
Adjusted R^2	0.005	0.007	0.022	0.034	0.004	0.014	0.042	0.049

Notes to Table A1: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary/Middle, Urban, and Percent White (Students), and models include flags for each imputed data point and unknown categories. Standard errors are clustered at the school level and are shown in parentheses. ~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2: Regression Results Comparing the Likelihood of Review Types, Conditioning on Achievement (Public Schools Only)

	(1) Physical Environment	(2) Resources	(3) School Staff	(4) Instruction and Learning	(5) Overall Quality	(6) School Culture	(7) School-level Features	(8) Unclassified
Reviewer Type:								
Principal	0.065***	0.076***	-0.110***	0.100***	-0.005	0.075***	0.074***	-0.022~
	(0.018)	(0.018)	(0.017)	(0.017)	(0.013)	(0.015)	(0.019)	(0.012)
Student	-0.072***	0.005~	-0.062***	-0.008**	0.017***	-0.071***	-0.023***	-0.204***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Teacher	-0.014***	-0.037***	0.004	0.042***	-0.009**	0.076***	0.012**	-0.038***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)
School Type:								
Charter	-0.003	-0.015**	-0.054***	0.114***	0.004	0.025***	0.069***	0.022***
	(0.006)	(0.006)	(0.007)	(0.005)	(0.004)	(0.005)	(0.006)	(0.004)
School Level:								
High	0.043***	0.053***	-0.082***	0.020***	0.008*	-0.019**	0.110***	-0.004
	(0.006)	(0.005)	(0.007)	(0.005)	(0.003)	(0.006)	(0.006)	(0.006)
Demographics:								
Herfindahl Diversity Index	-0.010	0.050***	0.035***	0.018**	0.017***	0.050***	0.005	0.023***
	(0.007)	(0.006)	(0.007)	(0.006)	(0.004)	(0.006)	(0.007)	(0.005)
Prop free lunch eligible	-0.035***	-0.045***	-0.004	-0.047***	-0.019***	-0.017**	-0.063***	0.019***
	(0.007)	(0.005)	(0.006)	(0.006)	(0.004)	(0.006)	(0.007)	(0.005)
Not Urban	0.016***	-0.023***	-0.000	-0.024***	-0.005**	-0.005~	-0.004	-0.003
	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Standardized Achievement	-0.007	0.017***	0.014***	0.018***	0.025***	0.010**	0.026***	0.007*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)

Constant	0.388***	0.322***	0.804***	0.593***	0.867***	0.704***	0.381***	0.884***
	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.005)	(0.004)
Observations	514,846	514,846	514,846	514,846	514,846	514,846	514,846	514,846
R^2	0.006	0.008	0.024	0.013	0.004	0.011	0.014	0.048
Adjusted R^2	0.006	0.008	0.024	0.013	0.004	0.011	0.014	0.048

Notes to Table A2: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic, excluding any private schools. The omitted categories are Parent, TPS, Elementary/Middle, and Urban, and models include flags for each imputed data point and unknown categories. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table A3.1 - Regression Results Comparing the Proportion of Overall Topics in Reviews that Reference Each Topic

	Topics							
	(1) Physical Environment	(2) Resources	(3) School Staff	(4) Instruction and Learning	(5) Overall Quality	(6) School Culture	(7) School-Level Features	(8) Unclassified
Reviewer Type:								
Principal	0.009***	0.008**	-0.033***	0.012***	-0.011***	0.016***	0.010***	-0.010**
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Student	-0.010***	0.007***	-0.000	0.010***	0.043***	-0.005***	-0.000	-0.044***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Teacher	-0.001*	-0.007***	0.000	0.008***	-0.005***	0.019***	0.001	-0.015***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
School Type:								
Charter	-0.004***	-0.004***	-0.023***	0.026***	-0.007***	0.001	0.010***	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Private	-0.017***	-0.009***	-0.032***	0.009***	0.005	-0.001	0.028***	0.017***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.004)	(0.003)	(0.003)	(0.004)
School Level:								
High School	0.006***	0.013***	-0.028***	0.005***	-0.003*	-0.007***	0.018***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Demographics:								
Herfindahl Diversity Index	-0.003**	0.007***	0.002~	-0.003*	-0.004**	0.005***	-0.003**	-0.003*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prop. FRPL Eligible	-0.002*	-0.007***	0.005***	-0.007***	0.004***	0.002**	-0.013***	0.018***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Not Urban	0.005^{***}	-0.004^{***}	0.001^{*}	-0.004^{***}	0.000	0.000	-0.000	0.001^{*}
	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.066^{***}	0.056^{***}	0.170^{***}	0.114^{***}	0.192^{***}	0.140^{***}	0.071^{***}	0.192^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Proportion of overall topics that referenced each topic	0.06	0.06	0.16	0.12	0.19	0.14	0.08	0.19
Observations	673,761	673,761	673,761	673,761	673,761	673,761	673,761	673,761
R^2	0.003	0.006	0.023	0.017	0.018	0.002	0.030	0.013
Adjusted R^2	0.003	0.006	0.023	0.017	0.018	0.002	0.030	0.013

Notes to Table A3.1: Coefficients from OLS linear probability models estimating the proportion of overall topics in the review that referenced each topic. Sample excludes 3,355 reviews that didn't reference any of the topics. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and include unknown reviewer types, urbanicity, and school level. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A3.2 - Regression Results Comparing the Proportion of Overall Subtopics in Reviews that Reference Each Subtopic: Physical Environment and School Staff

	Physical Environment			School Staff	
	(1) Location	(2) Building Quality	(3) Facilities	(4) School Staff	(5) Teacher Quality
Reviewer Type:					
Principal	0.003*	0.000	0.004*	-0.031***	0.002*
	(0.001)	(0.000)	(0.002)	(0.002)	(0.001)
Student	-0.008***	0.001***	-0.000	0.005***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Teacher	0.001**	0.000	-0.002***	-0.001	0.002***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
School Type:					
Charter	-0.007***	-0.001***	0.002***	-0.025***	0.001***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Private	-0.012***	-0.001*	-0.006***	-0.033***	0.004***
	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
School Level:					
High	0.008***	-0.000**	-0.002***	-0.027***	0.003***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Demographics:					
Herfindahl Diversity Index	-0.008***	0.000~	0.004***	-0.002	-0.001***
	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
Prop free lunch eligible	-0.004***	0.000*	0.002***	0.004***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Not Urban	0.006***	0.000***	-0.001*	0.003***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)

Constant	0.029***	0.002***	0.027***	0.137***	0.006***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Proportion of overall topics that referenced each topic	.022	.002	.029	.122	.006
Observations	673,761	673,761	673,761	673,761	673,761
R^2	0.011	0.000	0.001	0.029	0.004
Adjusted R^2	0.011	0.000	0.001	0.029	0.004

Notes to Table A3.2: Coefficients from OLS linear probability models estimating the proportion of overall topics in the review that referenced each topic. Sample excludes 3,355 reviews that didn't reference any of the topics. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and include unknown reviewer types, urbanicity, and school level. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A3.3 - Regression Results Comparing the Proportion of Overall Subtopics in Reviews that Reference Each Subtopic: Instruction and Learning and Overall Quality

	Instruction and Learning			Overall Quality			
	(1) Instruction	(2) Curriculum/ Curricular Materials	(3) Learning Experience	(4) Evaluation	(5) Postsecondary/Graduation	(6) Preparation	(7) Quality Indicator
Reviewer Type:							
Principal	0.006***	0.003*	0.009***	0.001	0.007***	0.004***	-0.017***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.003)
Student	0.002***	0.003***	0.009***	0.003***	0.003***	0.001***	0.043***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Teacher	0.012***	-0.003***	0.003***	-0.003***	0.001*	0.001***	-0.004***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
School Type:							
Charter	0.008***	0.011***	0.015***	0.001*	0.005***	0.003***	-0.015***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Private	-0.001	-0.003*	0.010***	-0.005***	0.006***	0.003***	-0.002
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)
School Level:							
High	-0.004***	0.005***	0.001*	-0.011***	0.026***	0.004***	-0.008***
	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Demographics:							
Herfindahl Diversity Index	0.001**	0.006***	-0.007***	0.004***	-0.003***	-0.000~	-0.006***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.002)
Prop free lunch eligible	-0.005***	-0.006***	-0.003***	0.003***	0.003***	0.000**	-0.000
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)

Not Urban	-0.001*	-0.005***	-0.001	-0.001***	-0.000~	-0.000**	0.002*
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Constant	0.024***	0.027***	0.069***	0.032***	0.005***	0.002***	0.151***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Proportion of overall subtopics that referenced each subtopic	.023	.030	.073	.033	.012	.004	.147
Observations	673,761	673,761	673,761	673,761	673,761	673,761	673,761
R^2	0.006	0.008	0.013	0.006	0.059	0.008	0.021
Adjusted R^2	0.006	0.008	0.013	0.006	0.059	0.008	0.021

Notes to Table A3.3: Coefficients from OLS linear probability models estimating the proportion of overall topics in the review that referenced each topic. Sample excludes 3,355 reviews that didn't reference any of the topics. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and include unknown reviewer types, urbanicity, and school level. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table A3.4 - Regression Results Comparing the Proportion of Overall Subtopics in Reviews that Reference Each Subtopic: School Culture and School-level Features

	School Culture				School-level Features		
	(1) Student Discipline	(2) School Safety	(3) Interpersonal Relationships	(4) School Environment	(5) Religious	(6) School Finances	(7) School Type
Reviewer Type:							
Principal	-0.002*	-0.001	0.003~	0.011***	0.009***	-0.007***	0.010***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Student	-0.000	0.004***	0.002***	-0.006***	-0.001***	0.000	0.006***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Teacher	-0.001*	-0.002***	0.009***	0.017***	0.002***	-0.002***	0.003***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
School Type:							
Charter	0.001***	-0.002***	0.001*	-0.007***	0.001***	0.002***	0.001~
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)
Private	-0.000	-0.001	-0.004**	-0.002	0.026***	0.007***	-0.001
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
School Level:							
High	-0.000	-0.001***	-0.004***	-0.007***	-0.001***	0.003***	0.013***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Demographics:							
Herfindahl Diversity Index	0.001*	0.002***	0.005***	-0.000	-0.002***	0.000	-0.001*
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Prop free lunch eligible	0.006***	0.007***	0.002***	-0.008***	-0.000	-0.000	-0.009***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)

Not Urban	0.000	-0.000*	0.000	-0.000	0.001***	0.000	0.001~
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.005***	0.006***	0.032***	0.089***	0.003***	0.007***	0.038***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Proportion of overall subtopics that referenced each subtopic	.008	.010	.066	.083	.008	.010	.037
Observations	673,761	673,761	673,761	673,761	673,761	673,761	673,761
R^2	0.002	0.005	0.003	0.003	0.103	0.009	0.008
Adjusted R^2	0.002	0.005	0.003	0.003	0.103	0.009	0.008

Notes to Table A3.4: Coefficients from OLS linear probability models estimating the proportion of overall topics in the review that referenced each topic. Sample excludes 3,355 reviews that didn't reference any of the topics. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and include unknown reviewer types, urbanicity, and school level. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim} p < 0.10$, $^{*} p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$.

Table A3.5 - Regression Results Comparing the Proportion of Overall Subtopics in Reviews that Reference Each Subtopic: Resources and Unclassified

	Resources		Unclassified			
	(1) Extracurriculars/Electives	(2) Offerings	(1) Students	(2) Family	(3) School Choice/Enrollment	(4) Other
Reviewer Type:						
Principal	0.004**	0.003	-0.020***	0.003	-0.013***	0.011***
	(0.001)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Student	0.011***	-0.001*	-0.041***	-0.028***	0.002***	-0.008***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Teacher	0.002***	-0.007***	-0.014***	0.001*	-0.018***	0.003***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
School Type:						
Charter	-0.002***	-0.004***	-0.004***	0.001*	0.003***	0.011***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)
Private	-0.003**	-0.008***	0.005~	0.008***	0.007***	0.004**
	(0.001)	(0.001)	(0.003)	(0.002)	(0.002)	(0.001)
School Level:						
High	0.016***	-0.001***	-0.002**	-0.014***	0.002***	0.003***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)
Demographics:						
Herfindahl Diversity Index	0.000	0.006***	-0.005***	0.004***	0.000	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prop free lunch eligible	-0.006***	-0.002***	0.019***	-0.010***	0.014***	-0.005***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)

Not Urban	0.001*	-0.004***	0.003***	-0.003***	0.001*	-0.001~
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Constant	0.014***	0.035***	0.141***	0.054***	0.031***	0.035***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
Proportion of overall subtopics that referenced each subtopic	.016	.033	.138	.045	.038	.036
Observations	673,761	673,761	673,761	673,761	673,761	673,761
R^2	0.026	0.003	0.016	0.029	0.006	0.009
Adjusted R^2	0.026	0.003	0.016	0.029	0.006	0.009

Notes to Table A3.5: Coefficients from OLS linear probability models estimating the proportion of overall topics in the review that referenced each topic. Sample excludes 3,355 reviews that didn't reference any of the topics. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and include unknown reviewer types, urbanicity, and school level. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim} p < 0.10$, $^{*} p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$.

Table A4.1 - Regression Results Comparing the Likelihood of Review Types, Controlling on the Number of Words in the Review

	Topics							
	(1) Physical Environment	(2) Resources	(3) School Staff	(4) Instruction and Learning	(5) Overall Quality	(6) School Culture	(7) School-Level Features	(8) Unclassified
Reviewer Type:								
Principal	0.020	0.030*	-0.137***	0.043***	-0.033***	0.040***	0.037**	-0.041***
	(0.012)	(0.012)	(0.012)	(0.011)	(0.010)	(0.011)	(0.012)	(0.009)
Student	-0.067***	0.005*	-0.060***	-0.017***	0.015***	-0.073***	-0.030***	-0.197***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Teacher	0.014***	-0.015***	0.017***	0.053***	-0.004~	0.088***	0.026***	-0.027***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)	(0.002)
School Type:								
Charter	-0.049***	-0.044***	-0.084***	0.082***	-0.011***	-0.010***	0.026***	0.002
	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
Private	-0.096***	-0.046***	-0.125***	0.041***	-0.029***	-0.013	0.119***	0.004
	(0.009)	(0.009)	(0.010)	(0.010)	(0.008)	(0.010)	(0.010)	(0.007)
School Level:								
High School	0.021***	0.052***	-0.094***	0.019***	0.005**	-0.029***	0.078***	-0.010**
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Demographics:								
Herfindahl Diversity Index	-0.041***	0.020***	0.018***	-0.024***	0.004	0.017***	-0.034***	0.006~
	(0.004)	(0.004)	(0.005)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)
Prop. FRPL Eligible	-0.032***	-0.061***	-0.018***	-0.063***	-0.042***	-0.028***	-0.090***	0.011***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)

Not Urban	0.029***	-0.014***	0.005*	-0.013***	-0.002	0.002	0.006**	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
# of Words in the Review	0.003***	0.003***	0.002***	0.003***	0.001***	0.003***	0.003***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.149***	0.144***	0.670***	0.404***	0.799***	0.529***	0.200***	0.777***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
% Reviews with the Topic	36.08%	33.03%	76.05%	63.64%	87.07%	71.72%	42.74%	87.52%
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.188	0.130	0.101	0.175	0.047	0.140	0.171	0.134
Adjusted R^2	0.188	0.130	0.101	0.175	0.047	0.140	0.171	0.134

Notes to Table A4.1: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and include unknown reviewer types. Standard errors are clustered at the school level and are shown in parentheses. ~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A4.2: Regression Results Comparing the Likelihood of Review Types, Controlling on the Number of Words in the Review: Physical Environment and School Staff

	Physical Environment			School Staff	
	(1) Location	(2) Building Quality	(3) Facilities	(4) School Staff	(5) Teacher Quality
Reviewer Type:					
Principal	0.030**	0.005	0.013	-0.146***	0.021**
	(0.010)	(0.004)	(0.011)	(0.013)	(0.007)
Student	-0.069***	-0.000	-0.020***	-0.061***	-0.013***
	(0.002)	(0.001)	(0.002)	(0.003)	(0.001)
Teacher	0.018***	0.002~	0.005~	0.018***	0.025***
	(0.003)	(0.001)	(0.003)	(0.003)	(0.002)
School Type:					
Charter	-0.063***	-0.008***	-0.004	-0.086***	0.008***
	(0.002)	(0.001)	(0.002)	(0.004)	(0.001)
Private	-0.085***	-0.004	-0.041***	-0.136***	0.032***
	(0.007)	(0.003)	(0.008)	(0.010)	(0.006)
School Level:					
High	0.052***	-0.002**	-0.024***	-0.104***	0.021***
	(0.002)	(0.001)	(0.002)	(0.003)	(0.001)
Demographics:					
Herfindahl Diversity Index	-0.068***	0.002~	0.013***	0.003	0.003
	(0.003)	(0.001)	(0.003)	(0.035)	(0.018)
Prop free lunch eligible	-0.045***	0.001	-0.007**	0.020***	-0.010***
	(0.002)	(0.001)	(0.002)	(0.005)	(0.002)
Not Urban	0.040***	0.003***	-0.001	-0.014***	-0.011***
	(0.002)	(0.001)	(0.002)	(0.004)	(0.001)

# of Words in the Review	0.002***	0.000***	0.003***	0.006**	-0.005***
	(0.000)	(0.000)	(0.000)	(0.002)	(0.001)
Constant	0.099***	0.008***	0.032***	0.002***	0.001***
	(0.002)	(0.001)	(0.002)	(0.000)	(0.000)
% Reviews with the Subtopic	17.40%	1.96%	23.82%	75.39%	5.60%
Observations	677,116	677,116	677,116	677,116	677,116
R^2	0.095	0.004	0.157	0.103	0.024
Adjusted R^2	0.095	0.004	0.157	0.103	0.024

Notes to Table A4.2: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary, and Urban, and models include flags for each imputed data point. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim} p < 0.10$, $^{*} p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$.

Table A4.3: Regression Results Comparing the Likelihood of Review Types, Controlling on the Number of Words in the Review: Instruction and Learning and Overall Quality

	Instruction and Learning			Overall Quality			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Instruction	Curriculum/Curricular Materials	Learning Experience	Evaluation	Postsecondary/Graduation	Preparation	Quality Indicator
Reviewer Type:							
Principal	0.067^{***}	0.036^{**}	0.066^{***}	0.013	0.063^{***}	0.043^{***}	-0.061^{***}
	(0.011)	(0.011)	(0.012)	(0.011)	(0.009)	(0.007)	(0.011)
Student	-0.011^{***}	-0.001	-0.020^{***}	-0.019^{***}	0.018^{***}	0.004^{***}	0.020^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
Teacher	0.099^{***}	-0.006^{*}	0.043^{***}	-0.004	0.018^{***}	0.011^{***}	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
School Type:							
Charter	0.049^{***}	0.061^{***}	0.077^{***}	0.001	0.032^{***}	0.025^{***}	-0.021^{***}
	(0.002)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Private	0.000	-0.005	0.083^{***}	-0.043^{***}	0.056^{***}	0.020^{***}	-0.035^{***}
	(0.008)	(0.008)	(0.010)	(0.008)	(0.007)	(0.004)	(0.009)
School Level:							
High	-0.029^{***}	0.034^{***}	0.015^{***}	-0.075^{***}	0.191^{***}	0.035^{***}	-0.007^{***}
	(0.001)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
Demographics:							
Herfindahl Diversity Index	-0.002	0.021^{***}	-0.043^{***}	0.014^{***}	-0.030^{***}	-0.006^{**}	$0.006^{\sim}$
	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.002)	(0.003)
Prop free lunch eligible	-0.053^{***}	-0.064^{***}	-0.048^{***}	-0.007^{*}	0.010^{***}	0.001	-0.056^{***}
	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)

Not Urban	-0.004*	-0.028***	-0.004*	-0.007***	-0.000	-0.002*	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
# of Words in the Review	0.002***	0.003***	0.003***	0.003***	0.001***	0.000***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.064***	0.046***	0.290***	0.083***	-0.022***	-0.007***	0.768***
	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)
% Reviews with the Subtopic	19.40%	24.71%	52.17%	25.60%	10.19%	3.42%	83.19%
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.094	0.148	0.162	0.126	0.106	0.025	0.041
Adjusted R^2	0.094	0.148	0.162	0.126	0.106	0.025	0.041

Notes to Table A4.3: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary, and Urban, and models include flags for each imputed data point. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A4.4: Regression Results Comparing the Likelihood of Review Types, Controlling on the Number of Words in the Review: School Culture and School-level Features

	School Culture				School-level Features		
	(3) Student Discipline	(4) School Safety	(5) Interpersonal Relationships	(6) School Culture	(6) Religious	(7) School Finances	(8) School Type
Reviewer Type:							
Principal	-0.020***	-0.003	0.026*	0.063***	0.064***	-0.057***	0.073***
	(0.006)	(0.007)	(0.011)	(0.012)	(0.009)	(0.006)	(0.012)
Student	-0.009***	0.014***	-0.016***	-0.085***	-0.013***	-0.005***	0.001
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
Teacher	0.004*	-0.007***	0.066***	0.106***	0.016***	-0.005*	0.033***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
School Type:							
Charter	0.006***	-0.019***	-0.004~	-0.031***	0.003**	0.006***	0.006~
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.001)	(0.003)
Private	0.000	0.004	-0.038***	0.009	0.200***	0.049***	0.002
	(0.005)	(0.006)	(0.009)	(0.011)	(0.010)	(0.007)	(0.010)
School Level:							
High	-0.003***	-0.009***	-0.033***	-0.036***	-0.011***	0.015***	0.087***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Demographics:							
Herfindahl Diversity Index	-0.001	0.010***	0.023***	0.004	-0.019***	-0.005*	-0.019***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)	(0.004)
Prop free lunch eligible	0.038***	0.044***	-0.013***	-0.063***	-0.001	-0.010***	-0.087***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)

Not Urban	0.003**	-0.001	0.001	-0.005*	0.006***	0.004***	0.004~
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
# of Words in the Review	0.001***	0.001***	0.003***	0.002***	0.000***	0.001***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	-0.015***	0.003*	0.070***	0.419***	0.001	-0.015***	0.139***
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.001)	(0.003)
% Reviews with the Subtopic	6.54%	7.70%	26.29%	57.19%	6.24%	8.48%	28.35%
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.050	0.030	0.138	0.106	0.141	0.075	0.097
Adjusted R^2	0.050	0.030	0.138	0.106	0.141	0.075	0.097

Notes to Table A4.4: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary, and Urban, and models include flags for each imputed data point. Standard errors are clustered at the school level and are shown in parentheses. ~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A4.5: Regression Results Comparing the Likelihood of Review Types, Controlling on the Number of Words in the Review: Resources and Unclassified

	Resources		Unclassified			
	(1) Extracurriculars/ Electives	(2) Offerings	(1) Students	(2) Family	(3) School Choice/ Enrollment	(4) Other
Reviewer Type:						
Principal	0.043***	0.025*	-0.071***	0.001	-0.113***	0.079***
	(0.009)	(0.011)	(0.010)	(0.012)	(0.011)	(0.012)
Student	0.051***	-0.033***	-0.238***	-0.215***	-0.022***	-0.075***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
Teacher	0.027***	-0.031***	-0.036***	0.012***	-0.104***	0.045***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
School Type:						
Charter	-0.020***	-0.039***	0.002	0.002	0.012***	0.064***
	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Private	-0.009	-0.042***	0.001	0.035**	0.037***	0.037***
	(0.007)	(0.009)	(0.008)	(0.011)	(0.009)	(0.009)
School Level:						
High	0.110***	-0.014***	0.003	-0.094***	0.018***	0.017***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Demographics:						
Herfindahl Diversity Index	-0.005	0.027***	0.001	0.012**	-0.010**	-0.005
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)
Prop free lunch eligible	-0.057***	-0.038***	0.034***	-0.072***	0.067***	-0.056***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)

Not Urban	0.005***	-0.022***	0.007***	-0.019***	0.008***	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
# of Words in the Review	0.001***	0.002***	0.002***	0.003***	0.003***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.033***	0.112***	0.702***	0.216***	0.050***	0.092***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
% Reviews with the Subtopic	13.22%	26.61%	82.99%	34.11%	29.29%	29.43%
Observations	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.076	0.113	0.054	0.047	0.011	0.022
Adjusted R^2	0.076	0.113	0.054	0.047	0.011	0.022

Notes to Table A4.5: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic. The omitted categories are Parent, TPS, Elementary, and Urban, and models include flags for each imputed data point. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim} p < 0.10$, $^{*} p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$.

Table A.5: Regression Results Comparing the Likelihood of Review Types, using Qualitatively Coded Categories

	(1) Physical Environment	(2) Resources	(3) School Staff	(4) Instruction and Learning	(5) Overall Quality	(6) School Culture	(7) School-Level Features
Reviewer Type:							
Student	-0.013	-0.046	-0.176~	0.046	0.218**	-0.115	-0.083
	(0.052)	(0.081)	(0.092)	(0.086)	(0.075)	(0.089)	(0.067)
Teacher	-0.052	-0.159~	-0.035	-0.108	-0.029	0.167	-0.052
	(0.060)	(0.090)	(0.112)	(0.104)	(0.122)	(0.122)	(0.103)
School Type:							
Charter	-0.098**	0.005	-0.180*	0.074	0.147*	0.031	-0.080
	(0.034)	(0.076)	(0.077)	(0.080)	(0.073)	(0.084)	(0.057)
Private	-0.069~	-0.086	-0.257***	0.138~	-0.010	0.182*	0.091
	(0.040)	(0.068)	(0.078)	(0.078)	(0.080)	(0.079)	(0.076)
School Level:							
High	0.032	0.106	-0.106~	0.062	-0.051	-0.085	-0.003
	(0.040)	(0.065)	(0.063)	(0.062)	(0.067)	(0.068)	(0.055)
Demographics:							
Herfindahl Diversity Index	-0.018	0.093	-0.014	0.074	-0.148	-0.049	-0.037
	(0.084)	(0.117)	(0.108)	(0.113)	(0.125)	(0.129)	(0.105)
Prop free lunch eligible	0.071	-0.002	-0.083	-0.100	0.060	-0.220*	-0.088
	(0.065)	(0.091)	(0.077)	(0.093)	(0.091)	(0.098)	(0.078)
Not Urban	0.054	0.042	-0.029	0.082	-0.076	0.110	0.033
	(0.052)	(0.067)	(0.058)	(0.069)	(0.069)	(0.069)	(0.061)
Constant	0.107~	0.227**	0.910***	0.141~	0.611***	0.609***	0.288***

	(0.062)	(0.086)	(0.071)	(0.085)	(0.087)	(0.093)	(0.079)
Observations	500	500	500	500	500	500	500
R^2	0.034	0.022	0.109	0.112	0.052	0.082	0.080
Adjusted R^2	0.002	-0.011	0.080	0.082	0.021	0.051	0.050

Notes to Table A5: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic, as determined by qualitative coding. The omitted categories are Parent, TPS, Elementary/Middle school, and Urban. All models include flags for each imputed data point and include unknown reviewer types. Standard errors are clustered at the school level and are shown in parentheses. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.6: Regression Results Comparing the Likelihood of Review Types, Using School Fixed Effects

		Topics							
	(1) # of Words in the Review	(2) Physical Environment	(3) Resources	(4) School Staff	(5) Instruction and Learning	(6) Overall Quality	(7) School Culture	(8) School-Level Features	(9) Unclassified
Reviewer Type:									
Principal	6.896***	0.039**	0.047***	-0.124***	0.051***	-0.027**	0.051***	0.042**	-0.036***
	(1.619)	(0.013)	(0.013)	(0.012)	(0.013)	(0.009)	(0.013)	(0.013)	(0.009)
Student	-1.392***	-0.058***	0.000	-0.038***	-0.025***	0.015***	-0.062***	-0.051***	-0.183***
	(0.312)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)
Teacher	-4.857***	0.005	-0.020***	0.005	0.037***	-0.007*	0.070***	0.003	-0.033***
	(0.436)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)	(0.002)
Constant	69.194***	0.341***	0.318***	0.762***	0.640***	0.861***	0.701***	0.426***	0.892***
	(0.867)	(0.007)	(0.007)	(0.006)	(0.007)	(0.005)	(0.007)	(0.007)	(0.005)
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.234	0.171	0.169	0.183	0.194	0.151	0.168	0.208	0.202
Adjusted R^2	0.126	0.054	0.052	0.067	0.080	0.031	0.051	0.096	0.089

Notes to Table A6: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic with school fixed effects. The omitted category is Parent. All models include a flag for unknown reviewer types. ~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7.1: Regression Results Comparing the Likelihood of Review Types, Using School Fixed Effects: Physical Environment and School Staff

	Physical Environment			School Staff	
	(1) Location	(2) Building Quality	(3) Facilities	(4) School Staff	(5) Teacher Quality
Reviewer Type:					
Principal	0.040***	0.008~	0.025*	-0.134***	0.025***
	(0.011)	(0.004)	(0.012)	(0.012)	(0.007)
Student	-0.068***	0.000	-0.008***	-0.037***	-0.012***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)
Teacher	0.013***	0.002~	-0.005	0.006~	0.022***
	(0.003)	(0.001)	(0.003)	(0.003)	(0.002)
Constant	0.170***	0.021***	0.216***	0.758***	0.043***
	(0.006)	(0.002)	(0.006)	(0.006)	(0.004)
Observations	677,116	677,116	677,116	677,116	677,116
R^2	0.173	0.141	0.161	0.186	0.131
Adjusted R^2	0.056	0.020	0.043	0.071	0.008

Notes to Table A7.1: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic with school fixed effects. The omitted category is Parent. All models include a flag for unknown reviewer type. ~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7.2: Regression Results Comparing the Likelihood of Review Types, Using School Fixed Effects: Instruction and Learning and Overall Quality

	Instruction and Learning			Overall Quality			
	(1) Instruction	(2) Curriculum/Curricular Materials	(3) Learning Experience	(4) Evaluation	(5) Postsecondary/ Graduation	(6) Preparation	(7) Quality Indicator
Reviewer Type:							
Principal	0.077***	0.053***	0.078***	0.031*	0.069***	0.042***	-0.056***
	(0.011)	(0.012)	(0.014)	(0.012)	(0.008)	(0.005)	(0.011)
Student	-0.017***	-0.012***	-0.022***	-0.021***	-0.000	-0.001	0.022***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)
Teacher	0.090***	-0.014***	0.025***	-0.012***	0.006**	0.008***	-0.004
	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.001)	(0.003)
Constant	0.190***	0.235***	0.524***	0.246***	0.094***	0.028***	0.822***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.004)	(0.003)	(0.006)
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.154	0.181	0.184	0.160	0.220	0.152	0.147
Adjusted R^2	0.035	0.066	0.069	0.041	0.110	0.032	0.027

Notes to Table A7.2: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic with school fixed effects. The omitted category is Parent. All models include a flag for unknown reviewer type. ~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7.3: Regression Results Comparing the Likelihood of Review Types, Using School Fixed Effects: School Culture and School-level Features

	School Culture				School-level Features		
	(1) Student Discipline	(2) School Safety	(3) Interpersonal Relationships	(4) School Culture	(5) Religious	(6) School Finances	(7) School Type
Reviewer Type:							
Principal	-0.011	-0.003	0.053***	0.071***	0.047***	-0.048***	0.086***
	(0.007)	(0.008)	(0.012)	(0.014)	(0.006)	(0.008)	(0.012)
Student	-0.020***	0.002	-0.025***	-0.055***	-0.016***	-0.008***	-0.025***
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.002)	(0.002)
Teacher	-0.003~	-0.014***	0.052***	0.090***	0.009***	-0.009***	0.014***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)	(0.003)
Constant	0.069***	0.058***	0.240***	0.561***	0.066***	0.073***	0.277***
	(0.004)	(0.004)	(0.007)	(0.007)	(0.003)	(0.004)	(0.007)
Observations	677,116	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.149	0.155	0.153	0.167	0.381	0.149	0.183
Adjusted R^2	0.029	0.036	0.033	0.050	0.293	0.028	0.068

Notes to Table A7.3: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic with school fixed effects. The omitted category is Parent. All models include a flag for unknown reviewer type. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7.4: Regression Results Comparing the Likelihood of Review Types, Using School Fixed Effects: Resources and Unclassified

	Resources		Unclassified			
	(1) Extracurriculars/ Electives	(2) Offerings	(1) Students	(2) Family	(3) School Choice/ Enrollment	(4) Other
Reviewer Type:						
Principal	0.051^{***}	0.040^{**}	-0.069^{***}	$0.024^{\sim}$	-0.104^{***}	0.082^{***}
	(0.009)	(0.012)	(0.010)	(0.013)	(0.013)	(0.013)
Student	0.028^{***}	-0.024^{***}	-0.229^{***}	-0.181^{***}	-0.027^{***}	-0.069^{***}
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
Teacher	0.019^{***}	-0.035^{***}	-0.047^{***}	0.009^{**}	-0.109^{***}	0.030^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	0.134^{***}	0.251^{***}	0.854^{***}	0.375^{***}	0.258^{***}	0.286^{***}
	(0.005)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)
Observations	677,116	677,116	677,116	677,116	677,116	677,116
R^2	0.186	0.160	0.202	0.202	0.165	0.166
Adjusted R^2	0.071	0.042	0.089	0.089	0.047	0.048

Notes to Table A7.4: Coefficients from OLS linear probability models estimating the probability a reviewer referenced each topic with school fixed effects. The omitted category is Parent. All models include a flag for unknown reviewer type. $^{\sim}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A8: Regression Results Predicting Star Ratings, by User Type and School Sector, comparing the Proportion of Overall Topics in Reviews that Reference Each Topic

		Respondent Type			School Sector		
		(1)	(2)	(3)	(4)	(5)	(6)
	All	Parent	Student	Teacher	TPS	Charter	Private
Physical Environment	-1.412***	-1.617***	-0.489***	-1.433***	-1.535***	-1.466***	-0.630***
	(0.030)	(0.035)	(0.105)	(0.154)	(0.035)	(0.107)	(0.063)
Resources	1.051***	1.089***	1.308***	0.884***	1.162***	0.691***	0.994***
	(0.029)	(0.035)	(0.087)	(0.150)	(0.035)	(0.102)	(0.063)
School Staff	0.077**	-0.052	0.868***	-1.335***	0.484***	-0.630***	-0.631***
	(0.027)	(0.032)	(0.083)	(0.151)	(0.032)	(0.088)	(0.059)
Instruction and Learning	0.557***	0.532***	0.842***	0.012	0.172***	0.769***	1.154***
	(0.026)	(0.032)	(0.075)	(0.148)	(0.031)	(0.089)	(0.063)
Overall Quality	0.791***	0.765***	1.371***	-0.079	0.936***	0.494***	0.413***
	(0.024)	(0.031)	(0.065)	(0.140)	(0.028)	(0.082)	(0.058)
School Culture	-0.208***	-0.262***	0.157~	0.296~	-0.528***	0.148~	0.632***
	(0.026)	(0.032)	(0.081)	(0.151)	(0.031)	(0.087)	(0.061)
School-level Features	0.204***	0.173***	0.677***	-0.257~	0.415***	-0.508***	-0.851***
	(0.029)	(0.034)	(0.091)	(0.147)	(0.034)	(0.099)	(0.060)
Constant	3.696***	3.780***	3.143***	4.499***	3.592***	3.814***	3.958***
	(0.016)	(0.019)	(0.059)	(0.095)	(0.019)	(0.054)	(0.037)
Observations	575,803	425,807	45,197	18,369	377,325	54,684	143,794
R^2	0.017	0.018	0.025	0.022	0.024	0.016	0.018
Adjusted R^2	0.017	0.018	0.024	0.022	0.024	0.016	0.018

Notes to Table A8: Coefficients from OLS models estimating the reviewer rating, conditioning on the proportion of topics covered in each category. We exclude unclassified reviews and reviews without ratings. Standard errors are clustered at the school level and are shown in parentheses. ~p<.10, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A9: Regression Results Predicting Star Ratings, by User Type and School Sector, Controlling on the Number of Words in the Review

		Respondent Type			School Sector		
		(1)	(2)	(3)	(4)	(5)	(6)
	All	Parent	Student	Teacher	TPS	Charter	Private
Physical Environment	-0.131***	-0.143***	-0.111***	-0.095***	-0.144***	-0.098***	-0.006
	(0.005)	(0.006)	(0.017)	(0.023)	(0.006)	(0.017)	(0.009)
Resources	0.343***	0.352***	0.284***	0.343***	0.398***	0.286***	0.266***
	(0.005)	(0.005)	(0.016)	(0.023)	(0.006)	(0.016)	(0.009)
School Staff	0.095***	0.100***	0.034$^{\sim}$	-0.189***	0.234***	-0.073***	-0.054***
	(0.006)	(0.007)	(0.018)	(0.025)	(0.008)	(0.018)	(0.010)
Instruction and Learning	0.281***	0.291***	0.111***	0.253***	0.175***	0.377***	0.475***
	(0.005)	(0.006)	(0.017)	(0.025)	(0.006)	(0.018)	(0.012)
Overall Quality	0.591***	0.599***	0.518***	0.351***	0.625***	0.471***	0.496***
	(0.008)	(0.008)	(0.030)	(0.034)	(0.009)	(0.026)	(0.015)
School Culture	0.092***	0.108***	-0.078***	0.341***	-0.019**	0.227***	0.339***
	(0.005)	(0.006)	(0.016)	(0.029)	(0.006)	(0.018)	(0.011)
School-level Features	0.181***	0.196***	0.095***	0.121***	0.224***	0.058***	-0.081***
	(0.005)	(0.005)	(0.016)	(0.022)	(0.006)	(0.017)	(0.009)
Unclassified	0.130***	0.266***	-0.245***	0.381***	0.060***	0.201***	0.365***
	(0.008)	(0.010)	(0.017)	(0.035)	(0.009)	(0.027)	(0.018)
# of Words in the Review	-0.007***	-0.008***	-0.005***	-0.006***	-0.008***	-0.007***	-0.006***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	3.376***	3.253***	3.814***	3.649***	3.403***	3.403***	3.220***
	(0.014)	(0.014)	(0.044)	(0.046)	(0.017)	(0.042)	(0.027)
Observations	578,578	427,262	45,681	18,457	379,365	54,971	144,242

R^2	0.070	0.026	0.040	0.034	0.034	0.020	0.032
Adjusted R^2	0.070	0.026	0.040	0.033	0.034	0.020	0.032

Notes to Table A9: Coefficients from OLS models estimating the reviewer rating, conditioning on the number of words in reviews. We exclude reviews without ratings. Standard errors are clustered at the school level and are shown in parentheses. ~p<.10, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A10: Regression Results Predicting Star Ratings, by User Type, using School Fixed Effects

		Respondent Type		
		(1)	(2)	(3)
	All	Parent	Student	Teacher
Physical Environment	-0.305***	-0.315***	-0.195***	-0.088**
	(0.005)	(0.005)	(0.020)	(0.029)
Resources	0.194***	0.189***	0.214***	0.110***
	(0.005)	(0.005)	(0.019)	(0.029)
School Staff	-0.009~	-0.007	-0.044*	-0.209***
	(0.005)	(0.006)	(0.021)	(0.035)
Instruction and Learning	0.076***	0.068***	-0.012	0.055~
	(0.005)	(0.006)	(0.020)	(0.031)
Overall Quality	0.451***	0.436***	0.431***	0.232***
	(0.007)	(0.008)	(0.030)	(0.041)
School Culture	-0.077***	-0.066***	-0.126***	0.230***
	(0.005)	(0.006)	(0.021)	(0.037)
School-level Features	-0.023***	-0.037***	0.015	-0.023
	(0.005)	(0.005)	(0.019)	(0.029)
Unclassified	-0.068***	0.078***	-0.287***	0.179***
	(0.007)	(0.010)	(0.022)	(0.044)
Constant	3.618***	3.536***	3.750***	3.844***
	(0.009)	(0.011)	(0.031)	(0.054)
Observations	578,578	427,262	45,681	18,457
$R2$	0.259	0.299	0.556	0.801
Adjusted $R2$	0.137	0.151	0.190	0.433

Notes to Table A10: Coefficients from OLS models estimating the reviewer rating, using school fixed effects. We exclude reviews without ratings. Standard errors are clustered at the school level and are shown in parentheses. ~p<.10, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix B: Methods for Traditional Qualitative Analyses

This section describes the methods used for the three separate analyses we did using traditional qualitative coding methods on a subsample of the reviews.

QUESTION 1

Can we get more information about what the Unclassified reviews are capturing? Are there differences by students and parents in the Unclassified reviews? Are the topics brought up in Unclassified reviews different by star rating?

METHOD

Sample: Obtain a random sample of 575 reviews coded as including Unclassified words that are within the 25th to 75th percentile of length (about 30-107 words) and have a star rating among the following categories:

1) were written by a parent and given a low rating (144)
2) were written by a parent and given a high rating (144)
3) were written by a student and given a low rating (143)
4) were written by a student and given a high rating (144)

Coding Scheme:

For each review, coders will:

- Step 1:
  - Read over the key words/phrases that were classified in L_1 as "Unclassified" and locate them in the review to determine what the words are conveying. Describe this meaning using the following categories:
    - Irrelevant: the unclassified reviews are not relevant to the meaning of the review
    - Already defined categories: Level 1: Physical Environment, Resources, School Staff, Instruction & Learning, Overall Quality, School Culture, School-level Features, Level 2: Evaluation, Postsecondary/Graduation, Preparation, Quality Indicator, School Choice/Enrollment, etc.
    - New categories: School Selection (related to how the reviewer picked the school), Physical Resources (school supplies and other resources available to students that are not already captured by the curriculum resources), Others that come up in the process.
  - Determine how important the unclassified words are to the overall meaning of the review: 0=not important (i.e. if we took the words out of the review, the meaning would stay the same), 1=somewhat important (i.e. the words provide important context for another category represented, but they themselves are not adding much to the review), 2=important (i.e. these words signify a meaning to the review that would otherwise be missed if we took them out)
    - Put any words/phrases that are used in a way that may not be picked up by the NLP (eg. other uses for words, misspellings, slang)
- Step 2:
  - Determine whether the review touches on topic areas not already included in the NLP coding scheme. 0=review does not cover a new category or topic, 1=review suggests a new category/topic to consider

- Step 3:
    - Examine patterns in highlights/categories/topics that differ between high and low ratings
    - Examine patterns in highlights/categories/topics that differ between parents and students
    - Examine patterns in highlights/categories/topics that differ between parents low ratings and student low ratings, parent high ratings and student low ratings

<u>Process:</u>

- Begin with two interns coding the same 20 reviews to refine coding scheme and ensure coder reliability. Discuss any differences in coding and resolve discrepancies.
- Split remaining reviews evenly between two interns. Discuss with supervisor any questions that remain to ensure consistent coding

QUESTION 2

Does our identification of topics have face validity? Are the codes included leading to the differentiation?

METHOD

<u>Sample:</u> Obtain a random sample of 500 reviews in the 25th to 75th percentile of length

- Provide a summary of how each review was coded by NLP (which categories are noted in the review?)
- Include what words in the review were picked up by NLP
- Include indicators of sector, school type, reviewer type, and star ratings

<u>Coding Scheme:</u>

- Step 1: Examine the words highlighted by ~ in the review to determine what the review discusses
- Step 2: Indicate any words not highlighted by ~ that help to determine what the review discusses
- Step 3: Determine which categories the review mentions
- Step 4: Compare the categories the reviews include as coded by NLP and by the interns
    - For reviews that are coded differently, pick out the words/phrases used to determine their categorization

<u>Process:</u>

- Use a review ID to distinguish reviews and remove indicators of what categories the reviews were coded as
- Begin with two interns coding the same 20 reviews to refine coding scheme and ensure coder reliability. Discuss any differences in coding and resolve discrepancies.
- Split remaining reviews evenly between two interns. Have them discuss any questions together to ensure consistent coding
- Merge back in the indicators of what the reviews were categorized as and compare coding

QUESTION 3

Why do different groups give different types of reviews? How do the reviews for the same school differ by user type? How do reviews for the same category of topic differ by user type?

METHOD

Sample: Start with reviews that are within the 25th and 75th percentile of length from parents and students, focusing only on high schools and only keep schools that have at least 5 parent and student reviews. Obtain a random sample of 48 of these schools, including 14 charter schools, 17 private schools, and 17 TPS. Within these 48 schools, get a random sample of 10 reviews (5 parent, 5 student) per school, with 480 overall reviews in the sample.

- Provide a summary of how each review was coded by NLP
- Include school-level characteristics in the dataset

Coding Scheme:

- Step 1: For each school, separate the reviews by the user types, school types, and categories
- Step 2: For each category, highlight keywords/phrases in the review to determine what the review discusses, using the NLP coding scheme as a guide.
- Step 3: Compare the words/phrases used by different kinds of users for each category along the following dimensions:

Description	**Values**
Proximity to classroom experience: Does the review discuss mostly classroom experiences (interactions with teachers, learning experiences, academic preparation, etc.) or does it discuss the entire character of the school (great school overall without mentioning learning) or other aspects of schooling (interactions with the office, communications, extracurriculars, etc.)?	0=Does not discuss classroom experience, 1=Discusses classroom experiences
Personal Example: Does the review discuss a personal story (interactions with a specific teacher, something that happened to them, etc.)? Or does it include more general information about the school?	0=no personal story, 1=includes a personal story
Objectivity: How much does the review capture objective aspects of education – i.e. numbers and stats and verifiable facts about the school – versus subjective aspects of education – i.e. feelings, tastes, opinions others may not agree with?	0=subjective, 1=objective
College Preparation: How much does the review discuss the school preparing students for college?	0=doesn't mention college preparation, 1=mentions college preparation
Preparing for Life: How much does the review discuss the school preparing students for their later lives, including being a good person or citizen or employee?	0=doesn't mention preparing students for life after education, 1=mentions preparing students for life after education
Broad: Does the review only include broad information about the school? Does it leave out specific aspects of the review?	0=specific examples, 1=broad statements

Academics: Does the review focus on academic excellence and instruction as essential?	0= does not discuss academics, 1=discusses academics
Caring: Does the review focus on caring and a welcoming environment as essential? Does it bring up social and emotional learning or the wellbeing of students? More than teachers who are dedicated, but who show they care about students.	0=does not discuss the caring/welcoming environment, 1=discusses the caring/welcoming environment
Comparisons: How much does the review compare the school to other schools or educational experiences to describe its qualities versus focusing on its own qualities without making comparisons? Does the review put the school in context with other educational options? Are there are comparison words used?	0=no comparisons, 1=includes a comparison

Process:

- Begin with all four interns coding the same reviews to refine coding scheme and ensure coder reliability. Discuss any differences in coding and resolve discrepancies.
- Split remaining schools evenly in two and assign two interns to each set of reviews. Have them discuss any questions together to ensure consistent coding
- After completing coding:
    - Compare the words/phrases used by different user types within categories across school types (i.e., how do users describe the same categories across school types?)
    - Compare the words/phrases used within different categories by the same user types across school types (i.e., how do the same kinds of users describe different categories across school types?)

Appendix C: Analysis of Sub-Topics

This section describes the analyses of the sub-topics by subgroup. Tables 5A-5D show the full main results and Appendix Tables A3, A4, and A7 display the alternative specifications. As for the main analyses, we focus on the results that are robust to these alternative specifications, but mention notable differences below. Rather than summarize findings for each subgroup-by-subtopic, we focus only on what we considered the most noteworthy patterns. We begin with an overall summary of the patterns we observed, followed by more detail description of patterns by broad topic area.

Summary of Results

The differences in sub-topic references that were the largest and most robust to different specifications emerged between school types. What aspects of education appear to be crucial for parents in selecting a charter or private school over a TPS? Reviews of charter schools discuss instruction, curriculum, evaluation, facilities, student discipline and interpersonal relationships more than reviews of TPS. These patterns suggest that charter school instructional methods and school culture set them apart from TPS, and are areas that charter schools have more autonomy over than TPS. In contrast, reviews of TPS tend to focus more on location and school staff. Reviews of both private and charter schools discussed the learning experience and how these schools prepare students for the future more than for TPS, but private school reviews focused less on facilities and interpersonal relationships and more on religion and teacher quality than TPS. Interestingly, school resources, including extracurriculars and offerings, did not appear to distinguish school types.

Differences between parents and students in the topics covered in reviewers were less robust to models that considered the number of categories discussed and review length. In these main results, students appear to write less often than parents about interpersonal relationships, school choice, evaluation, instruction, the learning experience, and school staff, but these relationships are all positive and significant in models that account for the number of topics discussed (appendix table A.4.1.2-5). We return to these relationships in our qualitative analyses.

One topic area that did differ significantly across all analyses was in resources. Parents wrote the least about extracurricular activities and the most (other than principals) about course offerings. High school reviews, not surprisingly, include much more about extracurricular activities and electives, which may play into part of these patterns given the majority of student reviews were for high schools.

Notable Results by Broad Topic Area

- *Physical Environment*: Students write the least about location and principals write the most. Charter and private school reviews involve less discussion about location; this may be because traditional public schools are tied to their neighborhoods and closer to home. have a more direct connection to their neighborhoods.
- *School Staff*. Principals write much less about school staff—presumably because they find it difficult to write about themselves—but they write more about their teachers. Teachers also write more about teachers, but, in that case, they can write about teachers collectively (e.g., “great

colleagues”). While it appears students write less about school staff than parents in the main models here, the relationship is flipped when we consider the number of categories in reviews (Appendix table A3.2). Charter and private school reviews have more text about teachers compared with TPS, and much less about school staff. Reviews of schools serving low-income students have less text about teachers.

- *Instruction/Learning*. Principals and teachers write more about modes of instruction and the learning experience in the classroom than parents who rarely spend time in classrooms and are less aware of these elements. Results for students are inconclusive, as models that consider the number of topics discussed in a review display the opposite results as those presented here. Reviews of charter schools include more about all three subcategories for instruction and learning than TPS. Private school reviews also include more about the learning experience than TPS.
- *Overall Quality*. Principals write the most about evaluation, graduation/postsecondary, and preparation–and the least about quality indicators. That is, it seems that principals are more specific than others in their comments on quality. Parents and students, in contrast, are more likely to use broad quality indicators. Charter and private school reviews are more likely to mention graduation/postsecondary and preparation.
- *Culture*. Principals and teachers write the most about interpersonal relationships and the school environment, and students write the least. Charter school reviews have more text about student discipline, which may reflect that they are known for being more strict. More racially homogenous school reviews have less text about school culture overall–but also every element of school culture.
- *Features*. Private school reviews are naturally much more likely to mention religion, but also school finances). Charter school reviews also discuss school finances more often than TPS, but to a much lesser extent than private schools.
- *Resources.* Parents write the least about extracurricular activities and electives, but more than students and teachers about offerings. Private school reviews also include less about offerings than reviews of TPS.
- *Unclassified.* Parents write the most about school choice, however, when considering the number of categories included in reviews, students write more about school choice than parents (see appendix table A3.5). Not surprisingly, reviews of charter and private schools write more about school choice and enrollment than reviews of TPS. They also include more about family.

Appendix D: Text Analysis Technical Details

Data Overview and Details:
Our dataset from GreatSchools consists of the following aspects which we consider important for analysis. As is common in NLP we will refer to a *corpus* which in the following we take to mean all the text within the reviews of the GreatSchools dataset.

Years: 2002-2019
Sector: Public, charter, private,
Source: Parent, student, teacher, principal, other, N/A
Level: Pre-K, elementary, middle, high
Number schools with text only reviews: 83,795
Number of text reviews with ratings: 578,667
Total number of text reviews: 677,210

We are interested in a set of specific **aspects** or slices of the data. Specifically:

- **School Type**: Charter, Traditional Public, Private
- **Review Quality**: Good Review (4,5), Bad Review (1,2,3)
- **School Quality**: Well Rated School (>= 4), Poorly Rated School (< 4).

Looking at the Distribution of Reviews we can see how the school's vary in number of reviews, school ratings, and review ratings when broken down by aspect.

Sub-domain	Number of Reviews (%)	Number of Schools with Reviews (%)	Mean Number of Reviews per School ± Std. Deviation	Median Number of Reviews per School	Max Number of Reviews per School	Number of Ratings (% of rated reviews)	Mean Rating ± Std. Deviation
Full Corpus	677210 (100%)	83,795 (100%)	8.08 ± 14.8	5	1685	578667 (100%)	3.89 ± 1.02
Charter	64054 (09.5%)	3987 (4.76%)	16.07 ± 25.75	8	658	54978 (9.50%)	3.84 ± 1.60
Private	162288 (24.0%)	17092 (20.4%)	9.495 ± 12.93	6	545	144260 (24.9%)	4.12 ± 1.49
Trad. Public	450868 (66.6%)	62930 (75.1%)	7.165 ± 13.99	4	1685	379429 (65.6%)	3.79 ± 1.58
Good Review	404488 (59.7%)	76084 (90.8%)	4.829 ± 0.2676	5	562	404488 (69.9%)	4.85 ± 0.36

Bad Review	174179 (25.7%)	55134 (65.8%)	3.159 ± 3.256	2	78	174179 (30.1%)	1.63 ± 0.79
Well Rated School	356378 (52.8%)	45303 (16.9%)	7.87 ± 18.0	4	1685	291507 (50.3%)	4.62 ± 0.39
Poorly Rated School	318218 (47.0%)	36326 (28.0%)	8.76 ± 9.64	6	204	287160 (49.6%)	2.97 ± .79

Note that the aspect of *Good Review*/*Bad Review* refers to the number of stars assigned by *a particular reviewer* (1 - 5). In order to have a proxy measure for quality and determine the *Well Rated School / Poorly Rated School* we take the *average of all reviews for a particular school* and call a Well Rated School a school with >=4 review average and a Poorly Rated School a school with a < 4 review average. This has a weak correlation with the SEDA Overall Average Rating when looking at public schools only. Note that this rating is over the whole dataset and cannot be sliced by year as there would be insufficient ratings for any particular school or year.

Additional Details on Data Cleaning from Section III.A General Methods

For *Step 1. Cleaning and Preprocessing the Data* in addition to the steps listed in the main document we:

Text Wrangling and Preprocessing - Our first step is to create a data table that includes all the meta-data about the reviews including school, location, who wrote it, etc., and the review text itself. This was achieved by matching review data from GreatSchools with the information in the NLSD (Carroll, Harris, Nair, Nordgren, 2023).

Text Normalization and Tokenization - Given a large volume of user generate text we need to remove the many "non-words" that appear, e.g., "49schtop" and others. To do this we first run a spell checker over the entire corpus and replace any misspelled word with the most likely correction (McCallum & Sondej 2021). After this we normalize the text by converting all letters to lower-case and removing all non-alphanumeric characters. For example, "Don't" becomes "dont", "STOP!" becomes "stop", and "Country-Day" becomes "country day." This has the effect of removing sentence boundaries but these are not needed in our later analysis. We then tokenize the text, that is, we break words up at each blank space character, and this gives us the full set of words or vocabulary of our corpus.

Removing Stop Words and Proper Nouns - We remove common stop words such as "is", "a", "the", and many others (Bird, Klein, & Loper, 2009; Manning, Rghavan, Schütze, 2009). We also use a list of common names and proper nouns to remove these words from the text and replace them with tokens that are common across all reviews, specifically <PLACE> and <PERSON>. This was done by (1) Taking all the city and state names in the NLSD, (2) Taking all the common city names in the US[26], and (3) taking the most common Surnames, male names, and female names from the US Census.[27]

For example "ms edwards" is replaced with "ms <PERSON>". This process is a bit noisy as many schools are named after specific people. We attempted to control for this by replacing any exact statements of the school name in the review with <SCHOOLNAME> but since many schools are referenced colloquially in the text, e.g., Forrest Elementary is referred to as forrest, this is not a perfect solution and many school names, person names, and place names are not captured exactly.

For *Step 2: Identify Keywords* we provide some additional details on how the log-likelihood keywords are found.

Using the full corpus as the reference corpus, we split our text into 6 sub-corpora each associated with a particular *aspect* of the data and performed this log-likelihood estimate on each of these aspects compared to the full corpus: Charter reviews, Trad. Public reviews, Private reviews, reviews of well rated schools, reviews of poorly rated schools, good reviews, and bad reviews. This left us with a list of 1,800 words that were deemed significant.

Using a log-likelihood estimate -- which is a score measuring relative frequency assigned to each word that compares the frequency of a word within a corpus with another reference corpus. Specifically, this highlights which words appear 'unusually' frequently, with 'usual' frequency being determined by the reference corpus. Using a p-value of .01 which is broadly used in the literature (Rayson & Garside 2000) -- statistically significantly relevant keywords were selected for each domain. These keywords are words that the frequency calculations have determined to be unusually frequent within our corpus, meaning they are indicative of that corpus relative to our reference corpus.

Log likelihood is calculated by:

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

[26] https://en.wikipedia.org/wiki/List_of_the_most_common_U.S._place_names
[27] https://namecensus.com/

'c' corresponds to the number of words in a subcorpus (related to a aspect), and 'd' corresponds to the number of words in the full corpus (all of our data). Terms 'a' and 'b' are observed values (O), and to calculate the expected values (E) we can use the formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

We have N1 = c, and N2 = d. So the expected value of a word, E1 = c*(a+b) / (c+d) and E2 = d*(a+b) / (c+d). Then get the log-likelihood value with the formula:

$$-2\ln \lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equates to calculating log-likelihood G2 as follows: G2 = 2*((a*ln (a/E1)) + (b*ln (b/E2))).

Log-likelihood calculations were done using AntConc (Anthony 2023).

For *Step 4: Annotate Reviews with Codes* some additional remarks about why we chose not to lemmatize our corpus is in order.

Note that there is a subtle issue with how these reviews are applied, consider the review sentence: "my son enjoys school. i am amazed how well school is ran with all the construction. the kids all seem to get along well and most of the teachers seem to care." Note that in our topic coding we make a distinction between the verb "teach" and the noun "teachers". Hence, we must match words at the word boundary (i.e., the spaces) completely, otherwise if we were to use a method that ignored word boundaries (e.g., when you search on a webpage) then the word "teach", which is an Instruction and Learning keyword, would match the word "teachers" (since the word teach is contained in the word teachers), and we would have erroneously applied the Instruction & Learning tag to the above review in addition to the correct School Staff tag for "teachers."

When applying topic codes to the reviews one must be careful about forms. In our coding the word "teach" is coded into L1 Instruction & Learning since the verb form implies the act of teaching. However, the word "teacher" is coded into L1 School Staff since the noun form is referring to the teacher themselves. Issues like this is why we did not apply the common NLP technique of *stemming* or *lemmatization* to our preprocessing of the text. In both stemming and lemmatization one is attempting to remove inflected forms of words so that only a root word remains. Stemming is a less intensive process and the most standard Porter stemmer (Porter 1980) applies a set of rules to reduce words to their basic forms, e.g., caresses would become caress. Lemmatization requires a full morphological analysis of the sentence and attempts to reduce both nouns and verbs to their most basic form, e.g., am, are, and is, would all be changed to the root verb be. In our analysis, the form of nouns and verbs are important, as we saw with the example of teach v. teachers above. Hence we chose not to apply this common technique. So when annotating the reviews we only apply the code if exactly the entire word is present in the review.

For *Step 3. Apply Topic Codes to Keywords* we provide some additional remarks about the packages and algorithms used.

In order to find these closest semantic neighbors for how words are used specifically in our corpus of review text, we used the popular *word2vec* method from the literature (Mikolov, Chen, Corrado, & Dean 2013; Mikolov, Sutskever, Chen, Corrado, & Dean 2013).[28] One of the most common examples of this technique is predictive text on your smartphone or email that is tuned to how you, individually, use language. In order to do this, one takes a large corpus of text and looks at the co-occurrence of sets of words, that is, how often different groups of words occur in the same sequence across all the reviews. To build this model, we create sentence fragments where, for every word in every review, we extract the 5 words before and after that word which gives the *context*. We then build a predictive model, using a neural network, that takes these 10 words and attempts to predict the missing target word, a "fill in the blanks" test. Once we are able to do this well (i.e., low error) we can use the vector embeddings learned in this process as words that are used similarly will have similar vector embeddings. We then define the *semantic neighborhood* of a word as its 10 nearest words according to cosine similarity of their respective vector embeddings. For example, the word "strict" tends to have the same words around it as words like "lenient" and "militant," which signals that the term is used to refer to student discipline.

We first used the simple sentence tokenizer from Spacy[29], a rule based sentence extraction method to extract all individual sentences across all reviews. We then used the GenSim package for Python and their built in Word2Vec model which makes use of a continuous bag of words (CBOW) model with an embedding vector size of 100 and minimum word frequency of 25 and a context window of 5 words before and after a target word.

Consider the sentence "The administrator regularly uses a bullhorn to yell at the students" if we were building a word2vec model using the CBOW methods then given a neighborhood set of 3, N = {-3, -2, -1, 1, 2, 3} we want to find MAX \Pi_{i \in C} Pr(w_i | w_j: j \in N + i) (Jurafsky & Martin, 2024), that is, we want to find a predictive model that maximizes the probability of a word given a neighborhood. So for the example sentence if our target word is 'bullhorn' then the context would be [regularly, uses, a, to, yell, at]. Once our model is fit using backpropagation we can use the resulting vector embeddings to determine similarity.

References

Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

---

[28] We used the GenSim Library: https://radimrehurek.com/gensim/models/word2vec.html which is implemented from an open source Google Project: https://code.google.com/archive/p/word2vec/.

[29] https://spacy.io/api/sentencizer

Carroll, J. M., Harris, D. N., Nair, A., Nordgren, E. (2023). Introducing the National Longitudinal School Choice Database. National Center for Research on Education Access and Choice (REACH).

Manning, C. D., Rghavan, P., Schütze, H. (2009). An introduction to information retrieval. Cambridge university press.

McCallum, J., Sondej, F. (2021). The Python autocorrect Package, Version 2.6.1. https://pypi.org/project/autocorrect/

Jurafsky, D., & Martin, J. H. (2024). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd Edition Draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6.