education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 29 Number 79 June 7, 2021 ISSN 1068-2341

The Value of Student Feedback in Open Forums: A Natural Language Analysis of Descriptions of Poorly Rated Teachers

Carlos Valcarcel

Jefferey Holmes

David C. Berliner

&

Mari Koerner

Arizona State University

United States

Citation: Valcarcel, C., Holmes, J. Berliner, D. C., & Koerner, M. (2021). The value of student feedback in open forums: A natural analysis of descriptions of poorly rated teachers. *Education Policy Analysis Archives, 29*(79). https://doi.org/10.14507/epaa.29.6289

Abstract: In this paper we used natural language processing to review hundreds of thousands of negative student reviews of their teachers submitted to the website RateMyTeacher.com. Our analysis identified several issues raised by students when rating teachers poorly, which adds to the literature that defines "bad teachers" from the student perspective. We also identify the language students used to describe these issues and notice a clear distinction between the language used to address teaching-related complaints and behavior perceived as unfit for teachers. We argue that digital forums can be valuable tools for schools and conclude with suggestions and examples of the type of policies that may be derived from this type of analysis of a digital forum for students.

Keywords: student evaluation of teacher performance; natural language processing; feedback (response); school policy; high schools; middle schools

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 20/11/2020
Revisions received: 3/18/2021
Accepted: 3/19/2021

El valor de las respuestas de los estudiantes en foros abiertos: Un análisis natural de las descripciones de los profesores que recibieron una mala calificación

Resumen: En este documento utilizamos el procesamiento del lenguaje natural para revisar cientos de miles de reseñas negativas de los estudiantes sobre sus maestros enviadas al sitio web RateMyTeacher.com. Nuestro análisis identificó varios problemas planteados por los estudiantes al calificar mal a los maestros, lo que se suma a la literatura que define a los "malos maestros" desde la perspectiva del estudiante. También identificamos el idioma que los estudiantes usaron para describir estos problemas y notamos una clara distinción entre el idioma usado para abordar las quejas relacionadas con la enseñanza y el comportamiento percibido como inadecuado para la enseñanza. Argumentamos que los foros digitales pueden ser herramientas valiosas para las escuelas y concluimos con sugerencias y ejemplos del tipo de políticas que se pueden derivar de este tipo de análisis de un foro digital para estudiantes.

Keywords: evaluación estudiantil del desempeño docente; procesamiento natural del lenguaje; retroalimentación (respuesta); política escolar; escuela secundaria; escuelas intermedias

O valor das respostas dos alunos em fóruns abertos: Uma análise natural das descrições de professores que foram mal avaliados

Resumo: Neste artigo, usamos o processamento de linguagem natural para revisar centenas de milhares de avaliações negativas de alunos de seus professores enviadas para o site RateMyTeacher.com. Nossa análise identificou várias questões levantadas pelos alunos ao avaliar mal os professores, o que contribui para a literatura que define "professores ruins" da perspectiva do aluno. Também identificamos a linguagem que os alunos usam para descrever essas questões e notamos uma distinção clara entre a linguagem usada para lidar com queixas relacionadas ao ensino e o comportamento percebido como impróprio para o ensino. Defendemos que os fóruns digitais podem ser ferramentas valiosas para as escolas e concluímos com sugestões e exemplos do tipo de políticas que podem ser derivadas deste tipo de análise de um fórum digital para alunos.

Keywords: avaliação do aluno sobre o desempenho do professor; processamento de linguagem natural; feedback (resposta); política escolar; ensino médio; escola secundária

The Value of Student Feedback in Open Forums: A Natural Language Analysis of Descriptions of Poorly Rated Teachers

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 different methods. These evaluations can be both formal such as those used by administrators, while others can be deeply personal, communal, and not shared, such as those done by students. Certainly, informal evaluations focus on different criteria, record qualities or accomplishments which are more difficult to measure, and typically 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 (Raufelder et al., 2016). Formal evaluations often put less of a premium on these characteristics. Despite challenges, student evaluations 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 (Polikoff, 2015).

In fact, several decades of studies 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 & Colianno, 2017). Yet, research on what they uncover remains scant (Polikoff, 2015), possibly because the surveys and interviews have been conducted with relatively small samples. Despite the inherent concerns about the reliability and validity of students' expressed judgments, there is value in understanding and using feedback from such evaluations as complements to the more formal criteria used for assessing practicing teachers. Student evaluations may paint a more complex—and more meaningful—portrait of a given teacher when combined with established evaluation practices. Evaluations of teachers by students can be useful for the development of small policy decisions at the local level. "Big P" policy (Berliner, 2019) may be concerned with national testing, school district funding, and integration by race and class. "Small p" policy 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—based on large numbers of written reviews by students who described their teachers as "bad". Policies emanating from such data may include recommendations for professional development for teachers whose students consistently and repeatedly report the same unwelcome behaviors, or such teachers might require closer monitoring of their classrooms, or even counseling for anger management.

Questions Driving This Study

1. How do students describe teachers they rate poorly on the RateMyTeacher.com website?
2. What common characteristics arise from these descriptions of teachers' rated poorly?
3. How can knowledge of these characteristics drive school and district (small "p") policy?

The study that follows aims to answer these questions through an analysis of the student comments that accompany low numerical ratings of teachers on the RateMyTeacher.com website. Our goal was to organize and analyze the comments submitted, mostly, by students to describe teachers they judged to be bad. We began by collecting more than 4.8 million publicly available student comments and ratings from the 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 perceived to be characteristic of teachers rated poorly. We end with some small "p" policy recommendations that we believe can lead to school improvement.

Our team included three educators, two with almost a 100 years' experience combined, who know that the issue of "bad" teachers is ever present in the talk about teachers and, often, in the policies made by schools and oversight boards. Discussions among colleagues, parents and students focus on proven or alleged accusations, on test scores, and on personal characteristics of teachers. Through this study we found out that students as informants offered common sense descriptions. Our analysis is a useful illustration of the methods schools can use to gain insight from forums that include students.

Conceptual Framework

Definitions of Good and Bad Teachers

There is, of course, a great deal of scholarship on the characteristics of *good* teachers, and such research leads to some general consensus of these characteristics. Gorham (1987, p. 3), using qualities of a good teacher as described by sixth 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 includes behavioral characteristics like not yelling, as well as more specific responsibilities of the teacher to intervene when necessary. Goodwin and Oyler (2008) argue that the most important quality any teacher must have is content knowledge as well as what they describe broadly as 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 light on our findings. Raufelder et al. (2016) had a similar goal to ours, but that research 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, which align closely with the findings we present here. Apparently, students' ratings of perceived bad teachers, as is true of good teachers, have common characteristics. From their interview data three prominent characteristics of bad teachers 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 assessment 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 by 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.

Similarly, Chang-Kredl and Cloannino (2017) examined the image of teachers in the public sphere—movies, television, and on the web. They analyzed 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. Demographic or other descriptive 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.

In each study discussed above, 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. We found something similar in our analysis; overall, students professed a desire to learn, and were frustrated when they were unable to do so because of perceived limitations or failures by their teachers. In the student comments summarized above, and our own findings, it becomes readily apparent that students really do want to learn and they do care about their achievements.

Table 1

Coding of Reddit Comments and Descriptions About One's "Worst" Teacher

Inductive coding for worst teachers.

#	Code	Examples	Total
Teaching/Professional Qualities			
1	Unintelligent, dull, boring, incompetent, unqualified	Stupid, no knowledge, boring, monotone, close-minded, unintelligent, unqualified, can't teach, useless, unable to teach, incapable, incomprehensible, thick accent	38
2	Poor judgment	Wants to save the black and brown students, insensitive Holocaust games, easily manipulated	13
3	No effort	Lazy, didn't even try, no effort, negligent, neglectful, apathetic, no investment, watched movies, didn't give a shit, stopped showing up, should retire	28
4	Unfair expectations, inconsistent	Hard marker, expectations not appropriate, punitive, inflexible, inconsistent, accused student of cheating (wrongly), grades based on drawings,	35
5	Plays favourites, biased	Biased/racist/sexist, religious & indoctrinating, has favourites, close-minded	26
Personal Qualities			
6	Bizarre, odd, immature, mentally questionable, inappropriate	Nut case, erratic, moody, odd behaviours (non-sexual, e.g., smells farts), bizarre, went nuts, inappropriate comments, over-sharing, wants to be a teen again, no social skills, alcoholic	29
7	Bad-tempered, nasty temperament	Mean, angry, yells, strict, vile	22
8	Superior, condescending	Cruel, snobbish, egotistic, self-centered, arrogant, vile, smug, haughty, pompous, power-mongering, requires own textbook	17
9	Physically unappealing	Fat, obese	7
Implications for Students			
10	Didn't learn anything (focus on student outcome)	Learned nothing	5
11	No self-confidence (lasting negative impact)	Became self-deprecating, felt stupid for years, lost confidence	4
Relational Qualities			
12	Discouraging, doesn't care about or support students	Discouraging, wouldn't help, insensitive, hypocritical, dismissive, inconsiderate, thoughtless, won't let child pee, doesn't look out for child, negligent in case of emergency, ignores sick complaint	18
13	Verbally abusive	Malicious, wants to terrorize the children, scares children, bullies, encourages/allows children to bully other children, cruel, humiliating, tells children Santa isn't real, embarrasses student	22
14	Physically abusive	Hits, locks child in closet, puts child in fridge, duct tapes kid's mouth shut, held student against wall, kicked kid's ass, sells kids drugs	24
15	Sexual advances or suggestive	Pedophile, sexual, creep, porn to class	12

Note: From Chang-Kredl and Cloannino, 2017, p. 48

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 related question for this study is "How many teachers are we talking about?" In fact, the numbers of teachers who are seen as "bad," as 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 the 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 et al., 2010; see also Konstantopoulos, 2014; and Popham et al., 2014). No lasting "P" or "p" educational policies have been derived from these studies.

Consistent unreliability in the measures used to assess teachers, strongly limits validity. Test-based accountability systems, especially VAM-based accountability, ought to be avoided (Pivovarova et al., 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.

Nevertheless, there are some observational instruments that are commonly used and are found by many educators to be useful in providing feedback to teachers (e.g. Danielson, 2008; Pianta et al., 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 et al., 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 et al., 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.

Other Modalities of 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) 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 (Berliner, 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 (SETs). 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 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, they are all limited. None of them adequately describe the overall quality of "teacher." In a sense, these various methods of teacher evaluation provide 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 SETs. Others, however, especially those in higher education (Lawrence, 2018) argue that the information obtained from SETs is invalid, and does more harm than good. We looked closer at this issue and found that student evaluations of teachers, 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 own 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 we study, and actively addressed.

Numeric vs. Descriptive Student Feedback

Stark and Freishtat (2014) studied the numerical rating system based on student evaluations of professors at UC Berkeley. They found, among other problems, unacceptable differences in response rates by students per faculty member. Thus, the confidence bands around any numerical rating were 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)." In what follows, this is exactly what we do.

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, particularly against female instructors. Ratings on such attributes as Caring, Enthusiasm, Feedback, and the like all showed this gender bias. However, Freishtat (2016) 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)."As such, SETs can provide valuable insight into the day-to-day experiences by students of their experiences with teachers and, in turn, can complement other forms of teacher evaluations to paint a broader "portrait" of a given teacher.

Because we are not using the information from SETs to influence decisions about teachers that demand greater trust in the validity of 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 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, in increments of ½ stars. 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 16, 2018.

Reviews were attached to specific teachers and schools and were collected from six English speaking countries (United States, Canada, United Kingdom, Australia, New Zealand, and Ireland). We limited our data set to only those originating in the US. We collected 4,884,479 reviews from the US. 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.

Natural Language Processing

Since natural language processing became popular in the 1980s 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 identify and 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 et al., 2013). The words that most frequently appear in the cluster can be used to characterize it and these can be interpreted as topics (Papadimitriou et al., 2000).

We used a latent dirichlet allocation (LDA) model to generate clusters, from which we derived topics. 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 et al., 2005) but instead, we chose to use the concept of coherence to estimate an optimal number of K-categories for subsequent analysis.

Our first step in preparing the data was processing the text 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 et al., 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 et al., 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 (e.g. "as_soon_as_possible"). Cleaning the data was an iterative process. One other experimental parameter was a lower bound on review length. We found that setting a minimum number of characters dramatically increased the probability that any given review contained non-trivial information, but higher minimums also decreased the size of the dataset. We experimented with 80, 100, 125, 150, and 175-character thresholds. Ultimately, we found 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.

Model Selection using 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 (Mimno et al., 2011; Stevens et al., 2012).

In our analysis, we trained models for different numbers of topics (from 3-32 in steps of 1 for a total of 29 different models) and used Gensim to estimate coherence scores for each topic generated by each model. We then calculated the average coherence score for each model and plotted these against the number of topics to create what is known as an elbow plot (See Figure 1). We selected our model based on two criteria (Syed & Spruit, 2017). First, average coherence usually increases with the number of topics up to a certain max, selecting the number at which the rate growth of average coherence drops (where the elbow is sharpest) usually yields the most interpretable results. Second, while a model with a higher number of topics may result in a high coherence score and very specific subtopics, if the same key-words appear in multiple topics this suggests the presence of redundant topics which can make thematic analysis more difficult (see Evaluation of the Language Model section below).

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.

Thematic Analysis of Results

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 and mostly appeared in the topic and nowhere else. We were able to use our model to predict the probability that any given review belonged to a topic (which was interpreted as the presence of a given topic in that review) and assigned each review to its "most representative" topic. This process gave us an idea of the "share" of each topic.

Afterwards, we sorted the reviews for each topic from most to least representative and drew a sample of more than100 sample reviews from the top half of the distribution. We coded

the sample reviews based on shared, meaning-based patterns or themes. We used a reflexive approach to thematic analysis meaning that the codes were refined in an iterative process (Braun et al., 2019). The key-words yielded by the LDA model framed this analysis by providing expected linguistic patterns.

Finally, we articulated a description of each topic and visualized the results of our linguistic model using the pyLDAviz module which illustrated the topic distributions by plotting them in a 2-dimentional space that illustrated both their share of the corpus and linguistic similarity (Sievert & Shirley, 2014). This provided us a more intuitive view of the relationships between the language and themes which we used to critique the patterns we highlighted in our coding process.

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. We did not know if a student was lashing out against a teacher for a bad grade. The common biases against instructors who are female or persons of color, discussed above, were surely present, as well.

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, Angel, 2009; Burdick, 2009; Burdick & Sandlin, 2010). Online forums 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.

In all, we felt safe to assume that the dataset represented a diverse array of motives, contexts, and incentives for submitting negative reviews about teachers. Most teachers in our sample received mixed (mostly positive) reviews, while a small minority received mostly bad reviews (see Description of Score Data in Appendix). 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 five 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 reviewer. In fact, the results of the analysis increased our confidence that this was the case.

Ultimately, any conclusions drawn from this study can only add to the literature around the issues related to students' capability of observing and communicating information about their teachers. We also believe this study is another piece of evidence that supports the value of student feedback. And it illustrates the usefulness of an online digital forum. Nevertheless, this is not a formal evaluation of teachers or teaching and this method of obtaining information should not be used to dictate promotion and salary. We hope this serves as a primer for future research into innovative methods that incorporate student feedback to make classrooms and schools better environments for students.

Results

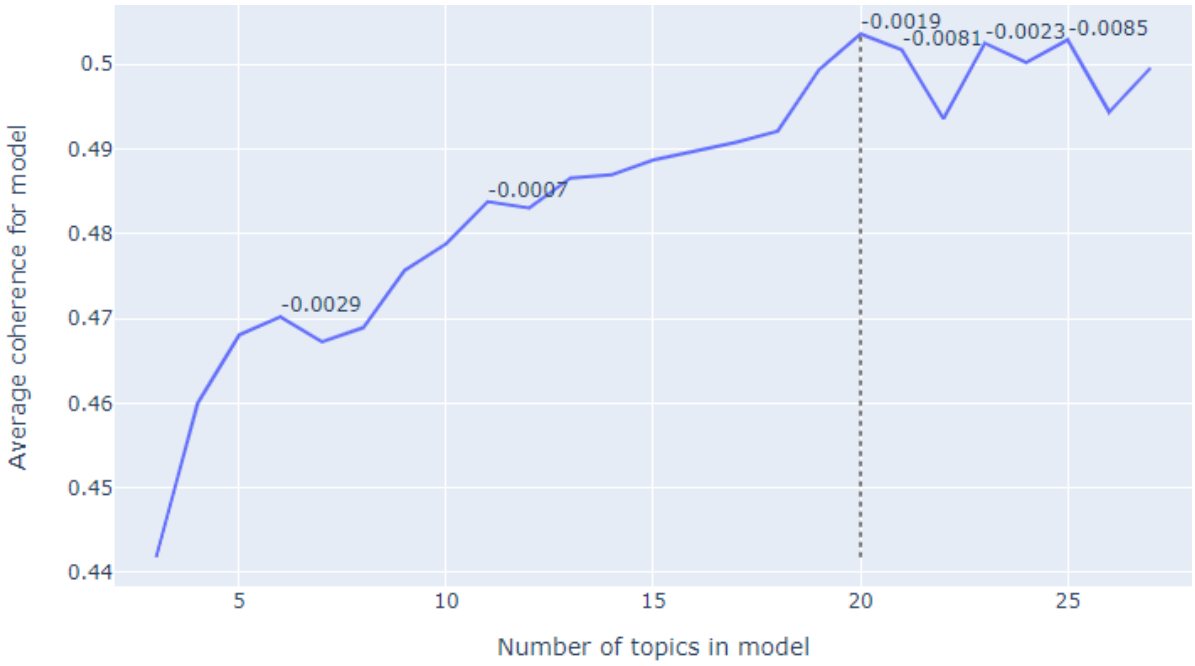
The results revealed one data cleaning procedure that consistently resulted in higher coherence scores. This procedure used both a lemmatizer (removes pluralization, conjugation) and stemmer (changes words to their root). Given the informal nature of our dataset, it made sense that the strictest cleaning method resulted in more concise and coherent topics. The 100-

character restrictions described above resulted in considerably higher average coherence scores. Of the 359,387 "worst" reviews, 211,224 met the 100-character threshold.

Figure 1 shows the elbow plots for the 100-character minimum models, the average of average coherence scores can be seen on the left. The moments where the slope from one topic model to the next were negative are highlighted in the graph. The models with more than 20 topics were reviewed and discarded based on our second model-selection criteria. The 20-topic model, which yielded an average coherence score of 0.503, was selected. This was lower than the 0.52 benchmark for unstructured datasets in Rosner, 2014. We describe the implications of this in the discussion.

Figure 1

Elbow Plot



Topic Descriptions

In this section we present the results of our thematic analysis. We iterated the coding process three times before arriving at a final list of themes and sub-themes that we used to code all the topics from the language model. We provide the final list of codes below in Table 2. Summary results of the coding process can be found in the first table of the Appendix. Table 3 displays each topic's coherence score, key-words, and the percentage of reviews in which they were present. A selection of sample reviews can be found in the second table of the Appendix, we highly recommend reading these along with the topic descriptions. The full body of sample reviews and results of the analysis can be found on the GitHub page we set up at https://github.com/losDaniel/Student-Voices/. We skip topics 5, 17, and 19 because we were unable to find consistent themes for these, this is discussed at the end of this section. For clarity, in our topic descriptions themes and sub-themes are *italicized* while key-words are shown in "quotations."

Table 2

Themes and Sub-Themes Used to Code Sample Reviews

Personality / Rapport	**Bad / Not Teaching**
bad relationship	no learning / unprepared
parent relationship	boring
biased	busy work
picks on kids	waste / pointless
picks favorites	cannot explain / unclear
does not care / lazy	confusing
angry / short temper	rushed
hates questions	unhelpful
mean / no respect	ignores kids / questions
scary	discouraging
yells	off-topic
unprofessional	by the book / ppt
abrasive /frustrating	**Workload**
cool / nice	unrealistic expectations
treat like kids	too much work
strict	hard / unfair (grader)
Duties	bad feedback
lack expertise	unclear direction
makes mistakes	study guides
distracted	misaligned
does other things	easy a
chaos in classroom	
disorganized	**Specific Subjects / Jobs**
slow / no grades	
no follow-through	

Table 3

Topics Returned by the 20-Topic LDA Model Sorted by Coherence Score

#	CS	Key-words (changed from root-form to the word that most appeared; e.g. spel to spell)	% of Reviews
1	0.611	read, use, book, write, note, science, compute, copy, powerpoint, word, board, page, internet, us, textbook, power_point, lab, biology, slide, tech, print, technology, online, picture, number, spelled, quote, code, text_book, calculus	2.90%
2	0.601	grade, give, assign, homework, project, work, point, paper, essay, hw, lose, check, ton, due, turn, busy_work, test, gave, night, never, collect, late, time, due_date, take, load, rubric, zero, week, hand	7.30%
3	0.597	ask, question, answer, wrong, ask_question, help, tell, say, answer_question, will, something, look, problem, right, ignore, whenever, never, shel, figure, raise_hand, correct, always, even, repeat, get_mad, know, clarify, someone, everytime, reply	5.10%
4	0.547	got, last, said, told, went, gave, name, first, came, glad, remember, thank, took, sub, left, counselor, knew, friend, week, still, year_ago, 8th_grade, thought, class, kept, gone, 6th_grade, 7th_grade, lunch, cale	5.00%
5	0.543	been, college, rate, daughter, educ, allow, year, speech, professor, program, course, posit, graduate, proceed, receive, experience, guidance, apply, school, attend, given, longer, cause, state, sinc, star, disappoint, neg, past, interpret	3.30%
6	0.537	favorite, play, pick, pick_favorite, music, favor, unfair, band, chose, one, other, obvious, fair, popular, everyone_else, sing, song, drama, orchestra, choir, girl, certain, pet, definite, least, people, talent, concert, fav, boy	3%
7	0.534	test, fail, quiz, anything, study, learn, exam, taught, pa, teach, prepare, class, never, review, give, chapter, all, impose, worksheet, memor, unit, study_guide, cover, read_textbook, single_thing, packet, material, final, will, regent	7.10%
8	0.508	nothing, class, learn, ap, honor, absolute, english, waste, waste_time, spent, classes, avoid, taken, physic, drop, entire, freshman, take, regular, cost, dread, level, semester, time, regret, sophomore, free, junior, dropped, pointless	5.10%
9	0.498	teach, knowledge, subject, method, style, skill, rather, often, lack, material, por, student, cannot, interest, ability, effect, topic, curiculum, inform, rely, disorganize, seem, simply, actually, improv, focus, little, instruct, actual, subject_matter	5.30%
10	0.486	student, child, rude, respect, parent, care, disrespect, issue, condescend, unprofesion, help, toward, teacher, arog, concern, attitude, insult, deal, extreme, behavior, inappropriate, demand, individual, encourage, approach, manners, comun, unwilling	6.00%

Table 3 (Cont'd.)

Topics Returned by the 20-Topic LDA Model Sorted by Coherence Score

#	CS	Key-words (changed from root-form to the word that most appeared; e.g. spel to spell)	% of Reviews
11	0.483	talk, joke, time, sit, life, story, spend, watch, minute, laugh, period, hear, whole, phone, class, movie, half, al, eat, go, ramble, tele, quiet, random, tell, listen, shut, min, hour, rom	5.90%
12	0.479	opinion, history, art, agree, mind, world, term, polit, draw, view, push, disagree, form, belief, religion, short, rock, know, character, smith, creative, brown, health, memorize, obnoxious, debate, judge, discus, bias, artist	2.00%
13	0.477	nice, person, god, great, teacher, luck, lady, side, really, sweet, bad, best, latin, greatest, over, nicest, awesome, however, intent, wise, sense_humor, friendly, kind, although, sometimes, heart, horrible, may, otherwise, teach	4.20%
14	0.475	explain, understand, math, hard, confuse, well, help, tutor, fast, dificult, lesson, geometry, teach, algebra, problem, clearly, concept, make, material, tri, harder, extra, sense, class, go_fast, struggle, rush, complicated, explain, cannot	6.40%
15	0.470	bore, easy, pretty, fun, really, col, class, funny, super, fall_asleep, hard, sleep, weird, alright, sometimes, chill, ok, pay_attention, monotone, stay_awake, fell_asleep, lame, anything, learn, gue, monotone_voic, easiest, wana, stuff, entertain	4.30%
16	0.469	worst, teacher, horrible, probably, far, terrible, aw, possible, life, fire, worse, switch, teach, bad, absolute, entire_life, honestly, met, math, anything, know, even, encounter, science, dumbest, world, never, without_doubt, quite_posibly_worst, all	4.90%
17	0.460	school, run, coach, principle, around, gym, head, sport, office, team, money, dance, pe, rule, away, job, athlete, rune, football, pleas, administr, care, kid, library, game, hit, build, dean, back, throw	4.50%
18	0.441	like, alway, felt, act, control, kid, seem, treat, class, make, people, bad, felt, really, cannot_control, old, enjoy, tri, smarter, kindergarten, bad_mood, nobody, moody, sound, babies, fun, dumb, look, want, late	6.20%
19	0.434	need, know, cannot, teach, spanish, think, french, speak, learn, language, stop, everything, maybe, expect, english, better, already, retire, dosent, anything, new, actually, speak_spanish, stuff, german, even, spell, tri, say, help	6.50%
20	0.421	hate, mean, think, yell, annoy, reason, omg, ugh, cannot_stand, pl, voice, people, everyone, say, wat, yell, meanest, really, evil, always, ah, detent, scream, strict, scare, cuz, ok, scary, plain	5.00%

Topic 1 (coherence score: 0.611, explicit in 2.9% of reviews)

Almost 70% of sample reviews complained about the teacher not teaching. Of these, a majority were coded under *by the book / ppt* sub-theme which described teachers whose classroom activities primarily consisted of note-taking, showing power-points, and reading from the textbook or who allowed absolutely no deviation from the class material. Key-words such as "powerpoint," "textbook," "note," and "board" gave us a good idea of what to expect. These reviews also described teachers as *unhelpful* or *hard* and *no learning* in the classroom.

Topic 2 (coherence score: 0.601, explicit in 7.3% of reviews)

Over three quarters of reviews in this topic were coded under the *workload* theme. The core complaint revolved around teachers giving *too much work / too little time*, *not enough direction* and being *unfair graders*. These were very often associated with descriptions of teachers being *late or not returning grades* or assigning *busy work* and providing *unclear explanations* of the subject. The key-words made it clear that students would be complaining about grades, but the thematic analysis revealed that key-words like "lose," "ton," "check," and "rubric" were pointing at causal descriptions of teachers failing to fulfill their *duties* or more generally, *not teaching*.

Topic 3 (coherence score: 0.597, explicit in 5.1% of reviews)

While 70% of sample reviews were coded under *not teaching*, and more specifically *ignoring students and their questions*, over 40% were also coded as expressing issues with teachers' *personality*, which described as hostility, specifically towards questions. This topic concentrated reviews that described environments discouraged questions because students felt the teacher was 1) not capable of providing good answers, 2) would ignore them, or 3) would make them feel bad about asking a question. Sample reviews describing the latter were often heartbreaking.

Topic 4 (coherence score: 0.547, explicit in 5.0% of reviews)

Most comments under this topic highlighted issues with teachers' *personalities* in distinctive ways that revolved around *unprofessionalism*, such as forgetting students' names, using inappropriate language, or teachers *yelling*, being *strict* or *abrasive* and making students feel bad. The key-words led us to expect reviews from parents or former students, we found this to be the case for many but not a majority of sample reviews which helped clarify why a topic with distinctive language covered relatively inconsistent themes.

Topic 6 (coherence score: 0.537, explicit in 3.0% of reviews)

More than 90% of sample reviews from this topic were coded as *personality* issues, 75% were specifically addressed picking favorites. What was remarkably interesting about this is that so few reviews tied this to any other complaint. In other words, while students expressed frustration and criticized the teacher's behavior, they rarely tied this to unfair grading, for example. They key-words revealed what we found in the coding process which was that music, band, or drama appeared more often here than in other topics (though only in a minority of sample reviews).

Topic 7 (coherence score: 0.534, explicit in 7.1% of reviews)

The key-words for this topic like "test," "fail," "quiz," "memorize," and "study guide" suggested complaints surrounding evaluation. In fact, this topic was split between two general themes, about 50% were coded as *bad / not teaching* and 60% were coded under the theme *workload* with 40% under the sub-theme of *misaligned* classwork and evaluation. These two overlapped over a third of the time either one of them appeared. Generally, sample reviews showed students' concern over understanding the structure of the class, inspiring descriptions of teachers as *disorganized*, *lacking expertise*, or being *nice* but a *bad teacher*.

Topic 8 (coherence score: 0.508, explicit in 5.1% of reviews)

Another subset of *not teaching* (72%) sample reviews from this topic explicitly mentioned *no learning* about 50% of the time. One of the more common descriptions was *waste of time*, a sub-theme we expected given the key-words "waste," "waste time," "pointless," "learn," and "nothing." Interestingly, few reviews highlighted an issue with teachers' personalities or performance of duties. We felt this topic demonstrated the value students placed on genuinely learning in school.

Topic 9 (coherence score: 0.498, explicit in 5.3% of reviews)

Over 60% of sample reviews for this topic were coded as *not teaching*, these overlapped considerably with *personality* issues which were found in just over 40% of sample reviews. The key-words "knowledge," "skill," "method," "disorganize" and "subject matter" framed an expectation that teacher's ability was being called into question. The sample reviews provided clarity, they described teachers as *lacking in expertise* or being *disorganized* and often providing *unclear explanations*. Other descriptions of bad teaching such as being *confusing*, *unmotivating*, *by the book*, and *unhelpful* were common.

Topic 10 (coherence score: 0.486, explicit in 6.0% of reviews)

Ninety-four percent of sample reviews were about teachers' personality. Above 40% described teachers as *mean or rude*, other descriptions include *unprofessional*, *uncaring* about students or teaching, or *getting angry*. Noticeably, about 15% of samples had been submitted by parents. The key-words such as "parent," "child," "rude," "respect," "disrespect," "care," "condescend," "inappropriate," and "insult" were very aligned with what we found in coding. These included little mention of teaching ability.

Topic 11 (coherence score: 0.483, explicit in 5.9% of reviews)

About 60% of sample reviews in this topic were coded under failure of *duties* and *not teaching*. They described teachers who were *off-topic* and *did other things* besides teaching in the classroom. About 40% of sample reviews addressed those sub-themes directly, other reviews complained about teachers' *abrasive* or *unprofessional personalities*. Key-words for the topic were in line with description of teachers spending entire class periods talking about themselves, "rambling," or doing things like being on their computer or phone.

Topic 12 (coherence score: 0.479, explicit in 2.0% of reviews)

Fifty-five percent of sample reviews addressed biased teachers. The key-words "opinion," "history," "political," "view," "belief," "disagree," and "debate" eluded to what we found in the coding process. The students who wrote these reviews rated their teachers poorly because they felt as though they were close-minded or radical. Sometimes these complaints seemed legitimate, though mostly they told us much more about the prejudices held by students themselves. Unsurprisingly, these addressed history or social science teachers more often relative to other topics.

Topic 13 (coherence score: 0.477, explicit in 4.2% of reviews)

Nice person, bad teacher. Seventy percent of sample reviews made positive remarks about their subjects being "nice," "friendly" or "sweet" but bad teachers. This juxtaposition was so frequent that exception language like "however" and "although" were highlighted among the key-words.

Topic 14 (coherence score: 0.475, explicit in 6.4% of reviews)

A remarkable 83% of sample reviews coded for this topic fell under the *bad teaching* theme. The sub-theme *cannot explain* was present in 60% of sample reviews. Teachers were also

described as *unhelpful*, *rushed*, and leaving students *confused*. Key-words such as "explain," "understand," "confuse," "fast," "clearly," "rush," and "complicated" set clear expectations while the key-words "math," "geometry," and "algebra" showed that mathematics courses appeared in this topic more often relative to other topics.

Topic 15 (coherence score: 0.469, explicit in 4.9% of reviews)

This topic also revolved around *bad or not teaching* (70%). The sub-theme *boring* appeared in nearly 50% of reviews, *easy A* in 30%, and *no learning* or feeling *unprepared* in over 20%. Teachers were described as a "bore," "easy," and having "monotone voices" but students also expressed difficulty staying away as suggested by the key-words "fall asleep" and "stay awake." Over 40% of sample reviews commented on teachers' *personality* though these were split between negative traits and teachers being "cool" or "funny" but boring teachers.

Topic 16 (coherence score 0.469, explicit in 4.9% of reviews)

No sub-themes were particularly common in this topic and no general theme was present in more than 50% of reviews. The most consistent description in this topic was "worst" or "horrible" teacher. If explanations were offered, they covered a wide variety of sub-themes. These were mostly declarative statements which often included key phrases like "entire life," without a doubt," and "quite possibly the worst."

Topic 18 (coherence score 0.441, explicit in 6.2% of reviews)

Teachers' *personality* was a theme in 77% of the sample reviews, with descriptions split between the *frustrating, treats us like kids*, *picks on kids*, *does not care*, and *mean*. About 20% of the sample reviews also described *chaotic classrooms*. The key-words were not very clear at the outset, but specific words like "treat," "control," "baby," "moody" and "bad mood" stood out.

Topic 20 (coherence score 0.421, explicit in 5.0% of reviews)

The key-words "hate," "mean," "yell," "scream," "annoy," "cannot stand," and "evil" framed this last topic. About 80% of sample reviews were coded as *personality* issues that revolved around aggression like being *mean or disrespectful*, *short-tempered*, *scary* or *strict* and often *yelling* or screaming.

Omitted Topics 5, 17 & 19

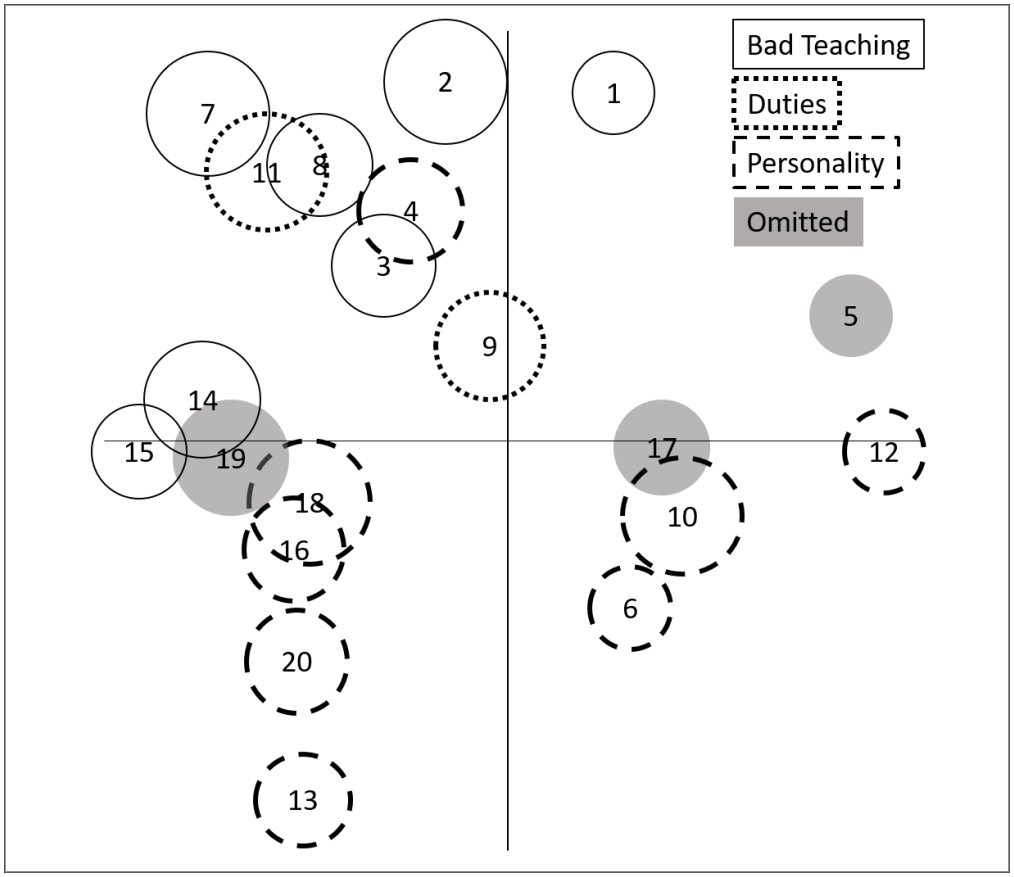
In each iteration of the coding process, we were unable to identify consistent or relevant themes or sub-themes for these topics. Topic 5 which yielded a relatively high coherence score collected reviews with distinctive language, but no consistent theme. Topic 19 was also inconsistent. Topic 17 was only slightly more consistent, however it addressed administrators and staff, often in a non-teaching capacity which we deemed irrelevant to our study.

Language Plotted in Two Dimensions

The LDA model grouped topics based on word frequency but also the order of the words in sentences (a proxy for grammar). It kept track of the unique words in the corpus, such that each word was a variable in the analysis. Topics were generated by randomly picking coordinates in the resulting vector space and iteratively grouping the closest data points together. These shifted the coordinates until the groupings minimized the predicted error of which point belonged in which group. We then plotted these groupings as circles whose size was proportional to the number of data points in the cluster. We color coded this graph to highlight the general themes that were most present in each topic (*Workload* was grouped into *Bad Teaching*) which helped us examine how well linguistic patterns adhered to the meaning-based patterns we identified.

Figure 2

Topics Plotted in Two Dimensions. Frequency by Size, Theme by Styling



Discussion

Evaluation of the Language Model

The average coherence score for the 20-topic model was 0.503 which was low compared to the 0.52 benchmark for very unstructured datasets set out in Röder et al (2015). Coherence is a relative measure based on consistent and distinctive language. It may help select a model and compare topics to one another but is not necessarily indicative of whether a model is useful or should be discarded. Sample reviews for topics with lower coherence scores generally included more reviews that had little to do with the themes highlighted. Thus, we understood coherence as a measure of "vagueness" since the main themes did not disappear but simply became less frequent or central in reviews. For our results, this implied that the percentage of reviews attributed to each topic was likely overestimated for less coherent topics.

The 2-dimensional topic plot gave us an idea of how language and meaning had been parsed out by our model. There was a clear distinction in the language students used to describe teachers' personalities vs. the language they used to describe their teaching ability or classwork. It was also surprising but reassuring to find that the topics that did not have consistent or relevant themes (Topics 5, 17 & 19) seemed to define the border between these two types of expression.

Small "p" Policies Based on Student Descriptions

Our analyses suggest that students used different language when they attempted to distinguish between teachers interpersonal and academic skills. We make the case that this value students place in their own learning lends itself to the usefulness of a digital forum.

Behind the Curtain of Unfair Grading

Topic 2 is illustrative of the type of policy that can be derived from these student descriptions of teachers they rated poorly. Revealed here is 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 students 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 regularly describe a particular 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. These are all potentially remediable behaviors if signaled to those with responsibility for the supervision of teachers.

Creating Nurturing Environments

Topic 3 revealed that when students' main complaint was about asking questions in the classroom, it was because their teachers created an unwelcoming environment. While many students felt their teachers could not answer their questions adeptly, others felt bullied or purposefully ignored by their teachers. Critically, this was one of the only issues where students complained about both teaching and teachers' personalities. Because questions are among the most productive and essential forms of teacher-student interaction, behaviors that discourage them can spiral into other problems that may only be perceptible to the student. Those responsible for the supervision of teachers should view complaints about not being able to ask questions as meaningful "red flags."

Contribution to the Literature on "Bad Teachers"

Overall, we found some overlap with themes described by Raufelder et al. (2016) and to a lesser extent with Chang-Kredl and Cloannino (2017). However, our findings add considerable detail to how we might choose to articulate definitions for bad teaching or behaviors inappropriate for teachers from the point of view of students. Topics that defined bad teaching (Topics 1, 2, 3, 7, 8, 14 & 15) accounted for 38.2% of bad reviews. Topics that addressed teachers' personalities (Topics 4, 6, 10, 12, 13, 16, 18 & 20) accounted for 36.3% of bad reviews. Topics that fell between these described teachers as unwilling or unable to fulfill their duties (Topics 9 & 11, a bit of Topic 2) which accounted for at least 11.2% of bad reviews while another 14% were grouped into topics that were omitted from the thematic analysis (Topics 5, 17 & 19) because they were too broad and inconsistent or irrelevant.

Bad Teaching

Bad teaching was the central theme for a number of our topics. As highlighted above, Topic 2 (present in 7.1% of all bad reviews) focused on incomplete direction and feedback while Topic 3 (5.1%) described environments that discouraged asking questions. Topic 1 (2.9%) expressed students' frustration towards teachers that adhered religiously to the text-book or were dependent on power-points and notes to do their teaching for them. Raufelder described this as *teacher-centered instruction*.

Students value for learning was clearly expressed in the phrase "waste of time" highlighted in Topic 8 (5.1%). This topic collected comments where students rated teachers poorly because they felt they had not learned anything. Students also rated their teachers poorly when they felt they had not been taught what appeared on their evaluations as described in Topic 7 (7.1%). Some of these students criticized teachers for being disorganized, others felt they could not understand teachers' explanations.

Bad explanations were a critical issue for students. They dominated 6.4% of bad reviews under Topic 14. Language around being left confused or teachers being unclear was common

across all *bad teaching* topics, which explained its low coherence score. We point this out because the thematic analysis for this topic was very consistent despite its low score. Topic 15 (4.9%) showed that students like to be challenged, or at least kept awake during class, another dimension of effective explanation. Topics 14 & 15 coincided with Raufelder's sub-theme *incomprehensible teaching.*

For its part, Topic 13 (4.2%) related to teachers' amenable personality, but recognized, as well, their inability to teach. Students' comments often included many of the teaching sub-themes discussed above. Topic 11 (5.9%) also addressed poor teaching, however these focused on teachers who were constantly off-topic or literally not teaching and doing something else. These teachers were described as disinterested in teaching, which coincided with the *no effort* theme highlighted by Chang-Kredl and Cloannino (2017).

Suggestions that a person was not right for or qualified for the job of teaching was also a central theme in Topic 9 (5.3%). Here, reviews commented on mistakes made by teachers or attributed bad explanations to a lack of knowledge on the subject or how to teach it. Next to Topics 3 and 13, Topic 9 was the only other one of these topics that also mentioned teacher's personality frequently. The finding that students felt antipathy when they described their teachers as lacking knowledge echoed Raufelders categorization of a lack of knowledge as a driver of antipathy under his theme of *student-teacher relationship*.

Behaviors Unfit for Teachers

Interestingly, in our analysis, most of the topics that expressed strong antipathy or attacked teachers' character had lower coherence scores. Topics 10, 18, and 20 (6.0%, 6.2%, and 5.0% respectively) all described what Raufelder called *relational aggression*. Reviews described teachers who became angry, raised their voices, showed little respect for or actively picked on some of their students. Topic 4 (5.0%) also provided anecdotes along these lines.

Another barrier to the student-teacher relationship was described in Topic 12 (2.0%) where students and teachers often disagreed about social or political issues. The most obviously biased student comments appeared under this topic. Topic 16 (4.9%) also illustrated a poor student-teacher relationship, however, it only did so through student remarking how some teachers were "the worst."

The only other behavior that was highlighted when students rated their teachers poorly was favoritism, under Topic 6 (3.0%). Sometimes this involved aggression but often it meant what Raufelder described as *injustice*. There was notable dislike of teachers that showed favoritism through their demeanor, actions, or grading. Though, in some cases students may have been being a bit dramatic.

Conclusions

The most outstanding insight gained was about how generative students' evaluative comments could be. We found that students' evaluative language frequently split relatively evenly between criticism of teachers' teaching ability, and criticism of their personality. We found that most negative reviews which addressed teacher's behavior or personality used language that was less consistent and more difficult to categorize. On the other hand, reviews that discussed teachers' knowledge, ability, and commitment, appeared to be more thoughtful, sophisticated, and consistent than we had expected. This suggests that students can be critical and insightful evaluators of their teachers, *if they are asked*!

Evaluative commentary from those closest to classroom instruction (i.e. students themselves) can inform simple policy changes that improve teaching, affect student learning, and affect, as well, student and parent satisfaction with schooling. In this paper we provide two brief examples of such policies, derived from our results. We argue that students' evaluations of their life in classrooms, while surely sometimes hard to accept, are capable of promoting new policies

for in-service education; for the counseling and guidance of teachers; for union negotiation of the numbers of visits to classrooms, and by whom; for teacher grade level and subject matter assignments; and more.

Our findings also support the policies that establish online forums as a means to collect student feedback. Many issues highlighted in our data are undetectable in other forms of teacher evaluation. For example, teacher anger is rarely expressed when a supervisor is visiting a classroom. Unpreparedness for a lesson will not be apparent if a supervisors' classroom visit is known in advance. But students may see anger and unpreparedness frequently. Student evaluations also turn a spotlight on teachers who wasted students time, something many students resent.

In fact, among the most surprising values reflected in the topics we found is that students 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.

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 as a *complementary form* of evaluation. Without being determinant of promotion and salary, students' analyses appear quite capable of providing extraordinarily rich data about *their* lives in *their* 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 can use to improve their schools. Thus, we urge further investigations into using student feedback in formative ways to improve instruction, and possibly, over time, 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.* American Educational Research Association.

American Statistical Association (2014, April 8). *ASA Statement on Using Value-Added Models for Educational Assessment*. 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*. Routledge. https://doi.org/10.4324/9780203409909

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). https://doi.org/10.14507/epaa.v20n12.2012

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.

Berliner, D. C. (2018). Between Scylla and Charybdis: Reflections on and problems associated with the evaluation of teachers in an era of metrification. *Education Policy Analysis Archives*, *26*(54). https://doi.org/10.14507/epaa.26.3820

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*. 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*. 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*. 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. https://doi.org/10.1177/074193259001100609

Braun, V., Clarke, V., Hayfield, N., & Terry, G. (2019) Thematic Analysis. In: P. Liamputtong (Ed.), *Handbook of research methods in health social sciences*. Springer. https://doi.org/10.1007/978-981-10-5251-4_103

Butrymowicz, S. (2014, June 16). How many bad teachers are there? *The Hechinger Report*. 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.

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). 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). 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/2012/12/11/students-evaluating-teachers/

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

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

Freishstat, R. L. (2016). *Expert report on student evaluations of teaching (SET)*. 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., & Lavigne, A. L. (2018). *Looking in classrooms* (12th ed.). Pearson.

Goodwin, L. (2016). Defining teacher quality. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teacher education* (pp. 399-403). Routledge.

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

Gorham, J. K. (1987). *Sixth grade students' perceptions of good teachers.* (Report No. SP 034 604). Curry School of Education, University of Virginia (ERIC Document Reproduction Service No. ED 359 164)

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*. Springer.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores, The 14th William H. Angoff Memorial Lecture*. Educational Testing Service. 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.* Routledge.

Hattie, J. (February, 2015). High impact leadership. *Educational Leadership*, 36-40.

Hosgorur, T. (2015). 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*. Teachers College Press.

Jacobs, H. H. (2012). *Curriculum 21: Socrates fails teacher evaluation.* 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 from http://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.* 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). Lawrence Erlbaum.

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

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *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

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, *18*(23). http://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, F. (1978). School psychology graduate students' perceptions of effective and ineffective teachers. *College Student Journal, 12*(4), 360-363.

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 D. H. Gitomer & C. A. Bell (Eds.). *Handbook of research on teacher education*. Routledge. from https://www.emerald.com/insight/content/doi/10.1108/QAE-07-2014-0033/full/html#loginreload

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

Pivovarova, M., Amrein-Beardsley, A., & Broatch, J. (2016). Value-added models (VAMs): Caveat emptor. *Statistics and Public Policy*, *3*(1), 1-9. https://doi.org/10.1080/2330443X.2016.1164641

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

Popham, W. J., Berliner, D. C., Kingston, N. M., Fuhrman, S. H., Ladd, S. M., Charbonneau, J., & Chatterji, M. (2014). Can today's standardized achievement tests yield instructionally useful data? Challenges, promises and the state of the art. *Quality Assurance in Education, 22*(4), 303-318. https://doi.org/10.1108/QAE-07-2014-0033

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. 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*. 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). 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*. arXiv preprint arXiv:1403.6397

Scriven, M. (1994). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research, and Evaluation, 4*, Article 7. 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. https://doi.org/10.1007/BF00972261

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

Stark, P. B., & Freishtat, R. (2014, 29 September). An evaluation of course evaluations. *ScienceOpen Research*. 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 (pp. 952-961). Association for Computational Linguistics.

Stronge, J. H. (2007). *Qualities of effective teachers* (2nd ed.). 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). http://dx.doi.org/10.14507/epaa.v22.1590

Syed, S., & Spruit, M. (2017, October). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)* (pp. 165-174). IEEE.

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* (pp. 1385– 1392). 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.

About the Authors

Carlos Valcarcel
Arizona State University
losdaniel@berkeley.edu
Carlos Valcarcel is currently a graduate student at the University of California, Berkeley but formerly held a position as a Research Analyst at the Center for the Art and Science of Teaching (CAST) at Arizona State University (ASU) under Dr. Mari Koerner. His research focus began in environmental economics and later shifted toward education and educational policy. He has published articles in the academic journal *Applied Economics Letters* and worked on numerous grant-funded research projects in his role as Sr. Researcher at the Digital Teaching and Learning Action Lab at ASU. He has also published articles on data-science, programming and cognitive-gaming for dyslexia. His current research interests focus on community organization and technology transfer in social and economic development contexts.

Jeffrey B. Holmes
Arizona State University
jeff.holmes@asu.edu
https://orcid.org/0000-0003-3693-4297
Jeffrey Holmes is an Instructor in Film and Media Studies within the Department of English. He received his PhD in Rhetoric, Composition, and Literacy from Arizona State University. His research focuses on the changing landscape of teaching and learning both in schools and in informal settings, in particular through digital media and distributed teaching and learning systems. He has published in academic journals such as *On the Horizon* and *Well Played* and regularly presents at conferences such as the American Educational Research Association (AERA), Games+Learining+Society (GLS), Association of Internet Researchers (AoIR), and the Rhetoric Society of America (RSA). He has also consulted on projects around the globe, including UNESCO, the Mahatma Gandhi Institute of Peace and

Sustainable Development, the Toronto Public Libraries, State Libraries of New South Wales, and Pearson, LLC.

David C. Berliner
Arizona State University
berliner@asu.edu
David C. Berliner is Regents' Professor of Education, Emeritus, at Arizona State University. He is a member of the National Academy of Education, the International Academy of Education, and a past president of both the American Educational Research Association [AERA]and the Division of Educational Psychology of the American Psychological Association [APA]. He has won numerous awards for his work on behalf of the education profession, and authored or co-authored over 400 articles, chapters and books. He co-edited the first *Handbook of Educational Psychology* and the books *Talks to Teachers*, *Perspectives on Instructional Time*, and *Putting Research to Work in Your School*. He has interests in the study of teaching, teacher education, and educational policy.

Mari Koerner
Arizona State University
mari.koerner@asu.edu
Mari Koerner is Professor Emerita of Education at Arizona State University's Mary Lou Fulton Teachers College. Her prior service at ASU includes serving as Dean of the College of Teacher Education and Leadership at Arizona State University's West campus (2006-2010) and Dean of Mary Lou Fulton Teachers College (2010-2016). Professor Koerner's book, *"THE SUCCESSFUL DEAN: Thoughtful Strategies and Savvy Tips for Today's Evolving Leadership,"* was published by Teachers College Press in 2020. She has served as principal investigator or co-director of over $100,000,000 million of grant-funded programs.

Appendix

Description of Score Data

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.

Appendix Table 1

Thematic Analysis: Percentage of Sample Reviews by Theme & Sub-Theme for each Topic

Topics	**1**	**2**	**3**	**4**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**18**	**20**
Personality / Rapport	**0.15**	**0.25**	**0.43**	**0.58**	**0.9**	**0.19**	**0.26**	**0.41**	**0.94**	**0.29**	**0.68**	**0.8**	**0.17**	**0.46**	**0.48**	**0.77**	**0.79**
bad relationship					0.06		0.03			0.02	0.05		0.02	0.02	0.23	0.04	0.02
parent relationship	0.01	0.02							0.14								
biased					0.06	0.01	0.02		0.1	0.02	0.55	0.02				0.04	0.02
picks on kids	0.01	0.01	0.06		0.02	0.01	0.02	0.04	0.02						0.02	0.13	0.02
picks favorites	0.03	0.05	0.06		0.74				0.02						0.02	0.06	0.02
does not care / lazy	0.01		0.04	0.04		0.03		0.04	0.16		0.02			0.02	0.05	0.12	
gets angry			0.02		0.04	0.01	0.02		0.08			0.02	0.07	0.02	0.03	0.08	0.1
hates questions		0.02	0.15						0.02							0.02	0.02
mean / no respect	0.01	0.05	0.09	0.08	0.12	0.03	0.03	0.07	0.41	0.04		0.02		0.04	0.07	0.1	0.29
scary																0.04	0.06
yells		0.03	0.02	0.1		0.01			0.06	0.04				0.02	0.02	0.08	0.12
unprofesional				0.14	0.02	0.01		0.02	0.14	0.08	0.02					0.02	0.02
abrasive /frustrating	0.06	0.04	0.04	0.14	0.06	0.01	0.02	0.09	0.08	0.1	0.05	0.12	0.05	0.15	0.03	0.23	0.12
cool / nice	0.02	0.02			0.02	0.07	0.07	0.15		0.04		0.71		0.17	0.03	0.06	0.02
treat like kids / fools		0.01														0.13	
strict	0.03	0.03	0.02	0.08	0.02	0.01	0.02	0.04	0.04	0.02	0.07		0.02	0.02	0.02		0.23
Duties	**0.26**	**0.37**	**0.04**	**0.16**	**0.06**	**0.19**	**0.13**	**0.3**	**0.04**	**0.59**	**0.02**	**0.1**	**0.07**	**0.15**	**0.11**	**0.29**	**0.02**
lack expertise	0.1	0.01	0.02	0.02	0.06	0.07	0.02	0.14	0.02	0.08	0.02	0.05			0.08		
makes mistakes	0.1	0.02		0.02				0.04		0.02			0.02				
distracted		0.02		0.02			0.02	0.07		0.08				0.07		0.02	
does other things		0.02		0.04		0.04	0.02	0.01		0.43				0.02	0.02	0.02	
chaos in classroom	0.01	0.01				0.04	0.03	0.09		0.06			0.02	0.07	0.02	0.21	0.02
disorganized	0.01	0.14		0.08		0.06	0.03	0.13				0.05		0.02		0.02	

Appendix Table 1 (Cont'd)

Thematic Analysis: Percentage of Sample Reviews by Theme & Sub-Theme for each Topic

Topics	**1**	**2**	**3**	**4**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**18**	**20**
slow / no grades	0.02	0.25				0.02			0.02	0.02						0.02	
no follow-through	0.01	0.04	0.02				0.02						0.02				
Bad / Not Teaching	**0.68**	**0.38**	**0.69**	**0.18**	**0.08**	**0.48**	**0.72**	**0.62**	**0.18**	**0.55**	**0.16**	**0.2**	**0.83**	**0.7**	**0.36**	**0.19**	**0.15**
no learning / unprepared	0.09	0.01		0.04		0.1	0.48	0.02		0.06		0.02	0.02	0.22	0.1	0.02	
boring	0.04	0.04		0.04	0.02	0.02	0.05	0.05	0.02	0.1	0.07	0.02	0.02	**0.48**	0.07	0.04	
busy work	0.1	0.16				0.09	0.03	0.03		0.02					0.02		
waste / pointless	0.01	0.02				0.04	0.13	0.03		0.02	0.02						
cannot explain / unclear	0.04	0.09	0.11	0.02	0.02	0.06	0.02	0.22	0.04	0.02	0.02	0.02	0.6	0.04	0.02	0.04	0.08
confusing	0.03	0.04	0.11	0.02		0.01	0.02	0.07					0.07	0.02	0.02		0.02
rushed	0.01		0.02			0.01	0.02	0.02		0.02			0.1				
unhelpful	0.08	0.06	0.11		0.04	0.06	0.08	0.07	0.14			0.05	0.19	0.07	0.07	0.02	0.02
ignores kids / questions	0.04		0.52	0.02	0.02	0.02		0.07				0.05		0.02		0.02	
discouraging	0.01	0.01				0.01	0.07	0.08		0.06	0.05		0.02		0.05	0.06	
off-topic	0.02	0.01		0.02		0.05	0.03	0.04	0.02	0.41		0.05	0.02		0.02		
by the book / ppt	0.47	0.03				0.08	0.03	0.13			0.02		0.02	0.02	0.03		0.04
Workload	**0.26**	**0.77**	**0.09**	**0.1**	**0.06**	**0.61**	**0.28**	**0.13**	**0.08**	**0.1**	**0.09**	**0.02**	**0.05**	**0.37**	**0.03**	**0.08**	**0.08**
unrealistic expectations	0.01	0.07	0.02			0.1			0.02				0.02			0.02	
too much work	0.05	0.29	0.02			0.03	0.03	0.02	0.02	0.02	0.02	0.02		0.02		0.04	0.02
hard / unfair (grader)	0.07	0.28		0.1	0.06	0.03	0.1	0.07	0.04		0.05			0.07	0.02	0.02	0.06
bad feedback	0.02	0.06				0.01			0.02								
unclear direction	0.01	0.24	0.02			0.08		0.04	0.02				0.02				
study guides misaligned	0.06	0.08	0.04			0.44	0.1			0.04					0.02		
easy a	0.04	0.04				0.02	0.08	0.02		0.04	0.02			0.3			

Appendix Table 2

Selection of the reviews sampled during the thematic analysis

	Topic 1
(1)	lectures full of information not found in the book. the test is entirely book-based. but you have to take notes, to hand in your notebook each chapter. smart.
(2)	[redacted] is the worst teacher I have ever had. Biology is supposed to be a "lab science" not completely powerpoints, videos, and lectures. Sure we've had labs, but all except one were modeling labs without actual data. Also other biology teachers teach way more information and have multiple dissections when we have had none.
(3)	She doesn't teach anything.She writes the assignments on the board & you do them. EASY though you can use the book on the tests.
(4)	Monotone voice during lectures and just reads off a powerpoint projector - doesn't make class interesting at all. For his AP class, you pretty much are on your own; you have to teach everything yourse
(5)	It doesn't take much to do [redacted] job. All she does is make you read out of the textbooks and write 4 pages about it.
(6)	Ms. [redacted] does not actually teach. Everyday, we take notes that she directly copies onto the powerpoint for the students to copy and barely goes over the homework. We barely spend enough time on one section for the students.
	Topic 2
(1)	He is such a slow grader. He still hasn't graded essays from a month ago, even though he has had plenty of time over the numerous breaks.
(2)	Inconsistent w/instructions...gives last minute instructions on 9 wk projects when students have completed or nearly completed them...does not tell each class the same instructions...grades harshly & inconsistently
(3)	she never hands back papers she is not only inconsitent but she is contradictory she does give extensions but rarely explains what should be done. HORRIBLE TEACHE
(4)	She couldn't keep up with the assignments and the grades and didn't prepare us for our exam. I don't know what grade I ended up with in that class because she never updated edline.
(5)	Sporadically assigns unclear homework assignments, often contrary to those printed in her homework calendar. Yells at the entire class over trivial matters, which wastes time.
(6)	Can you say bussy work? OMG she asigned more homework projects a week than most teachers do in a year. I cant believe she still gets paid.

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

	Topic 3
(1)	She doesn't help us at all and when u ask her a question she doesn't want to help and she doesn't let u correct her because she thinks she is always right even when she isn't and she doesn't let us go
(2)	She is so mean to you if you ask a question, she never gives you a straight answer and makes you feel stupid for trying to learn.
(3)	I asked her a legitimate question about a problem from the quiz, and she rolled her eyes and said "uhh....no.....moving on....". And I was like "wt*?." And then I asked Mr. [redacted] about the problem and he said that I was right and she was wrong. So basically, she seems to think she is always right, and she does not respect all her students.
(4)	Mrs. [redacted] most of time can't give us answers to our questons. We ask her something and she goes off onto something else.Instead of knowing the material she has to go back in the book to check what we are doing.She does not know how the program works. Harry
(5)	She insults other students and ignores some when they have question or they dont understand something.
(6)	My daughter was confused the whole year because she did not take questions
	Topic 4
(1)	this guy hardly showed up for work my goodness come on i mean all we learned about is bach last time i saw him he was at [redacted]
(2)	Dude, not a good vice principal. I remember he kept saying [redacted] name wrong. I guess he was nice, as long as he wasn't yelling.
(3)	He was awful. I had him last year. He started screaming at our class and another teacher had to come in and calm him down.
(4)	He called us "douchebags" and "assholes". On the first day he told us "you're probably not brilliant if you're in this class." He made a sixteen-year-old girl cry, and he made a girl in my class who had just moved from China and knew less than rudimentary English read complex scientific passages. It was his first year of teaching, so maybe he's changed, but this was my experience.
(5)	she is the worst!!!!!!!!!!! she made me come into her office 4 times because she forgot she already met with me!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! :(
(6)	She talks way too much about her nails. We were trying to cheer one of my friend up because she was upset. Everything was going fine until... Mrs. [redacted] came barging in. Made my friend her feel worse.
	Topic 5
(1)	She doesn't seem to be doing her job. I still don't have my attendence fixed and it's been many months now.
(2)	It is sad when a well respected educational institution is forced to continue to employ personnel of such low quality purely in the name of tenure. [redacted] is a shining example of the failures of the tenure system and a true detriment to the excellence

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

(3)	She is very unhelpful and should be eradicated from her current position at [redacted] high, she has repeatedly changed the complex rules and regulations to her own liking and had not been very beneficial to the successes of any department in regards to extra-curriculars.
(4)	Worst instructor in this world existing, highest grade in midterm was 92 / 100 for an engineer with us, 2nd highest was 54 !!!! he deal with students as if they are engineers and professionals, personally i knew that this instructor passed through horrible experience in the past and he is revenging >>>
(5)	can easily cross the line, hers is unfortunately a compulsory busy work heavy course. I don't care about Desperate Housewives!!!
(6)	[redacted] is a teacher NOT a professor according to my research so far. No teaching credentials found....this worries me. [redacted] students need degrees to excel in the workplace but [redacted] hires it's teaching staff based on experience only?
	Topic 6
(1)	Too moody, doesn't care about the choir at all. If only she cared more, then we'd be a much better group. Often plays favortites, and doesn't teach much at all.
(2)	She is a fave picker and doesnt let other students be dancers, or have main parts. we only get 2-3 lines.
(3)	She seems very picky and choosy... she doesn't like a lot of people and plays favorites. I don't enjoy her class even though I love playing my instrument...
(4)	She totally chose favorites and hated anyone who deny agree with her or who wasn't in cheer
(5)	i dislike her sooo much! she plays favorites and makes the non-favorites cry, and when they're crying, she yells at them more!she does'nt belong in a classroom!
(6)	completely unfair and picks favorites. if you are not one of her favorites, she picks on and insults you. very very mean. if you do one thing wrong, she will scream at you and embarass you on purpose.
	Topic 7
(1)	He puts stuff on tests that were never in our notes. I am wondering HOW ARE WE SUPPOSED TO GET THOSE QUESTIONS. i wonder why everyone is failing his class...
(2)	Very easy class but don't expect to learn anything. Lots of busy work. May have difficulty on exams because she has not taught what is on it very well.

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

(3)	She "taught' science, but didn't reallly know any phys sci (just some bio) and didn't care enuf to learn so she could teach us even tho that was part of the req. lessons.
(4)	doesnt teach very well in class because she cant control her students, so make sure to try to learn the information on your own before tests, because if you fail a test your grade goes wayyy down.
(5)	Her class is easy. But she DOESNT TEACH. At all, she writes on the board tells you to copy it. And her notes dont go with anything on the tests. TERRIBLE TEACHER.
(6)	Would like to switch out of AP and take it with another teacher. He is not preparing us for the exam!
	Topic 8
(1)	90% of the class is about himself, 10% is about physics. I learned more teaching myself than in his class. Unless you want your GPA to suffer, avoid his physics class at all costs!! It's sad that he is the only physics teacher in Gab. This guy should be fired and replaced
(2)	Worst teacher in history. If you want to have easy A, take [redacted]'s AP Physics Class. Don't take regular, because there is no difference between the two classes.
(3)	i didn't learn very much in his class we spent too much time talkign about the holocost and it is depressing. i wish he challenged me more. it was fun though.
(4)	dont listen to all these people. she may be a good person, but her class is all busy work. YOU LEARN NOTHING! its is a waste of time and will ruin your life
(5)	Her class is extremely easy... because you don't learn anything and she won't test you on much. Her exam questions are basic and very unhelpful. If you're aiming to get a 5 on the AP Physics 1 exam, you'll have to study on your own. She is one of the worst teachers I have ever had; you won't be able to learn much in class. If you wish to truly be prepared for AP Physics C, I recommend that you take Physics Honors or get tutored. Don't waste your time in this "challenging" AP class.
(6)	Took his AP Physics class. Homework consisted of writing down notes on videos he made on YouTube. Little to no learning in class. We usually completed problems on WebAssign with occasional help from him. All in all, he seemed more like a proctor than a true teacher
	Topic 9
(1)	The topics that we are covering seem easy, but she is making it 20 times harder!
(2)	One of the most disorganized teachers I have ever had. I thiink he gets rigid because he is so disorganized. Not a whole lot of intellectual steam.

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

(3)	Has her head up in the clouds. Very difficult to learn from. Very unclear. Often annoying. (But I suppose she is also a nice person.)
(4)	One of the worst teachers I have ever had. He has no idea how to teach and never answers questions, only leaves the students more confused. He also reiterates everything that isn't important. He once attempted to teach us something even he didn't know how to do, leaving all of the students angry and confused. He uses the textbook to teach. As in, he reads right off it. He has no structured class and cannot manage time. He mumbles to himself and is quite rude too sometimes.
(5)	He is absent from class often and is unclear and imprecise in his lessons. i have learned more on my own than from his class.
(6)	I strongly feel as though she doesn't ever listen to me and extremely unclear. She teaches with videos(um is that really teaching)
	Topic 10
(1)	Very poor attitude with students. Becomes angry easily and singles out students for mistakes. Loves to condescend. Several parents have had meetings with this person and no changes.
(2)	Ms. [redacted] is a insufficent teacher and is not capable of creating a safe and caring environment for the students she teaches. Please do not get the idea I don't respect her because the state of her
(3)	This is the only teacher my child has ever had an issue with in 11 years of schooling. She has zero compassion for students and their personal circumstances. She is unwilling to adjust her approach to a way that child might learn more effectively. My child and I have both reached out to her to try and get my child the extra help to succeed and she refused. This teacher marks assignments as zeros for my child when they are complete & correct because she assumes they aren't before checking.
(4)	Verbal and written communications sent home to parents are extremely condescending. Never had a problem with my kid enjoying school until her. I've never met a teacher so rude and talks down to parents if things are not done her way. Everyday my kid comes home I hear stories of how [redacted] has yelled at them and made the entire class run for "talking". The environment for children especially small children should be warm and welcoming not hostile. In a nutshell no I would not recommend her.
(5)	Terrible principal! Doesn't help the students when they need it. The vice principal does more work then her!
(6)	She is rather rude to certain parents whose children are core(zone) students, and very nice to parents whose children are magnet students.She's helpful if you are not hispanic or black,not a good person to be in her position, you have to treat all children the same fairness goes a long way,people tend to notice the difference quickly.Magnet parents are treated like GOLD and Core (ZONE KIDS) need I have to spell it out , it is without a doubt different in treatment at this school.

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

	Topic 11
(1)	she made no sense. she laughed at herself and talked about nothing the whole time. she was funny, but everyone in the class didnt get her.
(2)	All he talks about is his kids, and he has no idea what he is talking about. I didn't learn anything in his class, at all.
(3)	This guy the borin'est teacher in the math department, he tells stories that are irrelevant to the topic. ZZZZZzzzzzz
(4)	Always talks about all the stuff he did at Los Alamos. Doesn't do much in class, mostly talks about himself, sometimes goes over homework problems, takes forever to grade stuff. Shouldn't teach AC.
(5)	i dont think i've learned a thing in this guys class. all we do is talk about sports and watch pointless movies. (not that i'm complainig or anything)
(6)	he is very ..... never mind. his tests are easy, he is generous, but he does spend too much time talking nonstop.
	Topic 12
(1)	Liberal. Enough said.. She pushes her thoughts and opinions on others by leading questions and bias information. Such a hypocrite.
(2)	Can be over opinionated, does not like.to be contradicted about opinions or if she is misinformed continues to insist only on her views.
(3)	She is very opinionated and does not teach objectively. If you disagree with her, you will struggle in her class.
(4)	She sux. she cant even pronounce terms right. Her information is biased, she forces her religious beliefs on students.Theres a reason 4 sparating church & state
(5)	worst teacher i have ever had by far. feminist who does not tolerate anyone else's views, and to get a good grade, you have to pretend to have the same political views. also very arbitrary
(6)	She has biast ideas on subjects in civics and teaches her own ideas. She had us watch all these news stations BUT FOX news which is the only conservative one.
	Topic 13
(1)	she does not really teach students, however, she is overall a fine person
(2)	God-awful teacher. The only way to make it through IB bio is to get on her good side by following her quarky and whimful instructions. Good luck...
(3)	She didnt awnser my questions, and my grades fell in response. Not a very good teacher. She is nice, just didn't teach very well.
(4)	Ok oK. Your nice and all but you gotta explain some things betta then the way that you do.Otha wise you would be good.

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

(5)	hes the nicest guy in the wohle world hes just not a good teacher we never stay on topic in class....
(6)	she's a verryy nice person, but a verryy unclear teacher, i understand the material but i dont know what she wants from me...89!!!!!!!!! ahhh!!!
	Topic 14
(1)	Dr. [redacted] did not explain the material clearly, he easily loses his patience when asked to go over something they don't understand, and he does not make himself available for help outside of class.
(2)	Really struggle to understand this teacher, she means well but is not able to get ideas across, then gets angry and waves her hands a lot when no one understands.
(3)	[redacted] could be the worst teacher there is for Precalc. Doesn't explain the material well enough, no step by step when teaching problems, and doesn't care that the class isn't ready to be tested.
(4)	Mr. [redacted] fails to help the students when asked and quizzes and test were extreamly hard becuase he did NOT.. expalin enough
(5)	I have to help my son every day, because his teacher doesn't explain things clear enough for him.
(6)	she isn't very helpful and is not a very good teacher for this subject. she doesn't listen very well to students and does not try to help them.
	Topic 15
(1)	BORING, very boring! I don't really learn in his class, but he's nice. He never raises his voice, but he's the most boring teacher ever, he tries to teach, but he isn't clear and his tests are hard, but he does curve a lot, a 76% in a B in his class. overall, nice, but boring and you cant learn.
(2)	He's the most boring teacher ever. I received an "A" in his class but I didn't learned anything from him at all.
(3)	The class over all isnt so bad, its just so boring and dry, although Oviatt is funny...it's my least favorite class, but I dont like history anyway.
(4)	SHE IS SOOOO BORING!!! I CANNOT REALLY KEEP AWAKE IN THIS CLASS. SHE TELLS US TO TURN OFF OUR MONITORS AND IS SOOO BORING. The class is ez though.
(5)	whoever you guys are who think she is mean.you are wrong....she was boring and stuff but never mean and her class was the easiest class ever.i didnt learn anything
(6)	The most boring teacher i have ever had she dosent teach a thing, shes monotone, can't stand this class

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

	Topic 16
(1)	ms [redacted] doesn't know anything about teaching. she is without a doubt the worst teacher i have ever had. she doesn't even deserve to teach kindergarten, they'd probably hate her as much as our class does. no one respects h
(2)	with all the respect he deserves, hes the worst teacher i have ever had in the 6 years living in the u.s.. is my honest opinion
(3)	Bad psyche teacher, worse history teacher. Got fired because of that then re-hired because we needed a psyche teacher
(4)	WORST TEACHER EVER. She didn't teach me how to write a good essay. I had her as a Freshman and I am now a junior. Still the worst teacher ever
(5)	ms. [redacted] had made this year the worst year for everyone in ms. [redacted].she did not teach anything and she used to tell me i am not worthy.she is a bad teacher.
(6)	He is the absolute WORST teacher ever!!!Im not in his math class but people say they dont learn a thing.History has never been so boring and uneducational. I cant believe he is still teaching.bye bye!
	Topic 17
(1)	i personally dont think he is the best teacher or helper in the school. i would prefer to see ms. [redacted] or someone that runs a program at the school take charge. some of my friends also agree that h
(2)	ego gets the best of him, i had him as a football coach and most of the guys including myself didnt have the best of experiences with him.
(3)	SO GLAD I WILL NEVER HAVE TO SEE HER AGAIN!!!!!!!!!!!! WHAT COMES AROUND GOES AROUND!!TOO BAD I WONT BE AROUND TO SEE IT!!!!!!!!!!!!!!!!!!!!
(4)	Doesn't it seem that she wants to coach the "winning" team? Softball, Volleyball and now Basketball? Seems like what [redacted] wants [redacted] gets. WAKE UP [redacted] ADMIN!
(5)	He doesn't let us bring backpacks into the room and he doesn't move his neck or eyes. He is very stric and his teaching style is odd.
(6)	diver doesnt care about the students or what we think he took away the hat rule just when i was going into my freshman year and he ruined homecoming for everyon

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

	Topic 18
(1)	You don't seem to really care about what your teaching and it sometimes seems like all you want to do is get class overwith...how am i supposed to want to be there??
(2)	Unfortuntely Mrs. [redacted] does not know how to react with kids.If she did students would like her.Her class wasn't exactly a place I enjoyed learning in.
(3)	Pick favorites, always has an attitude and you are the rudiest teacher ever. So who should put forth more effort?
(4)	She is the worst! She is a control freak in some ways and doesn't take control where she needs to! She is only nice to the students she likes. And hey, how about a smile once in a while?
(5)	she likes to scream alot when nobody is doing anything at all!!!! if she doesnt hate you your going to get a 100.
(6)	She is like the grinch, you cannot smile with out getting written up, and nobody really cares that much. She thinks it is harder than it is.
	Topic 19
(1)	He has to lean how to be a better listener and learn how to relate to his students more so that they listen to him more.
(2)	you really need to learn how to teach music. Everytime you teach it you teach it different. your class is a living heck!!!!!!!!!!!!!!
(3)	She would be a better teacher is she actually pronounced the words right! andshe makes mistakes all the time she does "franglais"= french and english
(4)	She does not know her stuff and is very confusing. Should go back to farmington.
(5)	i wish she knew what she was talkin about. she only knows what's in the book. she should realize we're not preschoolers.
(6)	she can't spell for anything, says um WAY too many times and should learn the stuff she teaches before she tries to teach it
	Topic 20
(2)	I don't have u but every one says they don't like u and u r realy mean.mrs [redacted] is freaken cool though so it's all good-for her!
(8)	I don't like her. she's mean alot of times to certain people and yells at them for no reason. Shes ok though
(17)	She always yells at people for no reason. Some of her projects could be really confusing but she is sometimes nice.
(26)	she doesn't yell at people for something they've done per say. but yells random parts of sentences to get peoples attention. she has their attention. not in a good way. at all. BOOO.

Appendix Table 2 (Cont'd)

Selection of the reviews sampled during the thematic analysis

(29)	Picks favorites, is never clear in her instructions, is meaner in the afternoon than the morning
(40)	VERY moody and mean. Always told us she hated our class. She must have written her own good reviews. No one else would! BTW her new name is Ms. Pierce.

Note: Note all names have been redacted but all other grammatical markers, typos, and misspellings have been retained.