The Value of Student Feedback in Open Forums:  
Natural Language Analysis of Student Complaints

Authors

[Introduction 2](#_Toc62607797)

[Questions Driving Study 4](#_Toc62607798)

[Conceptual Framework 5](#_Toc62607799)

[Definitions of Good and Bad Teachers 5](#_Toc62607800)

[***Teacher-Student Relationship*** 6](#_Toc62607801)

[***Lack of Expertise in Teaching*** 6](#_Toc62607802)

[***Personal Characteristics*** 7](#_Toc62607803)

[How many teachers are rated as “bad?” 8](#_Toc62607804)

[Reliability of Quantitative and Observational Evaluations 9](#_Toc62607805)

[The Importance of Innovation in Teacher Evaluation 12](#_Toc62607806)

[Numeric vs. Descriptive Student Feedback 15](#_Toc62607807)

[Data and Methodology 16](#_Toc62607808)

[Preparing the Data 17](#_Toc62607809)

[Natural Language Processing 18](#_Toc62607810)

[Measuring Topic Coherence 19](#_Toc62607811)

[Interpretation of Results 20](#_Toc62607812)

[Limitations of the Study 20](#_Toc62607813)

[Results 23](#_Toc62607814)

[Topic 7 - Ineffective Grading (8.6% of reviews) 25](#_Toc62607815)

[Topic 11 – Why even ask? (6.5%) 26](#_Toc62607816)

[Topic 2 - Did Not Learn (10.5%) 26](#_Toc62607817)

[Topic 9 – Lack of Expertise (10.5%) 27](#_Toc62607818)

[Topic 6 – Wasting Time (9%) 27](#_Toc62607819)

[Topic 5 – Ruining the School (9.5%) 28](#_Toc62607820)

[Topic 8 – The Worst (8.3%) 28](#_Toc62607821)

[Topic 3 – Can’t Teach (10.4%) 28](#_Toc62607822)

[Topic 1 – Boring (11.6%) 29](#_Toc62607823)

[Topic 4 – Mean, Short Tempered (10%) 29](#_Toc62607824)

[Topic 10 – In Need of Change (7.7%) 29](#_Toc62607825)

[Discussion 1](#_Toc62607826)

[Evaluation of the Model 1](#_Toc62607827)

[Small “p” Policies Based on Student Descriptions 2](#_Toc62607828)

[Contribution to the Literature 3](#_Toc62607829)

[Conclusions 6](#_Toc62607830)

[References 8](#_Toc62607831)

# Introduction

Teachers have always been evaluated. Legend has it that judgments about the quality and performance of the teachers of Athens were posted in the Agora for public view. Socrates was judged to be a bad teacher. Apparently, he spent too much time asking his students to think. A walk-through evaluation by his supervisor determined that “sometimes Socrates’s students meander through endless dialogues examining challenging questions that do not have one right answer.” This led Jacobs (2012) to think that Socrates might be replaced, or perhaps be required to take an intensive summer professional development program in Sparta. This fable, sarcastically retold, illustrates both the importance of feedback from students and the flaws of high stakes evaluations in order to improve teaching.

The public nature of teaching invites school administrators, peers, students, and parents to evaluate the quality of teachers, although by means of different metrics. These evaluations can be formal, such as those used by administration, while others can be deeply personal, communal, and not shared, such as those evaluations done, with feedback daily, by students. Certainly, informal evaluations focus on different criteria, record qualities or accomplishments which are more difficult to measure and do not carry as much weight. For example, interpersonal relationships are high in importance to students (Uttio, 2012), as are concepts like fairness and being treated respectfully. Formal evaluations often put far less of a premium on these characteristics. And, since these attributes of teachers are often subjective interpretations proffered by adolescents and young adults, they may be seen as unreliable and are often dismissed. Despite these challenges, student observations of their teachers have demonstrated correlations to other measures of teacher effectiveness (Chaplin et al., 2014) and can be reasonably stable for a given teacher year-to-year (Polikof, 2015).

Millions of evaluations happen in schools across the country by those students closest to the performances of their teachers. Several decades of studies do indicate that students have insight into how teachers perform and how it impacts them (see, for example, Rodin & Rodin, 1972; Check 1986, 1999; Uitto, 2012; Hosgorur, 2015; Raufelder et al., 2016; Chang-Kredl and Colianno, 2017). Yet, research on what they uncover remains scant (Polikof, 2014), and what does exist is somewhat small in scope, with surveys or interviews done on dozens or a few hundred respondents. Despite the inherent variation in these students’ experiences and judgments, there is value in understanding and using feedback from such evaluations as complementary to the more formal criteria used for assessing practicing teachers. They may paint a more complex--and more meaningful--portrait of a given teacher. These evaluations of teachers by students have utility for development of small policy decisions at the local level. Big policy, “P” (Berliner, 2019), may be concerned with national testing, school district funding, and integration by race and class. Small policy, “p,” may be related to how often school nurses check on students' health, when to call parents if a child is absent for 3 or more days, choice of substitute teachers, and so forth. This empirical study explores the possibility of using digital forums to inform the development of small “p” policies--school or district level policies--by using a large number of written reviews--the type of data a system like this may produce--to arrive at consistent descriptions of teachers whose teaching behaviors could be considered “bad.” These policies may include professional development for teachers who students consistently and repeatedly are reported to have the same issues, monitoring that signals the need for more serious review and, even, counseling for anger management and other ways they could improve performance.

We seek to increase the scale--and perhaps the volume--of student voices by asking: How do students describe and define teachers they judge as being “bad?” How do student’s descriptions of these teachers inform us about the practices, attitudes and beliefs of those teachers? And, how can descriptions of characteristics of bad teaching inform good policy? We began by collecting more than 4.8 million publicly available student comments and ratings on the RateMyTeacher.com website and used natural language processing to identify common descriptors of teacher performances. We narrowed our focus to the worst rated teachers on the website, specifically to uncover those characteristics students deemed to fit in this category. Our analysis should be a useful illustration of the methods schools can use to gain insight from digital forums for development of small “p”, school-based policies that can lead to school improvement.

# Questions Driving Study

1. How do students describe teachers they rate poorly as recorded on reviews on the RateMyTeacher.com website?
2. What common characteristics arise from these descriptions in aggregate?
3. How can these aggregate characteristics drive (small “p”) policy?

# Conceptual Framework

## Definitions of Good and Bad Teachers

Looking at issues of teacher quality, there is some consensus about the desired qualifications of a teacher. Dorham (1987, p. 3), using qualities of a good teacher as described by 6th graders, says “Three distinct themes regarding the efficacy of teachers emerged from students' comments: (1) instruction; (2) personality; and (3) classroom management.” Good instruction means that teachers presented material in ways in which students could understand and they did it with patience and creativity. Personality was described mainly as being “nice,” “not yelling” and not looking as if they were bored. Classroom management again emphasizes, no yelling and also the responsibility of the teacher to intervene when necessary. Goodwin and Oyler (2008), on the other hand, believe that the most important quality any teacher has to have is content knowledge. These researchers go on to list other required characteristics: language proficiency and fitness to teach. Hattie (2015) writes about good teachers making the connection between student achievement and their teaching practices, while paying attention to what students are actually doing in their classrooms. Peneul & Shepard (2016) describe the good teacher as instantiating the Deweyin idea of “seeing on the horizon the full mastery of disciplinary knowledge and practices and translating that into intermediate understandings and ways of participating connected to the experience of the learner” (p. 787).

Less has been written about the qualities of perceived or genuinely bad teachers. Two specific studies about bad teaching, however, are closely related to our study and shed a light on our findings. Raufelder et al (2016) had a similar goal to ours, but was based on a much smaller sample. These researchers questioned 86 German junior high school students about bad and good teachers. They then organized, as we do, the responses into themes and sub-themes. What we found overlaps with their research. Apparently students’ ratings of perceived bad teachers, as is true of good teachers, have common characteristics. From their interview data three prominent bad teacher themes were identified:

### ***Teacher-Student Relationship***

The first was about the quality of the teacher student relationship. Three sub-themes were recognized. First was *relational aggression*. Teacher behaviors of this type included teacher yelling, and the teacher being insulting (e. g. calling students stupid). Vilification of the students was also categorized as relational aggression (e. g. showing little or no respect for what they had accomplished). Sub-theme two was labeled *injustice*. Bad teachers were seen as playing favorites in the classroom, or who had opaque and changing evaluative criteria. Sub-theme three was called *antipathy*. Antipathy by students toward their teachers developed out of a general dislike of the teacher. This theme developed because students found their teachers often to be incomprehensible in communicating subject matter content, or because their teachers really did not know the content.

### ***Lack of Expertise in Teaching***

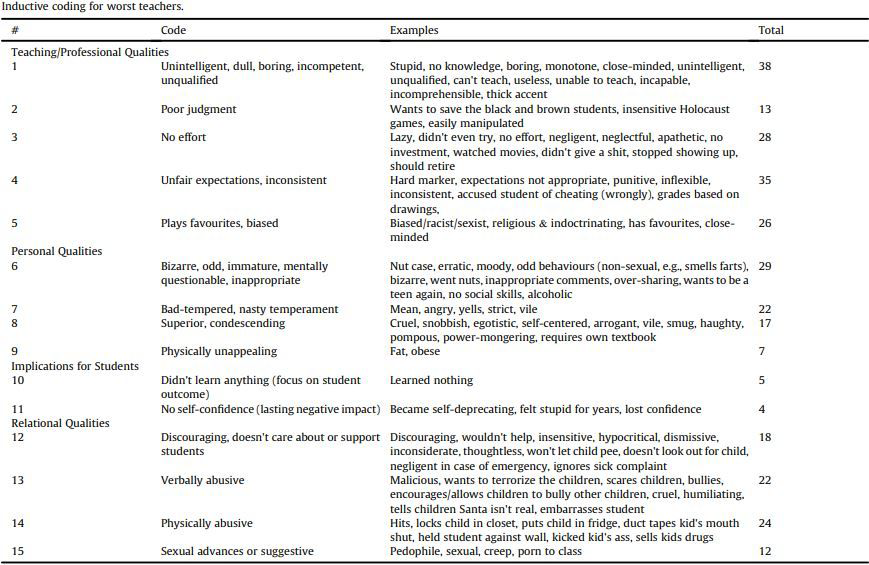
Theme two also arose out of the coding of the interviews, and was based on students’ concerns about their teachers’ lack of expertise in teaching. One subtheme was the perception of students that their teachers were *disinterested or indifferent* to the materials being presented, which was disappointing to the students. A second subtheme was labeled *incomprehensible teaching*. It is quite easy to understand why students would be harsh judges of this perceived teacher characteristic. A third subtheme was labeled *teacher-centered instruction*. The students resented copying material, or seatwork, where in both cases, the teacher has it “easy” and the students do the instructional work.

### ***Personal Characteristics***

Theme three consisted of comments about the *personal characteristics* of the teacher, for example, boring, repetitive, or disinterested. It included, as well, examples of teachers’ lack of assertiveness, and disorganized or incomplete presentations of the material. Student examples of this negative trait included allowing whispering or talking among students that should have been stopped, newspaper reading in class, disrespect to the teacher that was uncriticized, students throwing things in class, etc. In this study the students’ comments about their dissatisfaction with teachers were weighted more heavily by their teachers’ failings in terms of their interpersonal skills than by their academic skills.

In each study it is interesting to note that many students judge their teachers to be “bad” on the basis of their inability to teach them as much as they want to learn. The analysis in our study suggests that students used different language when they genuinely attempted to distinguish between interpersonal and academic skills. We aim to make the case that this value students place in their own learning lends itself to the usefulness of a digital forum. For example, a digital forum could be used to recognize whether the students in a school system consistently feel as though the course materials for specific subjects (such as computer science, history, or sexual education) are outdated or irrelevant. Our analysis suggests that grievances of this sort could be distinguished from the interpersonal failings of teachers that may confound and exacerbate problems. As a result, a potential policy could require teachers lacking in knowledge to demonstrate they have gained that knowledge.

In another study, Chang-Kredl and Cloannino (2017) were interested in examining the image of teachers in the public sphere—movies, television, and on the web. Analyzed in this study were the descriptive comments made about the best and worst teachers encountered by subscribers to Reddit, a popular website for social news aggregation, web content ratings, and discussions. Characteristics of the respondents are unknown, but consisted of individuals sufficiently motivated to post their comments on Reddit, often well after their personal encounters with “bad” teachers. Their analyses revealed many characteristics of bad teachers that mirror some of what we found in our data. Table 1 provides these data.



## How many teachers are rated as “bad?”

These descriptions, comments, and complaints make clear that for some, or even many students in a class, teachers who are not reaching students do exist, and in number, though certainly not in the numbers often bandied about in our media. So, a relevant question for this study is “How many teachers are we talking about?” In fact, the numbers of teachers who are seen as “bad” described in Table 1, and in our own work, actually appear to be rather low. Berliner (2014) estimates their numbers at about 3%. In the well-respected Hechinger report, Butrymowicz (2014) says that states such as Tennessee, Michigan, Georgia, Florida, and Pennsylvania, particularly in Pittsburgh, all provided estimates of “bad” teachers in this same low range. Danielson (2016), who has visited and coded many hundreds of classrooms, estimates the “bad teacher” number to be around 6%. That seems to be the high end of estimates from those who are experienced classroom analysts. In our own study we found that, out of more than 4.8 million reviews, using a 100 point scale, 55% gave a maximum rating of 100 (the best score), 75% gave a rating of 80, and 89% gave a rating greater than 50, reflecting compatibility with the estimates of the percent of bad teachers by those analysts cited above.

## Reliability of Quantitative and Observational Evaluations

While informal and unscientific evaluations are common, it is still difficult to find contemporary examples of teacher evaluation techniques that meet the standards for reliability and validity proffered by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education (2014). For example, test based accountability systems, such as value-added models (VAMs) generally are unreliable year to year (Amrein-Beardsley, 2014), from subject matter to subject matter (Amrein-Beardsley & Collins, 2012), and even unreliable from class to class in the same subject and in the same school year (Newton, Darling-Hammond, Haertel, & Thomas, 2010; See also Konstantopoulos, 2014; and Author, 2014). No lasting “P” or “p” educational policies have been derived from these studies.

Consistent unreliability in the measures used to assess teachers, negatively affects validity. Test-based accountability systems, especially VAMs-based accountability, ought to be avoided (Pivovarova, Amerein-Beardsley, and Broatch, 2016). Nevertheless, politicians and policy makers seem partial to test-based models of teacher evaluation, even when researchers point out that the amount of variance in student test scores that is attributable to their teachers is negligible (American Statistical Association, 2014). Demographic factors (family income, mothers’ level of education, abilities of the cohort that one goes to school with, etc.) are almost always the best predictors of performance on standardized tests of achievement: not teachers, nor schools (cf. Haertle, 2013).

Observational instruments to evaluate teachers have reliability problems as well, similarly limiting their validity. They almost always require more observers and more observation time than can be afforded by principals, peers, or the school systems that seek such data. Thus, their reliability is often questionable. Among their other drawbacks are the fact that observational instruments usually cover only a short period of teaching time, and cannot be trusted to be valid if consequential decisions about teachers are to be made on the basis of such limited observational data.

It should be highlighted that there are some observational instruments that are commonly used and found by many educators to be useful in providing feedback to teachers (e.g. Danielson, 2008; Pianta, LaParo, & Hamre, 2008), however, the results of test-based teacher accountability methods and observational methods of accountability are not substantially correlated. For example, in the multi-million dollar MET study, funded by the Bill and Melinda Gates Foundation (Kane, McCaffrey, Miller, & Staiger, 2013) four different observation instruments were correlated with the VAMs associated with math achievement test scores. Those correlations were .12, .18, .25, and .34., averaging about .22. With the VAMS derived from reading and language arts tests the observation instruments correlated .12, .11, and .09, averaging about .11 (Bill and Melinda Gates Foundation, 2012). A separate study using this data set found that the correlations between an observational measure of excellence in teaching, and two measures of excellence in teaching derived from VAMs, were trivial: .16 and .09, respectively (Grossman, Cohen, Ronfeldt, & Brown, 2014). Strunk, Weinstein, & Makkonen (2014) correlated observational data and VAMs for reading and math, over one year. They found correlations under .216. Similarly, Morgan, Hodge, Trepinski, and Anderson (2014), found correlations between observations of teachers and their pupil’s performance on tests that were roughly between .20 and .40, indicating, once again, that these two different measures of teacher competence have in common only between 4% and 16% of the variance observed. The latter investigators noted, additionally, that neither teacher performance in classrooms, nor teacher effectiveness as judged by test scores, were highly stable over multiple years of the study.

Since the variance in common between test-based accountability measures and observational measures is the square of the correlation coefficients just cited, these two methods of evaluating teachers are not measuring the same thing at all. They measure different constructs, or perhaps different aspects of what is sought. Each of these approaches to evaluation has problems: The test-based accountability systems do not look at teachers’ classroom behavior, and the observational systems do not assess learning outcomes. And neither has access to teachers’ thinking, which determines both their classroom behavior and, indirectly, their students’ likelihood of scoring well on tests.

## The Importance of Innovation in Teacher Evaluation

Although there is a public aspect to teaching and students' test scores provide important artifacts associated with classroom teaching, much of the most important part of the teachers’ job is cognitive. Thus, it is unobservable. Teachers make a number of decisions per day that cannot be easily captured from observational instruments or via student test scores. Borko, Livingston and Shavelson (1990) estimated that teachers make at least .7 consequential decisions per minute, 42 per hour, over 250 per day. Jackson (1990), writing at about the same time, believed that teachers engage in 200 to 300 consequential exchanges with students every hour (between 1,200-1,800 a day!). Most of these are unplanned and unpredictable and the thoughts that are behind them are typically unknowable. Choosing between the two most common and equally flawed evaluation systems (test-based and observational evaluations) is akin to being between Scylla and Charybdis (Author, 2018). Problem-free teacher evaluation systems do not exist.

But these are just the two most common ways to assess teachers. There are other methods, each with their own strengths and weaknesses. Scriven (1994), for example, has proposed that teachers be rated on the basis of their performance of the essential duties of a teacher. This “duties based” assessment has much to offer. Users can learn to use the system reliably in a short period of time, and its face validity is quite high. But duties-based evaluation systems are infrequently employed. It is thought that this form of evaluation is too removed from the heart of the teachers’ job, namely, interactive classroom teaching. Instead, a duties-based evaluation system focuses on other important aspects of the teaching job, such as showing up to class on time, giving students back their written papers or assessments with useful comments on them, communicating regularly with parents, and a host of other “duties” expected to be adequately fulfilled by teachers. This is an assessment system of important aspects of the teachers’ job—related to what happens in classrooms and on tests—but not directly assessing those factors. The correlation of duties-based evaluation systems with test-based or observational systems is unknown at this time, but it is likely to be low.

There is one other method occasionally used for evaluating teachers. We referred to it earlier: It is by means of student evaluations of their teachers. Such evaluations are most likely to be used at the college level, where raters are thought to be mature enough to engage in this activity. Students are less likely to be used as evaluators in the K-12 system because of their purported immaturity. We think that is because all four approaches to evaluating teachers (test-based, observation-based, duties-based, and student-based rating systems) only deal with a piece of the teachers’ job but cannot adequately describe the overall qualities of “teacher.” In a sense, this limitation provides a contemporary example of the parable of the blind men and the elephant in which we can only tell a part of the whole through any one (or even several) types of evaluation. This fourth method discussed, like the other three, has advocates for its use in the K-12 system.

For example, Scriven (1995), offers nine reasons to consider student evaluations in a positive light:

1. The positive and statistically significant correlation of student ratings with learning gains.
2. The unique position and qualifications of the students in rating their own increased knowledge and comprehension.
3. The special situation of the students in rating changed motivation (a) toward the subject taught; perhaps also (b) toward a career associated with that subject; and perhaps also (c) with respect to a changed general attitude toward further learning in the subject area, or more generally.
4. The singular ability of the students to be able to rate observable matters of fact relevant to competent teaching, such as the punctuality of the instructor and the legibility of writing on the board.
5. The peculiar circumstances of the students in identifying the regular presence of teaching style indicators. Is the teacher enthusiastic; does he or she ask many questions, encourage questions from students, etc.?
6. Relatedly, students are in a good position to judge—although it is not quite a matter of simple observation—such matters as whether tests covered all the material of the course.
7. Students as consumers are likely to be able to report quite reliably to their peers on such matters of interest to them as the cost of the texts, the extent to which attendance is taken and weighted, and whether a great deal of homework is required--considerations that have little or no known bearing on the quality of instruction.
8. Student ratings represent participation in a process often represented as "democratic decision-making."
9. Students may be the "best available alternative" for learning about what goes on in some classrooms.

Similarly, Ripley (2010), and Cuban (2012) both make persuasive arguments for using student ratings and evaluations of teachers. Others, however, especially those in higher education (Lawrence, 2018) argue that the information obtained from student evaluations is invalid, and does more harm than good. We looked closer at this issue and found that student evaluations of teachers (SETs), particularly simple numerical rating systems, do have serious problems that may render them invalid for their traditional purposes. For example, multiple studies reviewed by Reid (2010) confirm that an anticipated grade in a course influences SETs. The higher the anticipated grade, the higher the ratings. More important for this study, perhaps, is Reid’s empirical study at the college level that demonstrated racial biases. The majority white student body in the institutions he studied rated white faculty significantly higher than Asian and Black faculty. Such biases, no doubt, have influenced our data set. But these studies should inform small “p” policy makers not frighten them. With this awareness, biases could be sought out in the *linguistically* descriptive type of feedback used in our analysis and perhaps lead to helpful activities to improve racial, gender or other sensitivities in a school..

## Numeric vs. Descriptive Student Feedback

Stark and Freishtat (2014) studied the numerical rating system used for professors at UC Berkeley, and found, among other problems, unacceptable differences in response rates per faculty member. Thus, the confidence bands around any numerical rating were bound to be quite different, but such issues were usually not addressed. These authors also rightly questioned whether treating the same numerical value for an instructor (say an .average of “8”) means the same thing in courses as different as physical education and physics, or in electives and required courses. They too discuss biases in SETs and conclude “We will never be able to measure teacher effectiveness reliably and routinely….But [among other things] we can look at student comments (p. 4).” This is exactly what we do. Our study is an analysis of the student comments that accompany low numerical ratings of teachers. Our goal was to organize and analyze these comments submitted, mostly, by students to describe teachers they believed to be bad.

Beside Reid (2010), cited above, Boring, Ottoboni, and Stark ( 2016) have also demonstrated clearly, with both European and American data, that SETs are consistently biased. In this case the bias found was against female instructors. Ratings on such attributes as Caring, Enthusiasm, Feedback, and the like all showed this gender bias. These identical attributes of teachers were frequently mentioned in the data set we analyzed, but they took the form of comments about a “Lack of Caring,” a “Lack of Enthusiasm,” or a “Paucity of Feedback.” Such gender bias in the SETs we analyzed should not have affected our work at organizing and attempting to make sense of the descriptors used when students judge teachers to be “bad.” Freishtat (2016) sums up our defense of SETs quite well. He notes that while SETs have unresolvable problems when used to determine excellence, or for the design of policies relating to promotion, tenure, pay, and other important aspects of a teachers’ career, they really are still reasonable measures of student satisfaction with their experience in a course. SETs, he says, do “give us insight into the student experience (p. 12).” In what follows, we have organized the text of students who regarded their teachers as “bad.

Because we are not using the information from SETs to influence decisions about teachers that demand greater trust in the data, we join with Scriven, Ripley, and others in our defense of the student data we do use. We use *commentaries* made about a particular subset of teachers that students (and some parents) had judged with their numeric ratings to be “bad” teachers. Such comments, *all over one-hundred characters in length,* provide useful information about teachers and teaching, independent of the validity problems that plague numerical ratings used for consequential decisions about instructors. Our goal is to help school administrators, in particular, interpret the meanings of the negative language that is used in SETs collected through digital forums which can lead to little “p” school or district policies.

# Data and Methodology

The data we used for this study were reviews of teachers submitted through RateMyTeacher.com, a website where students and/or parents can submit a review of a teacher along with a rating from 0 to 5 stars (½ stars are allowed). In 2018, the website changed ownership, and the data collected previously was removed from the site as was a rater’s ability to submit written reviews. We scraped the data in early 2018, when all of the reviews submitted since 2001 were still visible on the site. The last review in our dataset was submitted on January 16th, 2018.

Reviews were attached to specific teachers and schools and were collected from 6 English speaking countries (United States, Canada, United Kingdom, Australia, New Zealand, and Ireland). We limited our data set to only those originating in the U.S. We collected 4,884,479 reviews from the U.S. Each review was accompanied by a rating from 0 to 5 stars. Some were one quarter or half filled, so we scaled the ratings to account for the partial scores, and used a 0-100 scale. A one star rating, therefore, would be given a value of 20; a half star rating was given a value of 10. Thus, a one and a half star rating was given a value of 30; etc. The data were heavily skewed, with almost half of all reviews being 5-star ratings. This distributional skew held even when we only considered teachers with at least 50 separate reviews (about 7% of the total dataset). This study focused on a filtered sample from the 359,387 reviews rated 0-35 in our dataset. We used natural language processing methods, described below, which benefit from large sample sizes.

## Preparing the Data

Our first step was processing the text data so it could be analyzed by statistical models. The “cleaning” process involved removing stop words (e.g. “a”, “is”), coded characters (e.g. “\n”, “\r”), and infrequent words such as personal names. We removed conjugations and pluralization using the word lemmatizer from Python's Natural Language Toolkit (Bird, Klein & Loper, 2009). We then used Gensim for Python to identify phrases by using n-grams to identify words that co-occurred often enough to warrant a unique meaning (Röder, Booth, & Hinneburg, 2015). For example, if “laid” appeared next to “back” enough times, every instance of both words appearing in that order would be replaced by “laid\_back” such that it made a new, unique word. We allowed phrases of up to four words (i.e. “as\_soon\_as\_possible”). Cleaning the data was an iterative process. We experimented with the parameters of the cleaning process and found one experimental parameter that dramatically improved our model. That was a lower bound on review length. We found that limiting each review to a minimum of 100 characters dramatically increased the probability that any given review in the sample contained non-trivial information. Of the 359,387 “worst” reviews, 211,224 met the 100-character threshold. We also experimented with 125, 150 and 175 character thresholds.

## Natural Language Processing

Since natural language processing became popular in the 1980’s a seminal criterion used to evaluate algorithms has been text classification through information retrieval (Lewis, 1992). Information retrieval refers to a language model’s ability to identifyand retrieve words, sentences, or paragraphs that are alike. When one does not have reliable labels that can be used to characterize documents (in this case, reviews), these are analyzed using unsupervised learning. This refers to statistical methods that cluster documents together based on how close they are to one another in the metric space created by a language model (Mikolov, Chen, Corrado & Dean, 2013). The words that most frequently appear in the cluster can be used to characterize it and these can be interpreted as topics (Papadimitriou, Raghaven, Tamaki & Vempala, 2000).

We used a latent dirichlet allocation (LDA) model to generate clusters to derive topics from. In LDA, clustering works by randomly sorting the documents into K groups and then iteratively moving them around until the members of each cluster are closest to each other and furthest from members of other clusters. Some algorithms come prepackaged to help determine that number (Teh, Jordan, Beal, & Blei, 2005) but instead, we chose to use the concept of coherence to estimate an optimal number of K-categories for subsequent analysis.

## Measuring Topic Coherence

Coherence metrics measured the spread (or concentration) and orthogonality (or mutual exclusivity) of topics. The logic behind using these metrics was that 1) clusters that were more spread-out would be less informative than clusters that were more dense and compact, and 2) clusters that overlapped significantly would be less informative (more redundant) than clusters that were mutually exclusive (Stevens, Kegelmeyer, Andrzejewski & Buttler, 2012; Mimno, Talley, Leenders, Wallach & McCullum, 2011).

In our analysis, we trained models for different numbers of topics (from 3 to 32 in steps of 1). We used Gensim to estimate coherence scores for each topic generated by each model. Average coherence scores were then used in selection of the models that best fit the dataset. Average coherence usually increases with the number of topics. Selecting the number at which the rate growth of average coherence plateaus usually yields the most interpretable results and a model with a higher number of topics can result in more specific subtopics, but if the same key-words appear in multiple topics then a model with a lower number of topics is likely preferable.

Several coherence measures exist, we chose the “Cv'' measure derived in Röder et al (2015), in which they compared this rating to others from the literature and found that it yielded the results that were most highly correlated with results generated by humans. When reviewing the topics in the best fitting models, we focused on topics that had higher coherence scores since these were most likely to represent consistent sentiments expressed by reviewers.[[1]](#footnote-1)

## Interpretation of Results

We visualized the results using the pyLDAviz module which illustrated the topic distributions by plotting them in a 2-dimensional space that illustrated both their share of the corpus and their overlap with one another (Sievert & Shirley, 2014). This provided us a more intuitive view of the relationships between the topics. We selected key words using relevance metrics described in Carson & Shirley (2014) and included in the pyLDAviz module. We set the relevance parameter to 0.5 such that the words returned were 1) those which were most frequent in the text and present in the topic or 2) those which were most distinctive; mostly appeared in the topic and nowhere else. We were also able to use our model to predict the probability that any given review belonged to a topic. This resulted in a dataset that assigned each review topic percentages that were interpreted as the presence of any given topic in that review. Using these, we drew the 18th, 19th, 20th, 98th, 99th, 100th, 198th, 199th, 200th... up to the 2,100th most representative reviews for each topic seeking obvious themes. We combined insight from these sample reviews with topic key words to create comprehensive descriptions of each topic.

# Limitations of the Study

The insights that can be drawn from this study are primarily limited by the data collection process. The reviews were voluntarily submitted by students or (much less frequently) by parents. We did not know whether the students or parents were incentivized in some way. For example, a teacher may have offered extra credit to students for submitting a review or a parent may have been trying to get a teacher removed from a school or district. We also did not know the degree of bias in each review. For example, we did not know if a student was lashing out against a teacher for a bad grade they may have deserved, or if a student was pressured by classmates into submitting a good review despite feeling differently. The common biases against instructors who are female or persons of color, discussed above, probably affected the ratings given by students. But as noted, this should not be a problem because no consequential decisions are being made about instructors here. Instead, the comments have been analyzed to promote understanding the language used for describing teachers rated “bad” in a digital forum.

Opinions about RateMyTeacher.com in articles and teacher forums were split and illustrate some of the controversy around online digital forums in educational contexts. Much of the scholarly literature around RateMyTeacher emphasizes the unreliability of commenters and the quality of the responses (see, for example, Burdick, 2009; Burdick & Sandlin, 2010; Angel, 2009). Online forums (such as those at https://community.tes.com/threads/ratemyteacher.328021/) featured more diverse opinions from teachers and students themselves. Some took offense at negative comments or were concerned about personal information being posted. Others appreciated it as an open forum for students and argued that site administrators had guidelines for removing inappropriate comments.

These sorts of uncertainties rendered our dataset useless to answer many of the types of questions one might address with student feedback collected in a classroom setting, for example. Fortunately, the size of our dataset let us scrutinize some of these concerns in the context of our guiding questions. Our dataset contained over 4 million reviews, covering over 27,000 schools. Of the 128,344 teachers reviewed in our final sample:

* 68.5% received only 1 bad review,
* 17.6% received 2 bad reviews
* 13.9% received 3 or more bad reviews.

To put this in context, the average number of bad and total reviews for teachers in our sample was 1.6 and 11.3, respectively. Put another way:

* 11.5% of teachers in our sample received 100% bad reviews.
* 10% received 50% bad reviews.
* 42.3% of the teachers in our sample received at least 20% bad reviews.

These distributions were reflected in the average ratings for teachers. The average rating for the entire dataset was 92.8, where the average ratings for teachers in our sample was 79.8. Such a high rating made it clear that many teachers receiving negative reviews must have received a much better rating and better reviews by other students at the same or a different time. Naturally, this appeared to be especially true of teachers rated poorly for having classroom favorites.

In all, the assumption that the entire dataset was completely biased seemed less plausible than the assumption that the dataset represented a diverse array of motives, contexts, and incentives for submitting reviews about teachers. Most teachers received mixed reviews, while a small minority received mostly bad reviews. Furthermore, the reviews for teachers were submitted across long periods of time (the average time between the first and last review for each teacher was 5 years). Thus, the topics discussed by reviewers in our sample likely represented common “descriptions” that could be found in classrooms across the country and over time, rather than representing the views of any one type of mythical reviewer. In fact, the results of the analysis increased our confidence that this was the case.

One results detailed the variety of ways in which students expressed the behaviors they observed in teachers, and gave, as well, their thoughts about the origins of those behaviors. Furthermore, one finding in this study was that filtering reviews by length (100 characters) was essential for identifying consistent topics among the reviews. This enforcement of minimum length increased the proportion of meaningful reviews that were analyzed and decreased the likelihood of including reviews that required no effort or a lack of sincere thought to write.

Ultimately, any conclusions drawn from this study can only add to the literature around the issues students are capable of communicating or observing about their teachers. It may also illustrate the ability to recognize these issues in an online digital forum. Some may also find this study as another piece of evidence that supports the value of student feedback. Nevertheless, this is not a formal evaluation of teachers or teaching and cannot be used to instruct teachers on how to teach. At most, it can serve as a primer for future research into innovative methods that incorporate student feedback to make classrooms and schools a better environment for students.

# Results

At 11 Topics, reviews were fairly evenly distributed with average percentage of reviews per topic at 9.09% with a standard deviation of (0.014%). The results revealed one data cleaning procedure that consistently resulted in higher coherence scores. This procedure used both a lemmatizer (removes pluralization, conjugation) and stemmer (changes words to their root). Given the informal nature of our dataset, it made sense that the strictest cleaning method resulted in more concise and coherent topics (reviewing the effects of different cleaning approaches on interpretation is beyond the scope of this paper). The other important parameter we experimented with was restricting review length. We trained 4 different sets of models on the target dataset restricted by 80, 100, 125, 150, and 175 character minimums. The size of the dataset shrunk as the minimums rose. To decide between each set of models we averaged the average coherence scores of the models for each of 29 topic numbers we tried (3 to 32 in steps of 1). The 100-character restrictions resulted in considerably higher average coherence scores.

We plotted coherence scores against the number of topics to create what is known as an elbow plot with which we could easily visualize the rate of change in average coherence score. **Graph X** shows the elbow plots for the 100-character minimum models, the average of average coherence scores can be seen on the right. The sharpest drops in the rate of increase of average coherence occurred at the models for 6, 11 and 20 topics. However, when we reviewed the results of the 20-topic model we found many topics shared several key-words which we interpreted as redundancy. The 11-topic model resulted in a higher average coherence score and more topics with relatively high coherence scores than the 6-topic model.

|  |  |  |  |
| --- | --- | --- | --- |
| ***11-Topic Model*** | | | |
| # | CS | Key-Words % of Reviews | |
| 1 | 0.427 | class, bore, like, really, easy, fun, control, felt, pretty, hard, cool, art, nice, make, funny, want, learn, try, joke, pay\_attention, sleep, anything, super, enjoy, act, all, fall\_asleep, think | 11.60% |
| 2 | 0.504 | learn, nothing, test, anything, teach, class, taught, ap, study, prepare, note, book, read, worksheet, all, textbook, exam, quiz, biology, powerpoint, never, teacher, chapter, absolute, know, review, fail, worst, actually, physics | 10.50% |
| 3 | 0.449 | explain, teach, understand, math, teacher, nice, god, cannot, confuse, well, know, hard, person, material, bad, really, anything, horrible, help, worst, terrible, expect, cannot, problem, tutor, lesson, fast, clearly, concept, luck | 10.40% |
| 4 | 0.416 | mean, hate, favorite, like, play, yell, ok, people, think, pick, always, god, teacher, kid, reason, side, annoy, pick\_favorite, really, omg, cannot\_stand, play, say, bad, nice, all, sing, smile, yell, sometimes | 10% |
| 5 | 0.478 | child, school, student, parent, coach, band, principle, music, respect, issue, teacher, administr, daughter, kid, educ, disrespect, unprofesion, office, team, athlete, concern, treat, care, positive, support, program, choir, orchestra, met, hire | 9.50% |
| 6 | 0.480 | talk, life, time, spend, period, sit, minute, whole, story, class, computer, waste, spent, half, phone, room, watch, all, eat, teach, min, site, hear, movie, ramble, complain, around, start, hour, entire | 9% |
| 7 | 0.611 | grade, give, homework, assign, project, work, paper, test, hw, point, essay, quiz, lose, gave, check, due, ton, never, turn, week, time, hard, take, extra\_credit, late, collect, missed, finish, many, absent | 8.60% |
| 8 | 0.450 | got, went, worst, teacher, told, class, college, last, took, thank, freshman, year, glad, still, first, switch, came, english, school, dropped, senior, year\_ago, semester, left, drop, honor, ruin, gave | 8.30% |
| 9 | 0.498 | knowledge, often, student, opinion, lack, poor, style, teach, dificult, extreme, subject, provide, course, skill, ability, discuss, instruct, require, method, however, level, professor, curiculum, class, inform, complete, rather, base, intelligence, standard | 7.90% |
| 10 | 0.414 | need, student, help, spanish, care, speak, french, rude, teach, know, think, language, english, learn, job, better, stop, while, atitude, teacher, retire, understand, try, maybe, speak\_spanish, realized, german, instead, actually | 7.70% |
| 11 | 0.595 | ask, question, answer, wrong, ask\_question, word, say, tell, right, help, look, something, write, board, answer\_question, whenever, said, will, problem, shell, never, figure, even, book, repeat, always, correct, someone, raise\_hand, ignore | 6.50% |

The above table displays each topic’s coherence score, key-words, and the percentage of reviews in which they were present. What follows are the topic descriptions sorted by coherence. We included the fifteen sample reviews for some of the topics in the paper, the others can be found in the appendix or on our GitHub page.

## Topic 7 - Ineffective Grading (8.6% of reviews)

The key-words from Topic 7 made it clear the topic discussed grading around assignments and evaluations. When discussing this topic, we identified a few issues that consistently popped up in the sample reviews:

* *unclear expectations or instructions for work*
* *insufficient time given or overwhelming workloads*

Subjects which also popped up consistently but less often were:

* *strict or unfair grading*
* *teachers losing or returning assignments late*

This characterization of language resulted in an interesting grouping of topics where a clear causal link exists between the former two and latter two subject. In other words, teachers who lose assignments cannot provide feedback in time and students who feel they’ve been marked down point unfairly might reasonably complain about instructions, expectations, and preparation time.

## Topic 11 – Why even ask? (6.5%)

The key words for Topic 11 suggested that reviews grouped here discussed classroom interactions, specifically asking questions. Sample reviews described a diversity of reasons why teachers failed to ask questions but one sentiment was obvious in the students writing these reviews:

* *They were discouraged to ask questions* 
  + *because the teacher wouldn’t provide a helpful answer, or*
  + *because they felt hostility, that questions were unwelcome*

This was a small but very coherent topic which spoke to the importance of cultivating an open environment in the school. This was important enough to students that it motivated them to submit reviews, perhaps in a fit of rage, wave of sadness, or low simmer of disappointment.

## Topic 2 - Did Not Learn (10.5%)

The key-words for Topic 2 suggested that students were complaining about learning outcomes but the sample reviews provided much needed context. What became clear was that students associated teaching with learning. Some consistent subjects appeared in students’ descriptions of the numerous ways their instructors “couldn’t teach”:

* *concern over how lack of learning current and future grades*
* *busy work, book dependent; teacher-centered instruction* (Raufelder et at. 2016)

The key-words *test* and *quiz* also appeared in only one other topic, Topic 7. This and the other key-words each topic suggested that when students were preoccupied with evaluation, they associated it with the issues covered in these topics.

## Topic 9 – Lack of Expertise (10.5%)

The key words for Topic 9 were more general than for topics above but seemed to suggest a lack of expertise in teaching. The sample reviews revealed two general but consistent complaints

* *disorganized course or instructor*
* *lacking knowledge of subject matter*

However, at about the 600th sample review began to diverge more often. Reviews described teachers as imposing beliefs or opinions, being biased, playing favorites or generally described the teacher’s temperament as unfit for teaching and showed little consistency, with disorganization and lack of subject matter expertise appearing less and less frequently. The topic’s coherence score suggested high consistency in the most representative reviews.

## Topic 6 – Wasting Time (9%)

Key-words for Topic 6 suggested reviews that discussed things that happened in the classroom to waste time. The sample reviews confirmed that students were complaining about teachers who wasted time. The most frequent example wasting time were teachers who talked about themselves too much. Despite its lower coherence score, sample reviews for this topic were remarkably consistent. This suggests that the key-words were strong anchors for a diversity of complaints, and language such as *waste* or *spent the whole period talking* was a typical description of this problem.

## Topic 5 – Ruining the School (9.5%)

Based on its key-words, Topic 5 appeared to group reviews written by parents. This was true of the most representative reviews. These complaints focused on administrators and instructors of extra-curricular activities who was perceived to damage the school experience or reputation. The broader pool of sample reviews included descriptions by students and staff who expressed complaints that were consistently similar. Another subject that appeared less frequently were instructors who were judged to be the worst in school.

## Topic 8 – The Worst (8.3%)

Both the key-words and sample reviews suggested that students had a strong dislike of the teacher. We found in interesting that reviews generally contained descriptions of situations rather than descriptions of the teachers or their teaching. The main subject of this topic was interpreted as *stories about why they’re the worst*. These were, however, very varied and did not communicate any particular issue consistently.

## Topic 3 – Can’t Teach (10.4%)

The key-words in this topic that jumped out to us were *explain*, *teach, understand*, and *confuse*. Sample reviews often included subjects found in other topics which explained the low coherence scores. More representative reviews described teachers as being unable to or simply not explaining material, the reviews often included other complaints brought up in other topics but confusion communicated *confusion* or *not teaching* as issues. Interestingly, reviews in this topic captured positive language which could often be found in sentiments such as *nice person, but bad teacher*.

## Topic 1 – Boring (11.6%)

This topic was rather incoherent, key-words and sample reviews suggested that Topic 1 focused on boring teachers and perhaps those that could not control their classrooms. While boredom or *not being fun* did appear to be the most consistent subject among the sample reviews, a variety of other issues were frequently raised.

## Topic 4 – Mean, Short Tempered (10%)

The key-word that stood out in the Topic 4 was *mean*. While there were seemingly random reviews that made it into the sample, the majority of those we read described *relational aggression*. Only one other description appeared consistently under this topic, teachers who picked favorites. However, *favoritism* was also described in the other topics that addressed teachers’ character such as topics 1, 3, and 8.

## Topic 10 – In Need of Change (7.7%)

Topic 10 also discussed teachers’ temperament for teaching. *Need* was a very curious key-word but the sample reviews made it clear that reviews grouped into this topic often suggested that teachers *needed* to change. Mostly, they needed to change their behavior. This topic captured a variety of issues and examples but expressed strong antipathy towards the teacher.

|  |  |
| --- | --- |
| ***Topic 7 – Sample Reviews*** | |
| (18) | *Mr. \_\_\_\_\_ was not a very good teacher. He didn't collect a lot of the assignments he assigned, so we would work hard to complete them and get no credit for it. Also, he would give us "grammar packets" without even going over what was inside and* |
| (19) | *he only gives two out of three on homework and takes so much time going over homework we never get to finish the notes so you won't get the lesson until two days later after the homework a very bad teacher* |
| (20) | *Very unclear about what is expected. Blames you for her mistakes, I was forced to redo a lab report of mine which she lost. Her ownership point system is ridiculous as well, and labs unhelpful.* |
| (99) | *Unfair grading system for writing projects. Unrealistic reading expectations (12 books a trimester!!!!). Didn't like me at all.* |
| (200) | *all she does is makes us take notes... have a quiz/test on and then have a HUGE project. its so stressful...* |
| (299) | *A very bad at Giving directions. Loses alot of students work. gives confusing study guides and hard test oon short notice* |
| (400) | *She can be funny at times, but she is usually strict. She also gives too much homework. Not to mention she has asigned a book report every month. it's a little much.* |
| (800) | *she doesnt teach just makes us read from the book shes super unorganized so one day i would have my hw and she didnt collect it so 4 days later she decides to collect it and i dont have it.* |
| (999) | *I have Mrs. \_\_\_\_\_ right now at \_\_\_\_\_. She gives us too little time for assignments. She gives us too much work for one period, so you never finish.* |
| (1300) | *I didn't like you! You gave us too much homework and you gave us tests about stuff that we hadn't even learned about. plus there was a test every week minimum. good luck newbies.* |
| (1400) | *He gives more work and homework and without teaching. He is the Spanish teacher from Guatemala born and raised and always loses all of the work you turn in. I always turn in my work but he always loses the work I turn in* |
| (1600) | *He grades unreasonably and does not allow enough time for you to understand what we're learning. He fails to be clear on when assignments are due.* |
| (2100) | *She's an awesome person, but she's a horrible teacher. Honestly, having her for the second term is horrible. She is never clear on assignments, returns things late, assigns papers due within the next two days, meh.* |

|  |  |
| --- | --- |
| ***Topic 11 - Sample Reviews*** | |
| (18) | Hmmm....... that is the only way he respnds to you. Doesnt explain nuttin, all he says is look at the drawing. If you do not understand a drawing he will start accusing of cheating. |
| (19) | Dr. Laurie always says he's going to do something, but it never happens! He's a big talker. Never go to him with a problem he'll somehow make it your fault. |
| (20) | all you do is read pages from the textbook and highlight things, and if you ask a question she either doesn't answer or restates something that doesn't anwser the question |
| (98) | You can't understand Tritt at all. He doesn't follow the book and he doesn't answer anyones questions. |
| (198) | He doesn't know wht he's talking about. He always gets problems wrong and we're always correcting him. He'll have the answer book in front of him and still get them wrong |
| (300) | mrs flanders is so oblivious...also she never answers you when you ask questions and when someone i know asked for help she just said no!!! |
| (500) | i had him last year and he never answers the question you ask him but instead he says, "well you tell me?" and forgets that you asked a question. |
| (600) | she ignores me when i have questions and goes as far as to get in people's faces whenever they do something wrong. |
| (800) | great at not answering your questions, and giving you round about answers to anything you ask |
| (900) | She is constantly getting mad over small things and every time you need help she just says look at your notes or "your mom" so honestly I don't think she knows what she is doing...? |
| (1198) | Can't understand him, asking him for help is like talking to the wall. And if you ask for help his response: "Everyone else gets it" (laughing) |
| (1199) | She's really mean, doesn't make me feel comfortable asking a question. She always yells at me for not knowing the answer! I can't ask questions!!! |
| (1299) | Schwieder get mad about all of the little things that should just be ignored. Se makes and example outof you evn if yu get an answer wrong. |
| (1798) | she never listens to what her students say and when you ask for help she doesn't give it. also if you dont know how to read music she makes you teach yourself. |
| (2098) | yells all the time & never lets you ask a question without interrupting, and i can quote her with this: "i just love children!" ok, sure, whatever |

|  |  |
| --- | --- |
| ***Topic 2 - Sample Reviews*** | |
| (18) | His class is way toooooo easy. You dont learn anything in his class, and you will fail in anything that has to do with that class. AP BIO, you will fail the AP Test. Easy A tho. |
| (19) | Seriously the worst history class ever. All he does is give us busy work and he lets us use our books on EVERY test and quiz we take. He is lazy and a horrible teacher. |
| (20) | This teacher is absolutely terrible she doesn' even teach us anything. she just gives you a textbook and reads from slides. tests are horrible she never teaches us anything |
| (100) | She does not teach us anything, she does not know how to teach. She is not teaching us to write essay's or anything. I will not know anything when I go to the High School next year. |
| (199) | did anything but class topic... didn't learn anything in the class.. doesn't teach anything.. has you do random stuff! |
| (500) | Absolutely the most horrible teacher I've ever had, the only way I learned the material was by watching YouTube videos on the chapters |
| (600) | so useless. she never knows what she's teaching so she literally reads everything from the book, which is unhelpful. her tests are nothing like her reviews cause she gets it all off the web. horrible teacher |
| (700) | we have tests almost every class and we fail every one of them.... we go through chapters in less than a week... how are we supposed to learn anything if she teaches too fast??? |
| (800) | she doesnt teach, and we do pointless activities in class. I never learned a thing, she makes it impossible to study fot tests. |
| (1000) | Tests on things barely taught. Takes off for literally everything possible. Barely teaches to the point where anyone recognizes anything on the tests. Switch out immediately. You could have another teacher and not study, and do better than you could do with two hours of studying for Ms. Dassaro. |
| (1100) | All she ever does is give notes, and more notes, o ya, AND MORE NOTES. She dusnt xplain nothin, and then she expects u to ace her exam. |
| (1200) | We had a sub half the year, and she didnt teach us anything, everything that was on the tests we had to learn from worksheets or fail. |
| (1700) | ms johnson is the worst bio teacher i have ever had. she cant teach and her tests are based on her lectures and only part of the material that she teaches you is in the book |
| (2100) | She is definitely one of the worst teachers I've ever had. She never teaches in class and she expects us to get good grades anyways. She loses her temper so easily and can often be found yelling at a student. She expects us to know so much on stuff she hasn't covered at all. In our own research, she expects us to know everything that's going on even though we haven't learned about what to do. Worst teacher everr. |

|  |  |
| --- | --- |
| ***Topic 9 - Sample Reviews*** | |
| (18) | This teacher is very interested in her topic. However, she is unorganized, and her exams are extremely poor in design. She is also often unprepared for class, and cannot answer questions clearly. |
| (19) | Mrs. Ancell isn't well qualified for teaching Chemistry AP. She makes many mistakes but won't admit it when students correct her. Her incompetence makes the class a waste of time. |
| (20) | Great person and personality but unclear teacher. I also believe a reason for the lack of clarity could be the fact that TPSP is a fairly new class/subject and the curriculum is not developed yet. |
| (98) | Overall he made me my opinion on my favorite subject change due to his poor teaching and lack of knowledge. |
| (100) | Inflexible. Uncaring. Presents material, but doesn't lead students to understanding. Avoid. Avoid. Avoid. |
| (200) | Very disorganized. Made a lot of mistakes. Did not own up to her mistakes and misinterpretations. Did not seem to have a good working knowledge of the program. |
| (299) | If you are not a strong science student, then switch out. She does not teach, nor is she able to keep her political/religious views out of the classroom, making it very uncomfortable for students who do not agree with her views. DOESNT TEACH! |
| (300) | she doesn't structure her classes very well and is disorganized...She did offer could insight on books that would've worked better in a Philosophy class rather than English. |
| (399) | He's my current English teacher, and I believe much too laid back. I prefer intense, research based classes; his is one of analysis. |
| (400) | Expresses HIS opinions on the subject and not the facts about the subject. He consistently was improvising on psychology theories that you could tell he clearly wasn't up to date with. Humor is sub-par but makes class interesting a lot of the times |
| (499) | Mrs. Weiniger is a lovely human being I'm sure, but her knowledge, experience, and general ability to teach is severely lacking. DO NOT TAKE AP ECONOMICS if you value your current GPA. You will not le |
| (500) | This person is teaching English? She makes spelling and grammar errors and is totally unable to admit it when she is outright wrong! |
| (600) | Tried to impose his draft-dodging spinless view on the class. Not intelligent, completley out of his league compared to other teachers |
| (700) | Great with jokes and reading off curriculum provided documents, not too great at answering questions or in-depth analysis. |

|  |  |
| --- | --- |
| ***Topic 6 - Sample Reviews*** | |
| (18) | seriously the WORST teacher i could ever have in my LIFE!..she cant stay focused because she is always leaving the class to meet up with another teacher. Get the picture..u have a HUSBAND! |
| (19) | a waste of an hour and a half. why don't you try teaching something instead of just telling stupid stories about yourself and listening to all the cheerleaders! |
| (20) | worst class & teacher I've had, he doesn't teach (He starts "teaching" at the bell) and spends the majority of the class talking about his personal life |
| (100) | Very bad teacher. Spent at least 30 minutes daily talking about unrelated to science topics. |
| (300) | Dr. Farmer, if you ever read this, please stop talking about things that have NOTHING to do with this class |
| (598) | THIS IS NOT STORY TIME!!! Teach us something. What do we care about the first twenty years of your life? |
| (599) | She is only a temporary substitute because the other teacher had a a baby, but she is not goo at all. We spend more time in class talking about her then the actual material. |
| (800) | all she cares about is buffy the vampire slayer buffy this buffy that..... stop talking about buffy thats not suppose to b apart of the lesson |
| (1000) | carlson goes off on tangents all the time, never talks about government. class time is a waste, read the book because what she says won't help you. |
| (1099) | Bad teacher. Blames students for not correcting his mistakes. He literally rambles on about random stuff for the first 30 minutes of class everyday. |
| (1200) | this is one of the sorse teachers at Irv talks more abotu nothing than teaching and what he talks about 90% of the time is wrong like more americans vote for idol than in elections stop talking and me |
| (1300) | we do all the work in teens and tots-it's glorified babysitting and she just stands around and talks to other teachers or leaves us to clean up -BORING |
| (1400) | He talks and TALKS. I mean, when we're supposed to listen to the announcements, he talks about irrelevant things. He gives us work that we might work on for 30 min and that's all. |
| (1800) | he is the worst teacher ever. All he does is talk about being jacked, if he spent half of that time teaching, all the kids in his classes would have better grades. |

|  |  |
| --- | --- |
| ***Topic 5 - Sample Reviews*** | |
| (18) | he destroys what little love for orchestra the students have, demolishes what little confidence the students have, and kills not only the students' passion but also their mental stability as he repeatedly engrains a particular measure of music into students minds by playing the measure over and over and over again. The smiley face tuner taunts and mocks us as we struggle to play our instruments with our dead arms, fatigued by the traumatic unending rehearsals. |
| (19) | Ms. Taylor does not interact with parents or students. She makes school-wide decisions without consulting staff, students, or parents. She does not attend after school events and is not a consistent participant at advisory meetings. Ms. Taylor also does not support the gifted education program or gifted students. |
| (20) | Unfortunately I have had to deal with her for many years. I have never had a positive experience with her and have found she is very conniving. I hope this is the old example of promoting to your highest level of incompetence and she will be released. Unfortunately our children have to pay the price for her being here for so long. It looks like it is time to make big changes to the district office heads as well. |
| (99) | Horrible Principal. Horrible person. Ruined my daughters entire senior year. Has no business in the position she's in. My daughter was bullied all year and not only did she allow it, she led the bullies. |
| (199) | A very, very clever liar. Do not trust this woman. Puts on a great show, but does what she wants, playing favorites among staff and students while ignoring major staffing issues. |
| (200) | coach sirico tries to be better then saint, he will never be better than saint he is a terrible football coach look at our schools record its time for him to go |
| (300) | She has great hs dancers, but self-serving and doesn't build self esteem. Picks favorites and influences student votes. Encourages irresponsible behaviors likedieting and tanning. |
| (498) | He is totally irresponsible and does not respect his students. he is also very immature and does not act like a teacher. he should reprimanded and at worse, fired. |
| (600) | She was very judgmental, uptight, and unkind. After trying to win her favor through sincerity and kindness, I found myself disappointed. She's immovable. |
| (800) | FAITH HUGHES DOES NOT EXIST. she probably would be a good teacher if she did exist, but unfortunately she DOES NOT EXIST. |
| (1000) | This is the worst principal u could ever have at a skwl! He doesn't care about anyone but himself! no wonder we are the worst skwl ever! NEW PRINCIPAL PLEASE! |
| (1300) | Mrs. Chabala is one of the worst teachers I have ever had and it is an embarrassment to the school to have such a bad teacher under its employment |
| (1600) | He's taking away all the traditions that we have at Owen. That's totally wrong, especially considering that many of the teachers actully attended school at Owen |
| (1900) | Markwith obviously doesn't have the social skills to run a school, or be in any position of power (no matter how pathetic a position it is) but really Greg, this job is NOT for you |

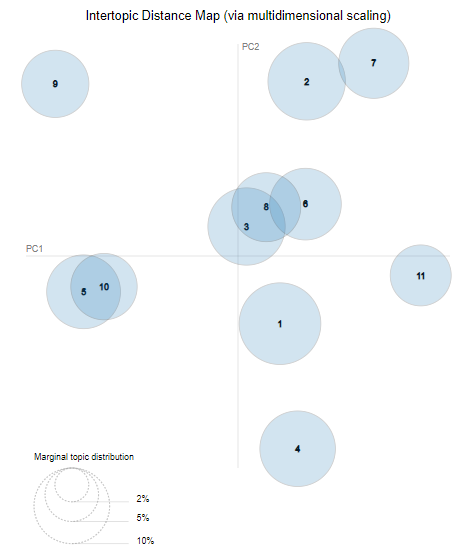
|  |  |
| --- | --- |
| ***Topic 8 - Sample Reviews*** | |
| (18) | Total airhead. I took her class in 1992; 14 years later, having finished graduate school (in science) I remember her as my second worst science teacher ever. |
| (19) | Anna says one thing and does another. I got an A in the class. Never went to class and passed my work in late. Got to love Quincy College. |
| (20) | There are 2 7th graders in her advanced 8th grade algebra class& the other day she told us both that we aren't smart enough to be in the class. This was completely inapprotie, espically since she's only a long term sub AND we are BOTH getting straight A's! |
| (100) | i hate him he failed me for no good reason and didnt recrmond me for adv so im back in ALEGERBRA 1 WHEN I ALREADY TOOK IT IN MIDDLE SCHOOL >:O AND I HAD HIM FOR GEO |
| (200) | i had her for 3 years...1st year-hated her..2nd year-hated her...3rd year- not as bad...but still hated her |
| (400) | i had her in 5th grade and left because of her i hated her she hated me she told me i wouldn't have a bright future |
| (600) | had keating for algebra 1 and didnt understand a thing --i always thought i hated algebra...then i got to high school and took algebra 2 and realized it wasnt algebra i hated. |
| (800) | Sabatini told me there wasn't a history regants in January and that I'd have to wate till June. When I found out it was to late to take it. Now I have only 1 more chance to pass it before graduation. |
| (1100) | Worst teacher ever. can not take her. called house 5 times during the year. emailed my mom with 5 days of school left. least favorite teacher ive ever had. good luck! |
| (1200) | She has been the worst teacher ever I got a concision at the beginning of the second quarter and I was 2 weeks behind in everything every other teacher helped me told me what I needed what I needed to do but ms Dias didn't care at all |
| (1300) | Most miserable man i have ever met, hates to help his students and hates it when you are absent or miss his class |
| (1800) | I'm a science teacher now because of him. Not because he was so wonderful, but because I hated his class so much and wanted to learn science so badly. |
| (2000) | Several years ago she told me I couldn't even get into a Cal State because I had failed Honors Trig. I ended up getting into a top 10 USNWR private school. Don't let teachers get you down kids! Upward and onwards! |
| (2100) | I hate her I she is so evil I am sure that you have heard about her 2007-2008 5th period class having to write that essay. Well I said BOOM and got a flippin ASD from her! I absolutly HATE her! |

|  |  |
| --- | --- |
| ***Topic 3 - Sample Reviews*** | |
| (18) | Ms. Novillo dosen't even know anything about what she is teaching. She dosen't even explain most of her material. She kind of expects you to come to her class knowing how to do the material. |
| (19) | shes a really nice teacher, but shes not that great at teaching. ive always been pretty good at math, but now im failing because she doesnt teach well at all. switch out if you get her! |
| (20) | She of one of the worst Earth Science teachers. If you do not understand it, you never will. She never goes over anything or explains topics into great detail, and teaches off of a SmartBoard. ---NJR |
| (200) | She is a really really nice lady but not a very good teacher. :( Youll have to get yourself a tutor unless you are a genius at pre-calculus. |
| (300) | awful teacher. can be sometimes funny but is overall a bad teacher. he doesnt explain anything. pray you don't get him |
| (400) | Worst math teacher ever. Nice person, but that doesnt help you in math. He has no idea how to teach effectivley what so ever. |
| (500) | No one can understand and she makes so many mistakes. But she is a really nice person, TERRIBLE teacher. |
| (600) | Nice person, terrible teacher. She doesn't seem to teach what she tests on. very unclear!!!!!!!!!!!!!!!!!!!!!!! |
| (799) | He's a good teacher if you already understand the material, but I didn't know what was going on, and he didn't teach me. |
| (899) | He's a funny guy but an awful teacher. He grades essays too harshly and he doesn't teach us anything but how to read a book and answer questions about it. |
| (1200) | Good luck if you get her. She makes everything so much more difficult than it has to be. |
| (1299) | SHE IS A HORRIBLE TEACHER!!! she's a really nice person but her teaching skills aren't to amazing if she tried to teach someone to add 2+2 they'd get 22 but she is a nice person |
| (1400) | One of the worst at Franklin. Not just because she's not 'cool' or 'laid back' or whatnot. Because she can't teach. She's a teacher, that can't teach. She is worse at teaching than any other teacher. |
| (1600) | horrible teacher, definitely can use improvement, can't teach, but otherwise really cool. she is a really nice person but just can't teach or doesn't know HOW to teach. |

|  |  |
| --- | --- |
| ***Topic 1 - Sample Reviews*** | |
| (18) | Mrs. Ugalde should be nominated for an Oscar. She acts like she cares, but all she cares about is making herself look good. She acts like we are dumb. She acts like she's a great teacher, but she doesn't know what that is |
| (19) | i dont really like her....im taking creative photo and we are like learning about like paintings and stuff...hello its not an art class its a photo class |
| (20) | L-Fels. Great guy... Not-so-great teacher. I wouldn't say you leave the class having actually learned anything. We never get anything done during class, either. |
| (100) | Doc. T is a decent teacher, however he is old fashion in his teaching and most of his classes are boring long lectures. He could be better if he was creative. |
| (200) | Everyone thinks he is cool, and he very well may be.. but he really isnt. I did not learn much from him and I dont think he is all that fun. |
| (300) | jerkkk. pretends to be your friend but all he really cares about is his 2 kids. has no control over the class either. |
| (400) | i dont rally like mr D and the most class that i really heat is his class sooooooooo Boringggggggg.Ijust heat that teacher and his class. |
| (500) | He makes us run for to long and lets us out late. He makes gym so serious and not fun |
| (600) | She is so boring!!!! She likes to think she is cool too. Really easy because she duznot like to work herself. |
| (700) | Its the easiest class ever. He doesnt teach at all, so if you arent trying to do any work. Take this class |
| (900) | He keeps in one monotone which makes class quite boring even though his class is like super easy. a big problem is that he acts like he's teaching kindergarten |
| (1100) | i didn't learn anything in this class. all he talks about is the 99 cent store. it seems like he doesnt know anything. i dont think he does. |
| (1300) | thinks he's cool, but not. tries too hard to be your friend and doesn't really "control" the school; clueless |
| (1400) | She try's hard but really just doesn't have real knowledge. Good activities though. Just not real bright. |

|  |  |
| --- | --- |
| ***Topic 4 - Sample Reviews*** | |
| (19) | mr. henson is annoying and weird and very grouchy. he is also kind of loopy. he gets mad for silly reasons and tries to be sarcastic but it sounds like hes serious. i dont like him very much. hes way to nice to the basoons and too mean to everybody else |
| (20) | mrs.nut is so weird, but I feel kind of bad for her because these kids made a song for her and it is kind of mean, but she dos'nt know so its all good. the only thing good about her is she gives candy |
| (100) | hes a lame teacher...if ur a jock ur set everyone else is dead...stoners r doomed even if u do everything he still hates u...kinda funny 2 mess with him tho |
| (200) | I really didn't like her. She was arrogant and pretty much thought she was always right...if you disagreed with something she said, then watch out. Snobbish and annoying. |
| (300) | this is the only person i think of as a bad person the nasty lady even said she thought of herself as always right |
| (400) | I'm sorry, i couldn't stand miss pyle. She had her favorites(they know who they are) and she never called on anyone else except her favorites. |
| (500) | mr hissong is realy weird. I think that he has mood swings. ond day he will be realy nice and the next he will bite you head off. he is very mean |
| (600) | Her name is Packevicz and i think she is the worst teacher in the world.. not funny at all and totally scitzofrenic i hate her so much |
| (900) | I had her 1987-1988. She was not a favorite of mine. If I remember right, she played favorites alot, and I was certainily not a favorite. She hated me. |
| (1100) | he is a really great guy but sometimes he doesn't teack at all. he is a bad teacher but a good friend |
| (1200) | omg. if she doesn't like you...ahem like me...she will unleash her wrath on you like there's no tomorrow...beware |
| (1698) | VERY VERY VERY NUTTY AND INSANE! Yells way to much and picks favorites! Is obsessed with the "Fountain of Air" |
| (1800) | u r mean but some times nice you have a really bad attiude i thought i liked u but i dont and u need to stop yelling at joey s. he is cool and very nice . |
| (1999) | this teacher is the wosrt teacher ever.she is so mean and yells at u for everything. she gives lunch detetions for every little thing and she always thinks she is right when most of the time she isnt |

|  |  |
| --- | --- |
| ***Topic 10 - Sample Reviews*** | |
| (18) | Insufferable know it all. Only knows how to teach the kids that would not need his help to learn. And calls the kids that need help dummies. |
| (19) | She needs to realize that it hurts people's feelings when she makes them feel stupid for not understanding something. C'mon Mrs. Sickler, help us don't hurt us! |
| (20) | he is the worst teacher i think i ever had.... he thinks he is all that, but in reality doesnt know how to teach at all... he doesnt get it when people just DONT UNDERSTAND something.... |
| (100) | She thinks she teaches well, but she doesn't even teach. All she says is: HELLO? she needs to retire!!! |
| (200) | She needs to chill out..needs to worry about more important things pther than getting students in trouble |
| (398) | Very uptight, always has an attitude, i don't believe there is a soul she is friendly to. I understand that you have to be strict with the students, but she takes it to a whole new level, i think she |
| (400) | she is really rude sometimes and I think she needs to be able to get to know her students better and everything will become much easier! |
| (699) | you embarass your students and make them feel stupid, you don't take the time to help them, although you offer, but you're very impatient |
| (700) | This teacher thinks she is a teenager. She needs to mind her own business and act her age. She is always in students business and talking about them behind her back. |
| (899) | You need to take a chill pill. If you put as much effort in your teaching as you do your screaming, we might learn something |
| (998) | I dislike her teaching method. She need to act kinder to students and maybe more students will like her. |
| (1098) | she is one of the most self centered teachers in schreiber. not only does she not care about the students well being she also doesnt care about her teaching skills. she will go on rants for an hour ab |
| (1298) | I think that she doesn't explain anything and claims that she doesn't have time to help you understand better. She is a teacher she should not REFUSE to help students learn better! |
| (1398) | Shes Getting a little bit better with her atitude but still shes really strik about her rules.mrs.raimo change your atitude a little bit and be nicer to kids |



# Discussion

## Evaluation of the Model

The average coherence score for the 11-topic model was 0.483 which was low compared to the 0.52 benchmark for very unstructured datasets set out in Röder et al (2015). We can trace this low score back to the shared language across topics and lack of structure in the data collection process. Coherence is a relative measure to help select a model and compare topics to one another and not generally indicative of whether a model is useful or should be discarded. However, we did get a good impression of how coherence scores measured meaningful deviation within a topic by reading samples from the 2,100 most representative reviews.

Topics with low coherence scores that were described as inconsistent included more reviews that had little or nothing to do with the subjects most frequently being discussed in the most representative reviews. Instrumentally, this expressed the diversity of contexts in which language highlighted in the topic was used. Thus, we understood coherence as a measure of sparsity since the main subject, if there was one, did not disappear but simply became less frequent or present in reviews. For our results, this implied that the percentage of reviews attributed to each topic was overestimated for less coherent topics.

## Small “p” Policies Based on Student Descriptions

The descriptors used by the raters, and the topics in which they were, provide valuable information to everyone with responsibility for the supervision, professional development, and evaluation of teachers. In particular, principals who note students talking about or describing a teacher using these descriptors should recognize that it may well be a signal that something about the classroom teacher, classroom practices, or classroom culture, is seriously amiss.

*Behind the Curtain of Unfair Grading*

Topic 7 was very constructive towards this type of policy. It revealed that one of the reasons students gave teachers bad ratings was because of how they organized or mishandled grading and feedback. Most notably, it revealed that they focused on the issues of “*unclear expectations or instructions for work*,” “*insufficient time given or overwhelming workloads*,”“*strict or unfair grading*,” and “*teachers losing or returning assignments late*.” This suggests that when administrators or other teachers find that students always describe a teacher as an “unfair grader” they may probe their students to see if they also found a problem with instructions, expectations or getting grades and feedback on time.

Topic 2 expanded on students’ concerns over grades. Generally, students rated teachers poorly because they failed to learn which they often tied to concerns over current or future grades. The example that was most readily provided in these reviews were those of *busy work*. This grouping revealed that students took issue with a lack of purpose in their work. One could also see how *busy work* might also cause or stem from issues raised in Topic 7. This analysis suggests that when students express concern over learning and grading, these issues may be at the root of the problem.

*Creating Nurturing Environments*

Some teachers were rated poorly because their students felt as if they were unable to ask questions in the classroom. This topic was contrasted by topics 3 & 9 which captured descriptions of teachers who were unable to explain material or lacked expertise. In other words, Topic 11 revealed that when students’ main complaint was about asking questions in the classroom, it was because their teachers’ behavior discouraged them from doing so. If complaints about a class suggest that students feel an inability to interact, this may suggest a strained student-teacher relationship with, at least, a part of the class.

## Contribution to the Literature

In this section we compared our findings to the literature on “bad teachers,” particularly Raufelder et al. (2016) whose themes and sub-themes we described in the introduction. We differentiate our findings from Raufelder’s by describing ours as topics and theirs as themes and sub-themes.

We found a number of topics that fit into the themes described by Raufelder. Topic 4, *mean, short-tempered* (10%) fit very well within the sub-theme of *relational aggression*. Topic 2, *did not learn* (10.5%) clearly overlapped considerably with the sub-theme *teacher-centered instruction*. Topic 3, *can’t teach* (10.4%) coincided with the sub-theme *incomprehensible teaching*, especially because we both found that students do not take failures to teach personally and such that a good student-teacher relationship can exist despite this issue. Topic 11, *why even ask?* (6.5%),which focused on behavior that dis-encouraged questions, fit neatly into the theme of *personal characteristics*.

Less coherent topics such as Topic 1, *boring* (11.6%), Topic 8, *the worst* (8.3%), and Topic 10, *in need of change* (7.7%), which clearly addressed teachers’ character but also described strained student-teacher relationships, would probably be split between the themes *student-teacher relationship* and *personal characteristics*. Interestingly, in our analysis, most of the topics that expressed strong antipathy and attacked teachers’ character or manner received lower coherence scores. This resonated with Raufelder who also found it difficult to identify consistent sub-themes under the theme of *personal characteristics*. Given that at least 34.1% of reviews (topics 1, 8, 10, and 11) fell under this theme, we also coincided with Patalano (1978) who found that most students valued their teacher’s personal qualities above their professional skills.

Other topics deviated more noticeably from the findings in Raufelder et al. (2016). Topic 9, *lack of expertise* (10.5%) discussed disorganization and lack of subject knowledge. These issues popped up together consistently in the most representative reviews for this topic (recall that its percentage was overestimated). Raufelder did not connect these two subjects and discussed a lack of knowledge as a driver of antipathy under the theme of *student-teacher relationship* and discussed disorganization in the commentary for *personal characteristics*. Where the students in our data suggested or perceived, rather intuitively, that disorganization and frequent mistakes stemmed from a poor handle of the material.

Topic 7, *ineffective grading* (8.6%) reflected an importance on and feedback grading that was mentioned in Raufelder’s descriptions of good teachers under the sub-theme of *appreciation*. Our findings suggested that students were considerably more critical of how teacher’s instructions, expectations and duties affected their ability to receive good grades. Topic 2 reflected a similar importance put on grading. We found that one of the reasons students rated teachers poorly was that they failed to provide students the opportunity to organize themselves for success.

Only two topics found in our study were absent from Raufelder’s study. Topic 6, *wasting time* (9%) described teachers who talked too much about themselves or did other things that had nothing to do with teaching. Some reviewers did not mind having distracted teachers but many took issue when they perceived their time was not being respected. This value for students’ time was not reflected in Raufelder et al. (2016) though we initially attempted to relate it to the *disinterested or indifferent to material* sub-theme. Interestingly, this topic was somewhat present in the *no effort* theme highlighted by Chang-Kredl and Cloannino (2017). Topic 5, *ruining the school* (9.5%) was also completely absent from Raufelder’s paper but that is because it mostly dealt with administrators and extra-curricular activities which was not relevant to their study.

Our work is also consistent with Dorham’s (1987), cited earlier, in that the topic we found could all be fit into the three themes they regarding the ratings of good teachers (instruction, personality, and classroom management). We found Raufelder’s themes and sub-themes useful. However, we would suggest adding two sub-themes based on topics 6 and 7 under the theme of *lack of expertise in teaching.* The first called *ineffective grading & workload management* which would address students’ concerns around setting expectations and getting feedback. The second called *failure to fulfill duties*.

# Conclusions

Perhaps the most outstanding insight we gained or regained is how much enjoyment students’ ideas and comments contain.  We were attracted to these data because we simply liked reading comments from the people who occupy our workplaces. And how they can point to simple policy and practice changes to improve learning for students and job satisfaction for teachers. Our paper explores two types of small “p” policies. The first type of policy is based in language, the second is based in the ability to collect comments online.

To develop policy based in language, in the discussion section we operationalized two key insights which were derived from our analysis of the language students used to describe teachers who were rated poorly. Consistent with Raufelder’s (2016) description of their analysis, we found that most negative reviews addressed teacher’s behavior or personality, and that these complaints used language that was less consistent and more difficult to categorize. However, reviews that discussed teachers’ knowledge, ability and commitment appeared to be more thoughtful, more sophisticated, and more consistent than we had expected. This suggests that students can be critical and insightful evaluators of their teachers, *if they are asked*!

We believe our findings also support the policies that establish online forums as a means to collect student feedback. Many issues highlighted in our data would be undetectable to other forms of teacher evaluation, some even went undetected in earlier studies which, again, speaks to the value of open forums for active students. For example, one topic we found highlighted teachers who *wasted students time* doing things other than teaching (this resonated with the duties-based evaluation suggested by Scriven, 1994). Perhaps the most surprising values reflected in the topics we found is that students also care deeply about their own learning. They expect to learn from the adults in school. We found that many students who voluntarily chose to comment about their teachers cared deeply about two particular things: *1) Achieving more in school, and 2) being treated fairly and with dignity*. Moreover, rating teachers as “bad” teachers because they stood in the way of student learning and, consequently, student achievement, is a finding that does not conform to the public’s image of contemporary K-12 students.

There are opportunities for students and teachers to learn better communication which will undoubtedly be one of the most important skills of not only the present but also the future. Schools need changes which are not costly, are supportive and take myriad variables into account. Administrators, we believe, should ask teachers themselves and their students to operationalize these themes and be able to identify them in the classroom.

Because students’ demonstrated competence as observers and critics, we recommend that more districts and schools develop policies to use students--those closest to daily classroom instruction-- to evaluate their teachers. Without being determinant of promotion and salary, students’ analyses appear quite capable of providing extraordinarily rich data about *their* lives in classrooms. We believe, as well, that with a little training, students could be even better prepared to provide both the positive and negative feedback that school administrators need to improve their schools.  Thus, we urge further investigations into using student feedback in formative ways to improve instruction, and as summative evidence of school improvement.

# References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Statistical Association (2014, April 8). *ASA Statement on Using Value-Added Models for Educational Assessment*. Retrieved December 9, 2019 from: https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. Philadelphia, PA: Routledge/Taylor and Francis Group.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, 20 (12). retrieved December 9, 2019 from http://epaa.asu.edu/ojs/article/view/1096

Angel, K. (2009). Be alarmed, be very alarmed: Federation urges teachers not to engage with the *ratemyteachers.com* website in any capacity. *Education*, 90(9), 14.

Author. (2014).

Author. (2018a).

Author. (2018b).

Berliner, D. C. (2019). Using the Social and Behavioral Sciences to Challenge the Political Roots of Educational Policy. Paper Presented at the XIV Congreso Chileno de Psicologia, November 13, 2019, Universidad de Tarapaca, Arica, Chile

Bill and Melinda Gates Foundation, (2012). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*. Retrieved December 10, 2019 from https://docs.gatesfoundation.org/documents/preliminary-findings-research-paper.pdf

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

Boring, A., Ottoboni, K. & Stark, P. B. (2016, January 7). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen*. Retreived October 1, 2020 from https://www.scienceopen.com/document/read?vid=818d8ec0-5908-47d8-86b4-5dc38f04b23e

Borko, H., Livingston, C. & Shavelson, R. J. (1990). Teachers' Thinking About Instruction. *Remedial and Special Education*, 11 (6), 40-49. Retrieved 12/10/2019 from<https://doi.org/10.1177/074193259001100609>

Butrymowicz, S. (2014, June 16). How many bad teachers are there? *The Hechinger Report.* Retrieved December 10, 2019 from https://hechingerreport.org/many-bad-teachers/

Burdick, J., & Sandlin, J. A. (2010). Inquiry as answerability: Towards a methodology of discomfort in researching critical public pedagogies. *Qualitative Inquiry*, 16(5), 349-360.

Burdick, J. (2009). The public construction/constriction of teachers: RateMyTeachers.com and the complicated pedagogies of the educational imaginary. *The Sophist’s Bane*, 5(1/2), 53-58.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology,* 3 (2), 77-101.

Chang-Kredl, S. & Cloannino, D. (2017). Constructing the image of the teacher on Reddit: Best and worst teachers. *Teaching and Teacher Education*, *64* (43-51). Retrieved December 10, 2019 from http://dx.doi.org/10.1016/j.tate.2017.01.019

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measure of teacher effectiveness in the Pittsburgh Public Schools*. (REL 2014–024). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Check, J.F. (1986). Positive traits of the effective teacher-negative traits of the ineffective one. *Education*, 106(3), 326-334.

Check, J.F. (1999). The perceptions of their former teachers by older adults. *Education*, 120(1), 168-172.

Cruickshank, D. R., & Haefele, D. (2001). Good teachers, plural. *Educational Leadership,* 58 (5), 26-30.

Cuban, L. (2012). Students Evaluating Teachers. [Larry Cuban on School Reform and Classroom Practice](https://larrycuban.wordpress.com/). Retrieved December 10, 2019 from https://larrycuban.wordpress.com/2012/12/11/students-evaluating-teachers/

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Washington DC: Association for Supervision and Curriculum Practice.

Danielson, C. (2008). The handbook for enhancing professional practice: Using the framework for teaching in your school. Alexandria, Virginia: Association for Supervision and Curriculum.

Dorham, J.K. (1987). [*Sixth Grade Students' Perceptions of Good Teachers*](https://search.lib.asu.edu/discovery/fulldisplay?docid=alma991037474109703841&context=L&vid=01ASU_INST:01ASU&lang=en&search_scope=MyInst_and_CI&adaptor=Local%20Search%20Engine&tab=Everything&query=any,contains,good%20teachers&offset=0)*.* S.l. : Distributed by ERIC Clearinghouse, Number: ED359164

Freishstat, R. L. (2016). Expert report on student evaluations of teaching (SET). Retrieved October 1, 2020 from: https://ocufa.on.ca/assets/RFA.v.Ryerson\_Freishtat.Expert.Supplemental.Reports\_2016.2018.pdf?utm\_source=OCUFA+Report&utm\_campaign=7bb120ce70-EMAIL\_CAMPAIGN\_2018\_07\_12\_01\_15&utm\_medium=email&utm\_term=0\_458512323c-7bb120ce70-&mc\_cid=7bb120ce70&mc\_eid=%5BUNIQID%5D

Good, T. L. and Lavigne, A. L. (2018). Looking in Classrooms, 12th edition. Pearson.

Goodwin, L. (2016). Defining teacher quality. In Drew H. Gitomer & Courtney A. Bell (Eds.)

*Handbook of research on teacher education.* NY: Routledge, 399-403.

Goodwin, L. & Oyler, C. (2016). Teacher educators as gatekeepers. In In Drew H. Gitomer & Courtney A. Bell (Eds.) *Handbook of research on teacher education.* NY: Routledge, 468-489.

Grossman, P., Cohen, J., Ronfeldt, M, & Brown, L (2014). The Test Matters: The Relationship Between Classroom Observation Scores and Teacher Value-Added on Multiple Types of Assessment. *Educational Researcher,* 43(6), 293–303.

Gurl, T. J. J., Caraballo, L., Grey, L., Gunn, J. H., Gerwin, D. & Bembenutty, H. (2016). *Policy, professionalization, privatization, and performance assessment: Affordances and constraints for teacher education programs*. New York: Springer.

Haertel, E. H. (2013). Reliability and Validity of Inferences About Teachers Based on Student Test Scores, The 14th William H. Angoff Memorial Lecture, Princeton, NJ: Educational Testing Service. Retrieved December 10, 2019 from https://www.ets.org/research/policy\_research\_reports/publications/publication/2013/jquq

Hattie J (2008) Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. Abingdon: Routledge.

Hattie, J. (february, 2015). High Impact Leadership, *Educational Leadership*, Association for Supervision & Curriculum Development, 36-40.

Hosgorur, T. According to former school students’ viewpoints, what aspects turn a bad teacher into a good teacher? *Anthropologist*, 19(3), 819-828.

Jackson, P. (1990). *Life in Classrooms*. New York: Teachers College Press.

Jacobs, H. H. (2012). Curriculum 21: Socrates Fails Teacher Evaluation. Retrieved December 10, 2019 from http://www.curriculum21.com/?s=Socrates

Kane, T. J., McCaffrey, D. F., Miller, T. & Staiger, D. O. (2013). *Have we identified effective teachers?* Validating *measures of effective teaching using random assignment.* MET Project Research Paper. Retrieved May 4 2017 fromhttp://k12education.gatesfoundation.org/resource/have-we-identified-effective-teachers-validating-measures-of-effective-teaching-using-random-assignment/

Lawrence, J. W. (2018). Student evaluations of teaching are not valid. *Academe*, May-June 2018, American Association of University Professors. Retrieved December 10, 2019 from https://www.aaup.org/article/student-evaluations-teaching-are-not-valid

Lewis, D. D. (1992). Text representation for intelligent text retrieval: A classification-oriented view. In P. S. Jacobs (Ed). *Text-based intelligent systems: current research and practice in information extraction and retrieval (*pp. 179-197). Mahway, NJ: Lawrence Erlbaum.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). Classroom instruction that works: Research-based strategies for increasing student achievement. Alexandria, VA: ASCD

Mimno, D., Talley, E., Leenders, M., Wallach, H. M., & McCullum, A. (2011). Optimizing semantic coherence in topic models (pp. 262-272). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language processing*. Edinburgh, Scotland: Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved December 10, 2019 from *arXiv:1301.3781*.

Morgan, G. B., Hodge, K. J., Trepinksi, T. M., & Anderson, L. W. (2014). The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. Education Policy Analysis Archives, 22(95). http://dx.doi.org/10.14507/epaa.v22n95.2014

Morgan, G. B., Hodge, K. J., Trepinski, T. M., & Anderson, L. W. (2014). The Stability of Teacher Performance and Effectiveness: Implications for Policies Concerning Teacher Evaluation. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, 22, 1-21. Retrieved December 10 2019 from https://www.redalyc.org/pdf/2750/275031898117.pdf

Newton, X.A., Darling-Hammond, L., Haertel, E., and Thomas, E. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability Across Models and Contexts. *Educational Policy Analysis Archives*, 18. Retrieved December 10, 2019 fromhttp://dx.doi.org/10.14507/epaa.v18n23.2010.

Özgüngör, S., & Duru, E. (2015). Course and instructor characteristics distinguishing highest and lowest student ratings of instructors. *Eurasian Journal of Educational Research*, 61, 118-136.

Papadimitriou, C. H., Raghaven, P. & Tamaki, H. & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*, (2), 217-235.

Patalano, Frank. "School psychology graduate students' perceptions of effective and ineffective teachers." College Student Journal (1978).

Pedersen, E., Faucher, T. A., & Eaton, W. W. (1978). A new perspective on the effects of first-grade teachers on children's subsequent adult status. *Harvard Educational Review, 48*(1), 1–31. https://doi.org/10.17763/haer.48.1.t6612555444420vg

Peneul, W.R. & Shepard, L.A. (2016). Assessment and teaching in In In Drew H. Gitomer & Courtney A. Bell (Eds.) Handbook of research on teacher education. NY: Routledge.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). Classroom Assessment Scoring System (CLASS) Pre-K Version. Baltimore, MD: Brookes Publishing.

Pivopirova, M., Amrein-Beardsley, A., & Broatch, J. (2016). Value-Added Models (VAMs): Caveat Emptor. *Statistics and Public Policy*, 3 (1), 1-9. Retrieved December 9, 2019 from https://doi.org/10.1080/2330443X.2016.1164641

Polikof, M.S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, 121, 183-212.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora.

Retrieved March 1, 2012 from https://github.com/RaRe-Technologies/gensim.

Řehůřek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. NLP Centre, Faculty of Informatics, Masaryk University.

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*(3), 137-152.

Raufelder, D., Nitsche, L., Breitmeyer, S., Kessler, S., Hermann, E., & Regner, N. (2016). Students’ perception of “good” and “bad” teachers—Results of a qualitative thematic analysis with German adolescents. *International Journal of Educational Research*, 75, 31-44.

Ripley, A. (2012, October). Why kids should grade teachers. *The Atlantic*. Retrieved December 10, 2019 from https://www.theatlantic.com/magazine/archive/2012/10/why-kids-should-grade-teachers/309088/

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Shanghai, China: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399-408. Retrieved December 10, 2019 from https://doi.  [10.1145/2684822.2685324](https://doi.org/10.1145/2684822.2685324)

Rodin, M. & Rodin, B. (1972). Student evaluations of teachers: Students rate most highly instructors from who they learn least. *Science*, 177, 1164-1166.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). *Evaluating topic coherence measures*. Retrieved 3 March, 2020 from arXiv preprint arXiv:1403.6397

Scriven, M. (1994). Student Ratings Offer Useful Input to Teacher Evaluations. *Practical Assessment, Research, and Evaluation, 4*, Article 7. Retrieved December 10, 2019 from https://scholarworks.umass.edu/pare/vol4/iss1/7

Scriven, M. (1994). Duties of the teacher. *Journal of Personnel Evaluation in Education, 8* (2), 151-184. Retrieved December 10, 20719 from DOI: 10.1007/BF00972261

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces, 63-70*. Baltimore, MD: Association for Computational Linguistics.

Stark, P. B. & Freishtat, R. (2014, 29 September). An evaluation of course evaluations. *ScienceOpen Research*. Retreived October 1, 2020 from: https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Stevens, K., Kegelmeyer, P., Andrzejewski, P. & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* Jeju Island, Korea, *952-961*. Stroudsburg, PA.: Association for Computational Linguistics.

Stronge, J. H. (2007). *Qualities of effective teachers* (2nd ed.). Alexandria, VA: Association of Supervision and Curriculum Development (ASCD).

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement." *Journal of Teacher Education*, *62* (4), 339-355.

Strunk, K., Weinstein, T., & Makkonen, R. (2014). Sorting out the signal: Do multiple measures of teachers’ effectiveness provide consistent information to teachers and principals? *Education Policy Analysis Archives, 22* (100). Retrieved December 10, 2019 from http://dx.doi.org/10.14507/epaa.v22.1590

Teh, Y. W., Jordan, M. J., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems* (1385– 1392). Vancouver, BC and Cambridge, MA: MIT Press.

Uttio, M. (2012). ‘Behind every profession is a person’: Students’ written memories of their own teacher-student-teacher relationships. *Teaching and Teacher Education*, 28, 293-301.

Appendix

One other way we evaluated how well the topic model represented the text was by evaluating how many reviews submitted by parents were grouped into Topic 5. Reviews submitted after August, 2015 sometimes had labels identifying whether they had been submitted by parents or students. We compared reviews dominated by Topic 5 to the “Submitted By” label in the data and confirmed that our model was able to accurately identify reviews submitted by parents. There were 4,448 parent reviews that were longer than 100 characters and rated between 0 and 35. Of these 4,448 reviews, 46.6% of them were dominated by Topic 5. Another 15% of the parent reviews had Topic 5 as the second most likely topic. Topic 9 was a distant second which dominated only 13% of the reviews submitted by parents. These findings also suggested that parents who were moved to submit reviews on RateMyTeacher.com were generally motivated by the strong antipathy reflected in Topic 5. The fact that our language model was able to accurately identify reviews submitted by parents increased our confidence that the language model we trained and selected yielded some consistent and coherent topics.

1. As a point of reference for coherence scores we used the results presented in Röder’s 2015 paper which set benchmarked coherence scores to human ratings of coherence for several datasets of varying complexity. The most difficult dataset to evaluate for all measures was a “Movie” dataset originally used by Rosner, 2014 in which the score for the “Cv” measure was .548, the next most difficult data sets yielded scores of 0.627, 0.655 and 0.671. Each was trained using the text being classified (other methods were also evaluated in the paper). [↑](#footnote-ref-1)