REACH
National Center for
Research on Education
Access and Choice

Can User Reviews Like Those on GreatSchools Improve Information for Schooling Choices?

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

Policy Brief | November 2024

Summary

Families are being given more and more choices of schools, from charter schools to virtual schools and private schools, through school vouchers. However, families often report not being well informed about these choices. Many report being awash in information, though not necessarily the information they want or need. State, district, and school websites, along with Facebook pages and word-of-mouth, provide additional information but not necessarily a coherent picture of the various schooling options available to families.

Massive school information and review platforms add another source of information. These platforms provide access to much of the same data found on other public websites–test scores, student demographics, graduation rates, and more–combined with text reviews and star ratings provided by students, parents, educators, and other stakeholders.

We studied the content and usefulness of user reviews from one of the most widely used platforms, GreatSchools, including 50 million words of text across more than 600,000 reviews written about 84,000 schools from 2009 to 2019. Using a combination of machine learning (natural language processing) and qualitative analysis, we draw the following conclusions about the content of the user reviews:

- Text reviews of all types have the advantage of bringing in rich data on actual user experiences, but they also suffer from a number of significant limitations, including presenting the views of only a small group of self-selected stakeholders who post on these platforms.

- The most common topics discussed in the text reviews are *Overall Quality* and *School Staff*. The reviews focusing on *Overall Quality* are typically vague, using words like "wonderful" or "awful," but provide little information about schools' specific qualities, strengths, and weaknesses. In other words, the potential richness of text reviews is rarely realized in practice.

- GreatSchools text reviews vary in length and number across user types (parents, students, teachers, principals), school sectors (traditional public, charter, and private schools), grade levels, and demographics, making it difficult for users to make sense of the reviews. For example, parents tend to write more than other users, and reviews of charter schools are longer than those of Traditional Public Schools (TPS).

- The topics discussed in the reviews vary significantly by user types and school sectors. For example, compared with TPS, 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 traditional

public schools.

- The relationship between the topics that users discuss and their star ratings also differ by user and school type. 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. This reinforces that reviewers base their reviews on different educational values.

Taken together, these findings suggest that user reviews have the potential to inform consumers about school choices. However, they face many hurdles, some of which have not been well recognized. We propose ways to improve school user reviews and discuss implications for school choice.

Background: User Reviews and GreatSchools

User reviews have become an important and popular way for individuals to learn about consumer goods, travel destinations, contractor services, and more. The rise of Amazon is the most widely used example of this revolution in consumer information and purchasing. While these reviews can be the source of relevant information, there are many reasons to expect that the reviews are not representative of all users. Reviews tend to be written by individuals with either a strong positive or negative opinion of the subject and those who have the time and resources available to write reviews. In addition, some users have a personal stake in the outcome. These users include social influencers and others who are paid to provide reviews. The organizations selling products organize "ratings drives" to build interest. Some negative reviews are fake and written by malicious actors.

The GreatSchools site includes not only the reviews provided by users but also test scores, graduation rates, college readiness, achievement gaps, and student demographics for the schools as a whole. In addition, toward the bottom of each school's page is the average 5-star rating given by users, a selection of text-based user reviews, and a section allowing users to add additional reviews. In this sense, information creates more information. People might go to the site to learn about a school and, in the process, be prompted to input information about other schools with which they might be familiar. Unlike other school user review platforms, such as Niche, GreatSchools does not offer rewards for those who post reviews.

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. The introduction of school choice policies has increased traffic to the GreatSchools site. The content of reviews is correlated with the racial/ethnic and socioeconomic composition of schools, even more so than measures of student academic growth. The content of reviews also predicts future changes in schools' demographic composition and test scores. However, the reasons behind these patterns are unclear.

In our analysis, we try to understand the typical content of the reviews and differences across reviews, especially by user and school types. Three potential reasons why we may find differences in the topics included in reviews and in the ratings associated with a given topic are:

- *Different tastes.* Users might vary in the elements they care about in education and whether they view these elements as positive or negative. For example, parents and teachers might prefer strict discipline while students prefer looser discipline. In this case, we could expect all kinds of users to write about discipline, but the relationship between the star ratings and discipline measures would go in opposite directions.

- *Different emphasis.* Even if tastes are in the same direction, they might vary in emphasis. For example, students, parents, and teachers might all have the same tastes for discipline, but students might see this as a less important issue compared with, say, the availability of extracurricular activities. Therefore, students may write less often about discipline than other groups.

- *Different expectations*. Different users might have different expectations for what schools should offer and what constitutes "good" school performance on any given dimension. Even if all groups viewed discipline the same way, students might be more easily impressed than parents by what schools do.

Our analysis provides evidence that all of these reasons 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.

In what follows, we discuss the results from our research using all the GreatSchools reviews from 2009-2019. Schools with at least one user review have an average of only 8 reviews in total, including all types of users (parents, students, educators, and others) and all years.

In the following sections, we attempt to understand the content of the reviews using natural language processing (NLP), an artificial intelligence (AI) technique based on the same technology as ChatGPT. AI-NLP allows us to code vast amounts of data quickly and automatically. We also read samples of reviews and analyzed them using more common qualitative methods. See additional details on the data and methods in the associated technical report and at the end of this brief.

User Reviews Include a Variety of Topics but are Frequently Vague

Based on prior research and the initial coding of a sample of reviews, we identified 7 main topics that reviews could cover: *Instruction and Learning*, *Overall Quality*, *Physical Environment, Resources, School Culture, School-level Features, and School Staff*. Each word was additionally sorted into one of 21 sub-categories (See Table 1). We then identified the most common keywords and placed them in these categories by reading the reviews and using the results from AI-NLP.

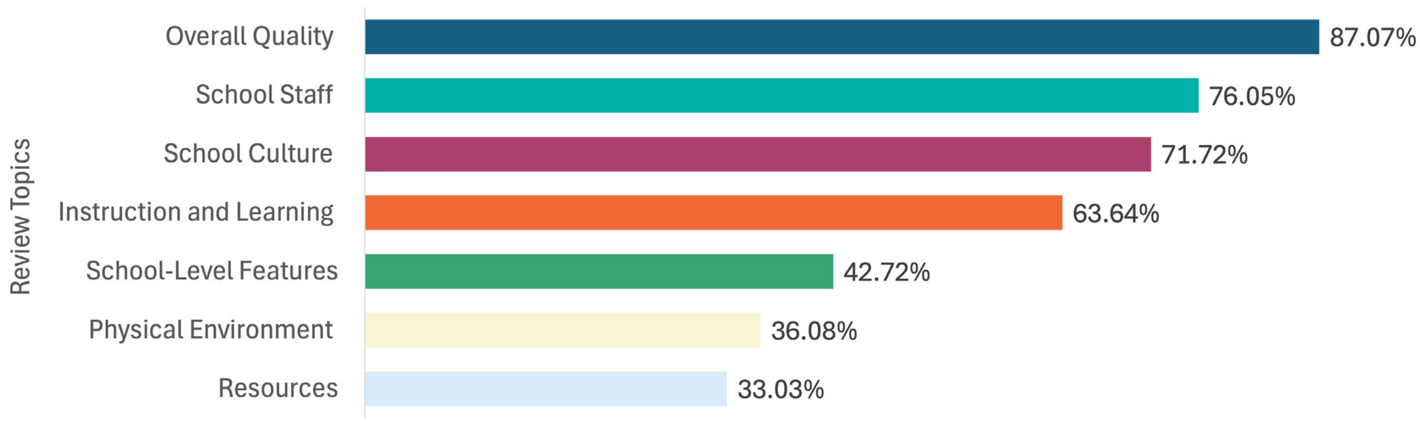
Table 1: Topics and Sub-topics

Topic Definitions	Sub-Topics
School Staff Words that describe teachers and other adults working in the school.	*Teacher Quality* and *Other School Staff*
School-Level Features Words related to the focus of the school or type of school.	*Religious, School Finances,* and *School Type*
Instruction and Learning Words about the ways students are taught and what they are taught.	*Instruction, Curriculum/Curricular Materials,* and *Learning Experience*
Physical Environment Words describing the school building and location.	*Location, Building Quality*, and *Facilities*
Resources Descriptions of what kinds of activities and courses are available in the school.	*Extracurriculars/Electives* and *Offerings*
School Culture Words that describe feelings about the school environment and relationships.	*Student Discipline, School Safety, Interpersonal Relationships,* and *School Environment*
Overall Quality Words (mostly adjectives) related to positive or negative judgment of the school.	*Evaluation, Postsecondary/Graduation, Preparation*, and *Quality Indicator*

Topics and Sub-topics in GreatSchools Data Analysis

The most discussed topics, shown in **Figure 1**, were *Overall Quality*, which includes words (mostly adjectives) related to the positive or negative judgment of the school, and *School Staff*, which includes words that describe teachers and other adults in the school. The next most common categories were *School Culture,* and *Instruction and Learning*. *School-level Features*, *Physical Environment*, and *Resources* were discussed least often.

Figure 1: *Overall Quality* and *School Staff* are the most frequently mentioned topics in reviews



Notes: Figure 1 shows the percentage of reviews mentioning each topic according to the AI-NLP analysis (2009-2019). Reviews that mention more than one topic are counted multiple times.

Topics in User Reviews Vary by User and School Types

Parents write the overwhelming majority of reviews (75%), followed by students (8%), teachers (3%), and principals (0.2%). Reviews are for TPS, charter and private schools in roughly equal proportion to the number of students nationally who attend schools in these categories. We examined how the reviews vary on these and other dimensions, especially school level (elementary/middle and high school) and student demographics.

Figure 2: Parents Contribute the Most to School Reviews and Principals Contribute the Least

Notes: Figure 2 shows the percentage of reviews written by each group. This does not reflect cases where the user type is missing.

Since we have multiple dimensions of subgroups, all of which are correlated, we estimated the relationship between each dimension and topic mentioned, controlling for all the other dimensions. For example, we study how the presence of each topic is related to each user type, controlling for the type of school, school level, and demographics. Below, we focus just on our main areas of interest–user and school types–and the more interesting findings associated with these dimensions:

- *Results by User Types*. Principals write more than parents about most topics. However, they write very little about *School Staff*, which can include both the staff they manage and themselves. Students write less than parents about almost everything, and they write even less about *School Culture* and *Physical Environment*. Teachers write less than parents about *Resources*, but more than parents on *School Culture* and *Instruction and Learning*. In the analyses of sub-topics within *Resources*, we find that parents wrote the least about extracurricular activities and the most (other than principals) about course offerings.

- *Results by School Sector.* The results also differed across traditional public, charter, and private schools:

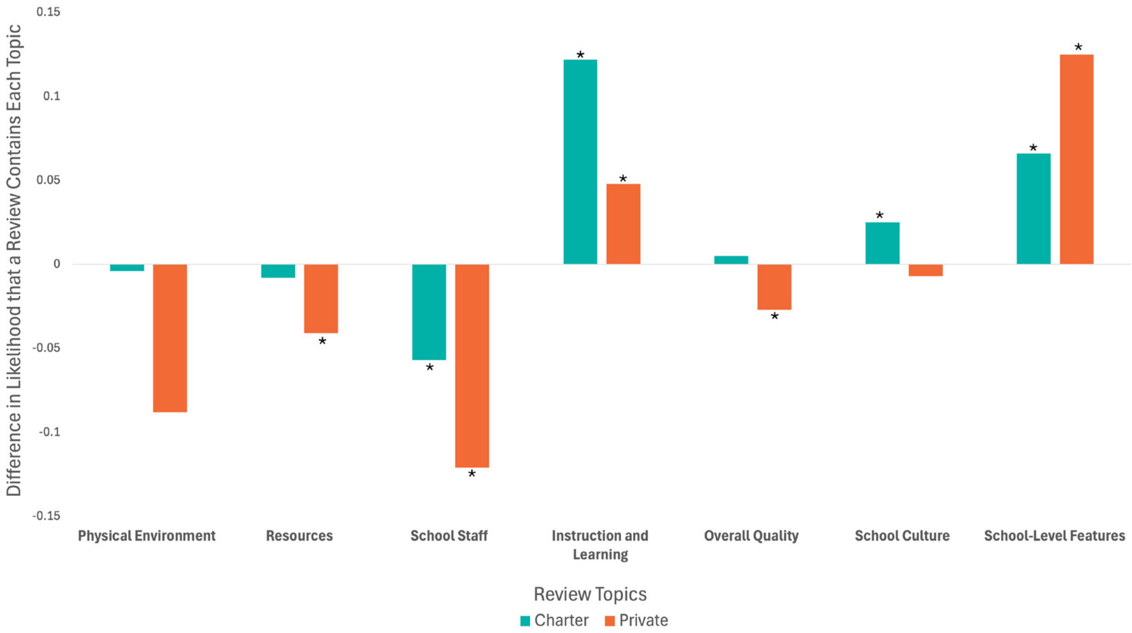
  ○ Charter reviews focus especially on *Instruction and Learning* and *School-level Features*, which may reflect that they often have specific themes (e.g., the arts).

  ○ In the analyses of *School Culture* sub-topics, 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. The sub-topic analyses suggest that these differences are driven by more of a focus on the location and facilities in TPS reviews than reviews of private schools.

  ○ The analysis of sub-topics within *Overall Quality* reveals that reviews of charter schools discuss *Evaluation* more than reviews of TPS and reviews of both private and charter schools discuss how these schools prepare students for the future more than for TPS.

Figure 3: Charter School Reviews Focus More on Instruction, While Private Schools Mention Physical Environment Less, Compared to Traditional Public Schools



Notes: Figure 3 displays results from regression analyses predicting the likelihood a review contains each topic, comparing TPS (the reference), charter, and private schools * indicates the difference is significant (p<.05).

The qualitative analysis also reinforced and deepened our understanding of differences by user and school type. 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 were also more likely than students to compare the current school to other schools, perhaps reflecting their roles as the predominant school "shoppers" and choosers. Parents are also more familiar with other schools because of adult neighbors and work colleagues, while students' social networks are mostly limited to school friends.

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 survey based research has found that charter school students report their teachers care less about them. Reviews of TPS 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 TPS.

What Users Value Varies by User and School Type

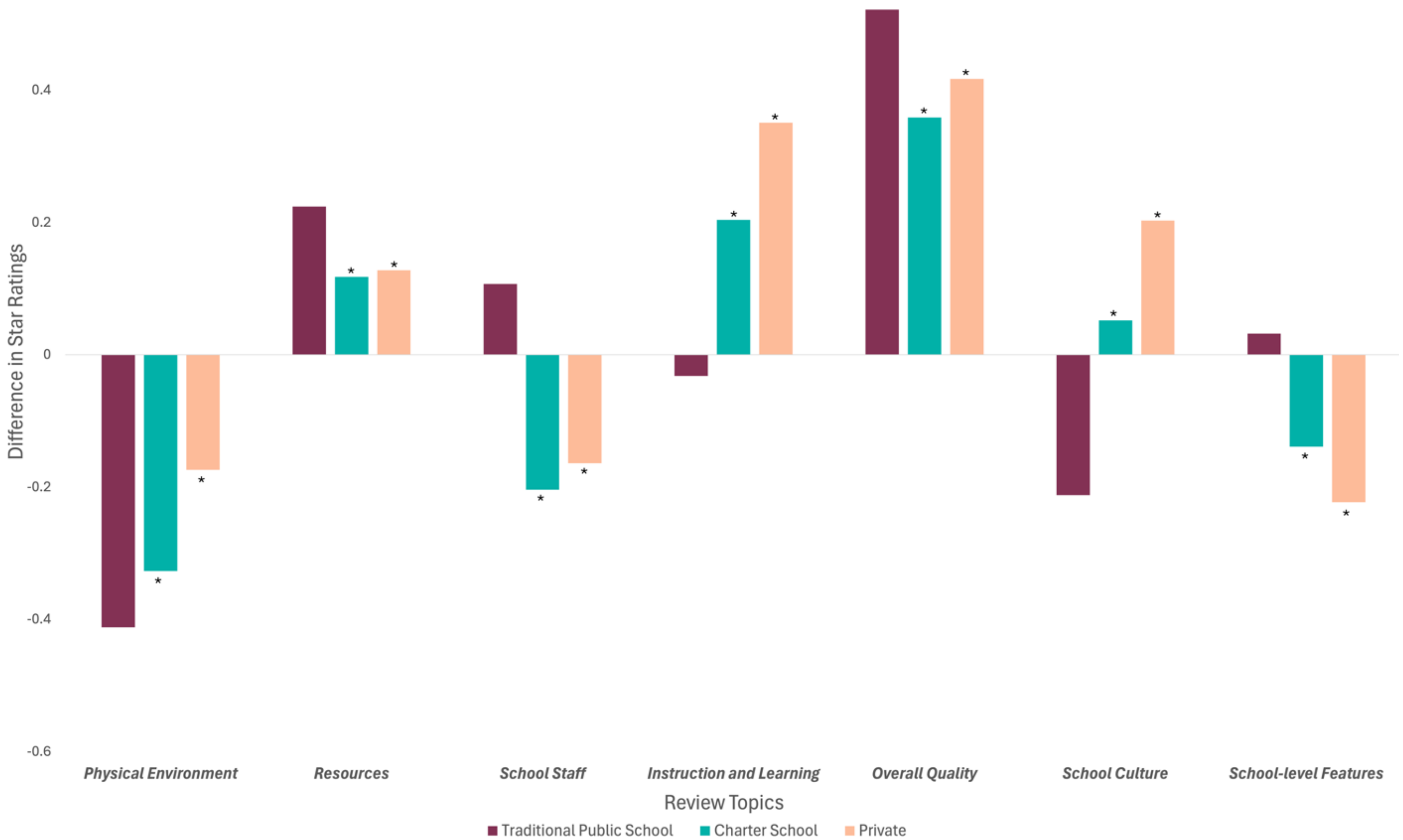
If the goal is to understand what users think is important about schools, then knowing the topics they write about is a good place to start. We would expect people to write about what matters most to them. However, we can go further by comparing the topics users write about with how they rate schools. We therefore used regression analysis to compare the star rating given by reviewers to the various topics discussed in the reviews. 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. Principals seem to recognize that giving themselves a star rating would be transparently self-serving.

- *Results by User Type:* Some of the results are consistent across parents, students, and teachers. [1] 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*.

- *Results by School Sector:* Across traditional public, charter, and private schools, we see starker differences. The relationship between star ratings and *School Staff* mentions is 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). The positive relationship between *Instruction and Learning* and rating is almost twice as large in private schools than charter schools.

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

Figure 4: Positive Ratings of TPSAre Associated with School Staff and School-level Features, Unlike in Private Schools



Notes: Figure 4 shows results from regression analyses of the relationship between star ratings and indicators for whether the main topics were discussed, done separately for each school type. * indicates there is a significant difference from TPS (p<.05)

Conclusion: The Limitations of School User Reviews

In this section, we summarize all of the limitations of school user reviews, including those recognized in prior research and those we have identified here. This is followed by a description of the potential ways that these issues might be addressed through improved platform design.

1. *Reviewers are not representative of all school stakeholders.*
   - Reviews tend to be written by individuals who either have a strong positive or strong negative opinion of the subject.
   - Only a tiny fraction of users post reviews at present, which reinforces the possibility that the reviews are not representative of the population. For example, 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%.
   - In other kinds of user platforms, some users have a personal stake in the outcome. People organize "ratings drives" to build interest.
   - Some negative reviews are fake and written by malicious actors. It is unclear whether these problems arise with school rating platforms.

2. *The vast majority of reviews are vague.* They mostly contain adjectives and do not describe the qualities, activities, or orientation of schools.

3. *Even where reviewers provide greater depth, they are difficult to interpret.*
    - Education is more complex than other goods and services. No reviewer could write about everything that users typically value about schools, or take the time to write about them clearly.
    - User review data is unstructured so it is nearly impossible for reviewers to describe schools in ways that reviewers can compare across reviews within schools or between schools.
    - Users are rarely clear in their reviews about their educational values and tastes, which makes it difficult to interpret even otherwise sophisticated reviews. We see measurable variation in the relationship between topics and star ratings, even between broad user types. The issue is likely much worse when we consider the additional variation within those groups.

4. *The above problems are not necessarily solved by increasing the number of reviews.* The share of reviewers is unlikely to ever reach a level where 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.

Implications for Improving User Platforms

There are three general, potential ways that user review platforms can solve these problems:

Option #1: Structured data collection (surveys)

The platforms could build in more structured data collection (e.g., surveys with close-ended questions) that would provide information that is more specific, standardized, and comparable across the schools from which families may be choosing. GreatSchools has implemented some of these changes since our analysis began. However, this approach may further reduce the share of stakeholders that engage with the sites, which is already a major limitation of the platforms.

Option #2: AI-NLP-generated summaries of all reviews, by school

As the number of text reviews increases, the user platforms could use AI-NLP to automatically summarize the information from all the reviews, not only in the aggregate, as we have done here, but for each school individually. This option, and the first one, both sacrifice some nuance that comes with open-ended text and qualitative analysis, but in a way that substantially reduces the cognitive complexity of the mass of unstructured data available. Our results suggest that AI-NLP is reasonably effective in capturing the topics that reviews cover and this is a good sign for the quality of the summaries.

These reviews could also be created by user type, so that parents could look specifically at what other parents think, although this comes at some cost. A user who is a parent might find what other parents say most relevant, but part of the value of such platforms is to provide information that users had not thought about, and this information is more likely to come from different types of users.

Option #3: Adaptive AI-NLP-generated summaries

Given what we have found in this study, a better version of the AI-NLP text summary approach would allow people to search for summaries that: (a) come from the same user types; and (b) focus on the specific topics that are important to them. 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. This would improve the likelihood that the information is actually useful. Amazon.com is already using a similar approach for product reviews. The platforms could also combine the methods and use data from targeted (and perhaps compensated) structured surveys to improve the AI-NLP algorithms that summarize the unstructured reviews.

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.

Implications for School Choice

The GreatSchools site is intended to improve school choice in a context where more and more options are available with the expansion of charter schools and vouchers. Given REACH's mission, we therefore reconsidered the above findings to see what implications they might have for school choice policies generally.

- We discussed the three main reasons that make reviews vary by user--differing tastes, emphases, and expectations. This has implications for the differences across school types. 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, reinforces prior evidence that "school choice" is not just about families choosing schools but about 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 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.

How We Did We Carry Out This Analysis?

Our dataset consists of all of GreatSchools' text review data from 2009-2019. We used two methods to analyze the data: a form of artificial intelligence (AI) called natural language processing (NLP) and qualitative analysis of small samples of text data. We briefly describe each below and refer readers to the technical report for additional details.

AI-NLP Methods

In all of our analyses using AI-NLP, we follow certain standard steps in preparing the text data for analysis.

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. Using a list of common place and person names we removed proper names.

Step 2: Identify Keywords. We then identified a set of approximately 1,000 frequent keywords and used them to identify which topics were discussed most often. When we compared the relative frequency of words by user types, school types, etc., we also built a set of log-likelihood keywords using a log-likelihood ratio test that measures relative frequency. 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 represent the range of subjects and concerns reviewers discussed or expressed about schools. We based the initial set of codes on what we know matters to people about their schools, based on prior research and existing parent surveys. 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 sub-topic codes (e.g., *Location* as a sub-topic of *Physical Environment*) to categorize the subjects or concerns that remained to be coded.

This is where AI-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. This method involves creating a prediction model that given a context, i.e., the words around a target word, predicts a target word. The language model allows us to identify the words that are most similar to each keyword.

Step 4: Annotate Reviews with Codes. After the keywords were coded into mutually exclusive topics, we then coded each review as containing a particular topic if the review contained one or more of the keywords associated with a given topic. Our coding of reviews does not distinguish the number of times a topic is discussed in a particular review, only which set of 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 sub-topics appear in the reviews (based on the above process).

To be clear, 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/sub-topics, none of the topics/sub-topics, and everything in between.

Qualitative Methods

While we are primarily interested in the potential application of AI-NLP, there are of course more standard methods we can use. First, to validate the coding done by the AI-NLP process described above we had members of the research team read and manually code reviews to check for validity and reliability. Since reading and coding reviews is time-consuming, we used random samples of 500-575 reviews that were of a reasonable length (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. We also used this qualitative analysis 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 AI-NLP-based analysis.

Comparing the AI-NLP and Qualitative Analyses

We checked whether the AI-NLP results were valid by having members of the research team read a small random sample of 500 reviews. The simple correlation between the AI-NLP and qualitative coding ranges from +0.24 to +0.75 across topics. Overall, The AI-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 and Learning*, 89% were also coded for *Instruction and Learning* under the AI-NLP. However, 56% of the reviews coded as not mentioning *Instruction and Learning* by the qualitative coding did have this code in the AI-NLP.

This disjoint between the AI-NLP and qualitative analyses is not ideal. First, it means the frequencies of topics reported in **Figure 1** are inflated. Second, our analyses for the subsequent research questions might be distorted.

How is this Research Related to Other REACH Research?

This is our second study of the GreatSchools platform. In our first study, *Online Reviews Are Leading Indicators of Changes in K-12 School Attributes*, we found some evidence that the language in text reviews is predictive of changes in school socioeconomic composition.

With the expansion of school choice policies, many different methods have been tried to improve information for families. These include "light touch" texting information to parents and providing them with school choice counselors, similar to college counselors, who help families understand their options and find the best fit. REACH has studies on both of these topics in process.

We have completed one study of school choice counseling, specifically, the services provided by the non-profit organization, Ed Navigator, *What Happens When Families Whose Schools Close Receive EdNavigator Support and OneApp Priority?*

About

The National Center for Research on Education Access and Choice (REACH), is the National Center for Research on Education Access and Choice. Our goal is to provide objective, rigorous, and applicable research that informs and improves school choice policy design and implementation to increase opportunities and outcomes for disadvantaged students. The research reported here was exclusively funded 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 or the U.S. Department of Education.

About the Authors

Douglas N. Harris is a Professor and Chair of the Department of Economics at Tulane University, and the Director of REACH.

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.

Nicholas Mattei is an Associate Professor of Computer Science at Tulane University.

Olivia Carr is a freelance educational consultant.