High-Resolution Course Feedback: Timely Feedback Mechanism for Instructors

Yunsung Kim
Stanford University
Stanford, USA
yunsung@stanford.edu

Chris Piech
Stanford University
Stanford, USA
piech@cs.stanford.edu

ABSTRACT

We study the problem of minimizing the delay between when an issue comes up in a course and when the instructors get feedback about it. The widespread practice of obtaining midterm and end-of-term feedback from students is suboptimal in this regard, especially for large courses: it over-samples at a specific point in the course and can be biased by factors irrelevant to the teaching process. As a solution, we release *High Resolution Course Feedback (HRCF)*, an open-source student feedback mechanism that builds on a surprisingly simple idea: survey each student on random weeks exactly twice per term. Despite the simplicity of its core idea, when deployed to 31 courses totaling a cumulative 6,835 students, HRCF was able to detect meaningful mood changes in courses and significantly improve timely feedback without asking for extra work from students compared to the common practice. An interview with the instructors revealed that HRCF provided constructive and useful feedback about their courses early enough to be acted upon, which would have otherwise been unobtainable through other survey methods. We also explore the possibility of using Large Language Models to flexibly and intuitively organize large volumes of student feedback at scale and discuss how HRCF can be further improved.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**.

KEYWORDS

Student Feedback on Teaching, Student Evaluations of Teaching, Timely Feedback, Course Survey, Course Improvement

ACM Reference Format:
Yunsung Kim and Chris Piech. 2023. High-Resolution Course Feedback: Timely Feedback Mechanism for Instructors. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3573051.3593391

L@S '23, July 20–22, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0025-5/23/07…$15.00
https://doi.org/10.1145/3573051.3593391

1 INTRODUCTION

Teaching is a complex organic interaction between students and teachers, and attending to student voices in this symbiotic relationship has increasingly become a routine practice in many higher education institutions [50]. Feedback from students about their course experience often provide rich and actionable insights about the course, and many instructors refer to them to reflect on their teaching and make meaningful and appropriate adjustments to their courses [16, 23, 28, 31].

In order to help instructors improve their teaching practice in time to benefit student learning, it is critical to obtain ongoing feedback from students in a timely fashion [11, 24, 26, 30]. Most universities regularly collect student feedback at the end of the term [39], and while end-of-term feedback suggests changes to be made for future offerings of the course, it offers less insight about the class currently in session. To this end, instructors often additionally collect midterm student feedback halfway through the course in order to aid adjustments for the remainder of the term [17].

This common practice, however, is still suboptimal at eliciting timely feedback for large courses. Most importantly, concerns that arise early in the course may remain unattended until halfway through the term unless students exert *extra efforts* to communicate with the teaching staff, at which point it could be too late to resolve the issue effectively. This delay can cause students to forget parts of their learning experiences or provide less detailed or distorted accounts of them. Such memory effects are commonly observed in retrospective surveys [22, 47] and could adversely impact the specificity and reliability of feedback. Also, the feedback students provide is often influenced by their expected academic performance [10, 18, 39], and when exams such as midterm or finals are scheduled close to the feedback period, feedback may be biased by factors unrelated to the teaching process.

Yet, blindly increasing the frequency of student feedback surveys does not effectively address these concerns either. Survey fatigue, also known as "over-surveying," can negatively affect the response rate and trustworthiness of surveys [20, 37], which can lead to an over-representation of the most vocal students in the class [3]. Moreover, not all students have new comments to make on a weekly basis, and surveys administered too frequently to all students may collect redundant opinions at the cost of repeated surveys. Rather than over-sampling at specific points in the course or requesting feedback from all students every week, the instructors could use the scarce resource of student feedback in a more intelligent way.

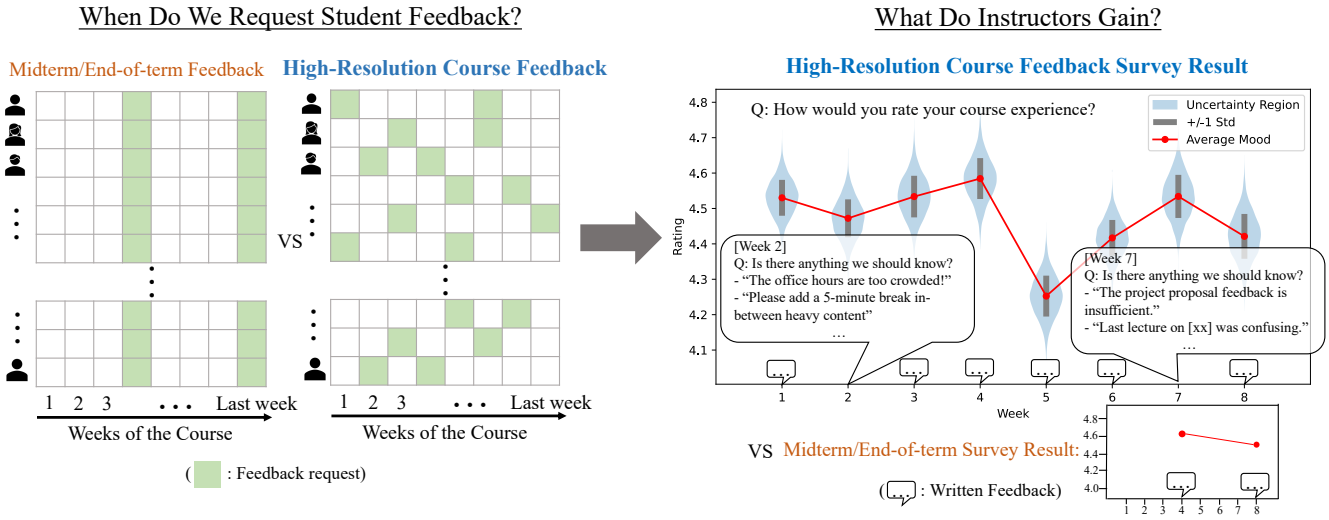


Figure 1: Schematic of the High-Resolution Course Feedback (HRCF) mechanism. HRCF makes the same number of survey requests to each student as the typical midterm/end-of-term feedback surveys, but evenly spreads the surveys across time. HRCF (1) reduces the time between when a class-wide issue arises and when the students are offered a chance to provide feedback about it, (2) allows changes to be made in time to benefit student learning, and (3) gives a high temporal resolution of the student opinions, whereas the common practice gives only a point estimate at two very specific points in the course.

In this work, we study the problem of obtaining timely feedback from students throughout the course while avoiding the complications of over-sampling. Our approach, which we call **High-Resolution Course Feedback (HRCF)**, is one that is surprisingly simple and straightforward: request feedback from a small random subset of the students each week, but survey each student exactly twice throughout the course (Figure 1). Despite the simplicity of this approach, what the instructors gain through it is immense: they can understand their students accurately at a much higher temporal fidelity and make necessary adjustments to the course early on.

Our study is based on the result of deploying HRCF to 31 college courses over 3 terms with a cumulative total of 6,835 enrolled students. Concretely, our work presents the following contributions:

- **The HRCF Insight.** We demonstrate both the qualitative and statistical utility of distributing feedback sampling over time without requiring extra efforts from students. In particular, we study the effectiveness of HRCF in obtaining time-sensitive comments from students and detecting significant mood changes throughout the course.
- **Public Release of the HRCF Tool.** We open-source our web-based implementation of the HRCF system.[1] Instructors are welcome to adopt the tool and use it for their courses.
- **Instructor Perspectives on HRCF.** We interviewed 9 users of HRCF to study how college instructors perceived HRCF and what tangible impact it had in their courses.
- **LLM Based Summarization of Comments.** We discuss the possibility of using Large Language Models (LLMs) to flexibly summarize student comments and organize them by semantic relevance.

We believe HRCF will lead to improvements for education at scale and serve as a valuable resource for timely feedback in classrooms.

1.1 Related Works

Student evaluations and feedback are widely used by many universities around the world to monitor the quality of education [8, 29], and various stakeholders in the education ecosystem rely on them for various reasons [28]. Many schools and school departments refer to them as measures of teaching effectiveness when making administrative decisions such as granting tenure, adjusting pays, and determining promotion [35]. For students, not only is student feedback a channel to make their voices be heard, but evaluations from other students can also provide them key information to be used in selecting courses and instructors [27–29]. Lastly, instructors and course designers often rely on student evaluations to analyze their teaching and make improvements to their courses [16, 23, 28, 31]. This work primarily concerns the last "formative" use of student feedback, and we leave the impact of our tool for other uses as future research.

The literature on the nature of student evaluations and feedback on teaching is rich, and researchers have engaged in profound debates over decades about its validity [2, 25, 31, 44], connections to academic achievements [10, 31], and potential sources of bias and correlations with factors irrelevant to teaching effectiveness such as course structure, grading leniency, instructor's gender, and student background factors [10, 15, 18, 39]. Overall, many studies support the idea that student evaluations generally convey helpful assessments of teaching performance and improvements to be

[1] https://github.com/yunsungkim0908/high-resolution-course-feedback

	Question Prompt	Intended Outcome	Response Type
Q1	"What did you like about the course so far?"	Encouragement to the teaching staff	Text
Q2	"Is anything from class still confusing to you?"	Self-reflections on the learning progress	Text
Q3	"Is there anything the teaching team should know?"	Foster an open-communication	Text
Q4	"How would you rate your course experience so far?"	To gauge the "mood" of the class	Qualitative Rating

Table 1: 4 Default questions used in HRCF surveys. "Class mood" refers to the average student rating for question Q4.

made [9, 29, 31, 36], and that they are capable of capturing multiple aspects of good teaching practice [44]. Instructors also often view student evaluations as useful resources for improving their teaching [4, 41].

Yet, the motivation of students to provide meaningful and authentic feedback is greatly contingent on their belief that their feedback will be valued by their teaching staff [6, 26, 44, 45]. [9] has shown that students consider improvements in the instructor's teaching practice as their most desired outcome for providing feedback, and improvements in course content and format as the second more desired. However, while many students possess the desire to express opinions and have influence on teaching, their lack of confidence on whether their feedback would be taken seriously by the teaching staff often results in their apathy towards providing careful feedback [43].

In this regard, end-of-term feedback is summative and retrospective by nature and often cannot benefit the current students providing feedback [24, 32], whereas *midterm* feedback can elicit useful formative feedback that can be acted upon earlier in the course [1, 10, 24, 34]. The literature supports the usefulness of formative feedback in improving teaching performance [9] and course content and structure [13, 42]. Midterm feedback has also been reported to have resulted not only in more favorable student ratings on instructional skills at the end of term [10, 34], but also in better learning achievements, more favorable affective outcomes [34], and increased student satisfaction with the feedback process [1, 48, 48]. [12] showed that student feedback obtained during the course also promotes a "two-way communication with learners on instructional design and decision making."

On maximizing the positive effects that formative feedback can have on the instructor's teaching quality and students' learning experience, studies [11, 26, 30] emphasize the importance of soliciting a well-timed and specific feedback on teaching behaviors and course structures. This study concerns the development of an instrument for improving timeliness of student feedback and its effects on improving the teaching and learning experience.

2 HIGH-RESOLUTION COURSE FEEDBACK SYSTEM

High-Resolution Course Feedback (HRCF) works by soliciting feedback each week from only a random subset of the students, enough to understand the common opinions and expectations in a classroom. Given the number S of total feedback surveys to request per student,[2] we schedule each student to be surveyed in S randomly chosen weeks, conditioned on the surveys being at least 2 weeks apart to reduce survey fatigue and prevent the previous week's response from affecting the current week's response. Assuming a uniformly random selection of survey weeks for each student, for a week in a W-week course with N enrolled students and response rate p, the number of feedback responses received R is

$$R \approx \frac{pSN}{W} \tag{1}$$

on average.

A single week of HRCF survey operates in the following 3 phases:

Monday (Noon): Send Survey Requests. The HRCF system sends out a survey email to all students scheduled to receive a survey that week. The email includes a link to an anonymous survey page where students provide answers to 4 default questions and additional custom questions chosen by the instructors each week.

Thursday (Noon): Remind Students and Instructors. Midway through the week, the HRCF system sends out a reminder email with a link to the survey page to the students who were sent a survey but haven't yet responded.[3] The HRCF system also sends out reminders to the instructors to update the custom questions and the roster of enrolled students for the upcoming week's survey through the survey settings dashboard page.

Sunday (4pm): Collect and Report Responses. At the end of the week, student surveys are closed and the responses are collected and sent to the instructors in a weekly digest email. The email reports the total rate of participation for that week, the visualizations and compilations of the collected responses, and a "class mood" graph (Figure 1) showing the estimated class mood across weeks. See the next section (Section 2.1) for more details about the digest email.

All responses are presented anonymously, but the frequency of each student's participation is reported to the instructors at the end of the term. Students are informed of this procedure in the notification and reminder emails they receive.

2.1 Survey Questionnaire and Weekly Digests

All surveys sent out to the students include the 4 "default" questions listed in Table 1, which were chosen to promote a constructive and civil interaction for both the students and instructors. In addition to the default questions, the instructors are given the option to add custom questions of their choosing each week. The custom questions were added to allow the instructors to ask about specific events in the course that they wished to focus feedback on. All

[2] While we set $S = 2$ to match the frequency of midterm and end-of-term surveys, the instructors were allowed to change S if desired. In our deployment, all instructors set $S = 2$ with the exception of one instructor who chose to set $S = 3$.

[3] Around 20% of the responses arrived after these reminders were sent.

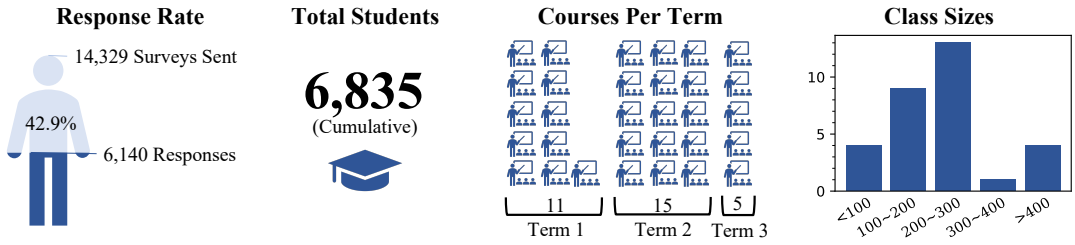


Figure 2: Deployment Statistics

questions asked in the survey are allowed to have either a text response, a numerical response, an integer rating between 1 ~ 5, and a corresponding qualitative rating from [Poor, Below Average, Ok, Good, Excellent].

At the end of each week, instructors receive a weekly digest email which contains the following 3 entries:

- **Weekly Participation Rate.** The total number of students surveyed and the number of students who responded are reported, along with their ratio.
- **Collection of Responses.** For each question, the responses are collected and listed in a single file. Responses to rating questions are additionally visualized as a histogram.
- **Weekly "Class Mood" Graph.** In HRCF, the average student rating for the question "How would you rate your course experience so far? (Q4)" is referred to as the weekly **"class mood."** The estimated weekly class mood is plotted (Figure 1, Right) for all weeks as a violin plot, along with the standard error of the mean (SEoM). To help contextualize the class mood, the plot also showed the average mood across all classes that use the tool along with its SEoM.

3 DEPLOYMENT

We deployed the HRCF system to 31 courses for 3 consecutive terms. 14,329 surveys were sent to a cumulative total of 6,835 enrolled students, and 6,140 responses were received, resulting in an overall response rate of 42.9% (Figure 2).

During the first 2 terms, a department-wide email was sent to the computer science faculty on the first day of classes with a description of the tool and instructions on setting up a course survey. During the 3rd term, no separate email announcement was made, and the tool was used only by instructors who directly reached out to the authors after having used the tool or hearing about the tool from colleagues. A typical course survey lasted for 9 weeks beginning in the 2nd week of term until the last week of classes.

4 ANALYSIS OF STUDENT RESPONSES IN HIGH-RESOLUTION COURSE FEEDBACK

In this section, we analyze the qualitative and statistical utility of High-Resolution Course Feedback. We will first analyze the timeliness of HRCF's written feedback and characterize HRCF's standard error of the weekly class mood estimates. We also discuss how to improve the accuracy of estimation when the number of observations is small and examine the ability of HRCF to detect significant class mood changes throughout the course.

4.1 Timeliness of Written Feedback

For HRCF to be helpful in quickly addressing course-level issues, it should be able to elicit feedback that instructors would prefer to be aware of promptly to allow a timely consideration. In this section, we address the question of how much of student-provided feedback in HRCF is "timely" from an instructor's perspective.

With HRCF, the instructors can choose to ask a different custom question each week, and this feature can be used to ask targeted questions of their choosing to obtain as timely and actionable feedback as they want. To set aside the effect of using targeted questions and analyze what students *proactively* inform the instructors, we selected a course with 273 students that did not use any custom question throughout the term and studied the student responses to the following two questions: "Is anything from class still confusing to you?" and "Is there anything the teaching team should know?"

The two authors made an initial pass through the comments together and discussed a coding scheme for timeliness. Although the notion of "timeliness" of feedback is necessarily subjective and open to interpretation, emphasis was placed on the following two criteria for a class-level issue that instructors would value being informed of without delay: (1) the issue is currently causing difficulties in a student's course experience, and (2) the issue can either be immediately resolved by the instructor, or should be brought to attention now to allow a timely resolution later if the instructor decides to take action. This resulted in the 3-way code shown in Table 2. Based on this coding scheme, the two authors independently coded each comment and resolved inconsistencies through discussions to arrive at a coherent labeling.

Figure 3 shows the number and fraction of student comments classified by timeliness for each week of the course. Most notably, 25%~50% of the feedback were considered timely for all but the last two weeks of the course, in which only 10~20% of the feedback were timely and a large fraction of it were deferrable. Topics brought up in timely comments from earlier in the course included instructor behaviors during lectures (e.g., asking to repeat questions that were asked quietly), requests for small adjustments in lecture style (e.g., pausing after heavy content, or having a short break during heavy lectures), inefficiencies related to the office hour structure, and early requests for reference materials that may take time to prepare. During the last two weeks of the course, most of the timely

No Issue: No course-level issue to be addressed by the instructor. *Constructed Examples*: - "Thank you for the thorough course materials!" - "X confuses me, but I'm yet to review the lecture." - "I'm a bit worried about the midterm, but that's normal."
Deferrable: Suggests a consideration for a non-immediate action, possibly in the long-term or for future course offerings. *Constructed Examples*: - "The course contents are dense." - "The first homework was a bit hard." - "I found the course somewhat fast-paced."
Timely: Expresses a difficulty that can be resolved immediately or should be brought to attention now for timely resolution later. *Constructed Examples*: - "Please repeat the questions when they're asked quietly." - "The lecture recordings are really bad. I couldn't hear anything that the professor was saying." - "The office hour wait is incredibly long. I had to wait 2 hours in line and still got insufficient help!"

Table 2: Codebook used for assessing timeliness of feedback.

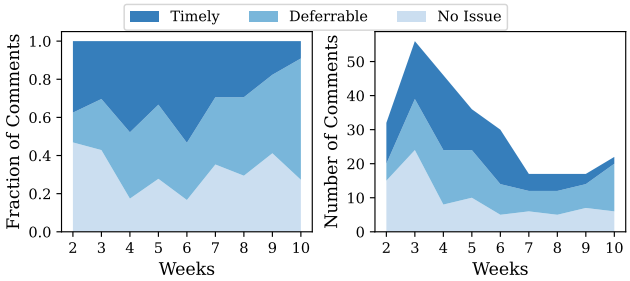


Figure 3: Fraction and number of comments for a select course by timeliness scale. Notice the spike in the number of observed responses in weeks 2-5 due to upsampling. (Section 4.1)

comments were focused on issues or requests specifically related to exams or final projects.

Notice in the right plot of Figure 3 a spike in the number of responses received in weeks 2 through 5. This spike is due to upsampling these weeks while downsampling the later weeks when selecting survey period. Since timely feedback arrives more often during the earlier stages of the course, this has the effect of increasing the volume of feedback received in the more timely and opportune phase of the course while still keeping constant the total number of surveys requested per student.

4.2 Accuracy of Estimated Class Mood

Recall from Section 2 that the average student rating for the question "How would you rate your response so far? (Q4)" is referred to as the "class mood." Since HRCF samples responses from a small subset of the entire class, it is natural to ask: How reliable are these mood estimates, and how much error should we expect from the

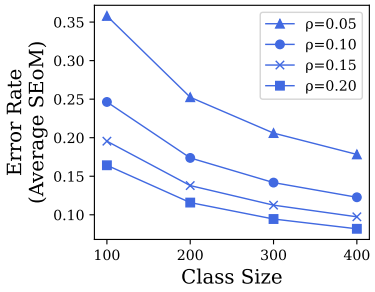


Figure 4: Average estimated standard error of the HRCF weekly class mood estimates for different class sizes N and ratio $\rho = \frac{R}{N}$ of number of observed responses to class size.

them? In this section, we characterize the average estimation error of HRCF's weekly class mood estimates and demonstrate that the error of estimation is small even when ratings from only a decent fraction of the total class is observed. This motivates the HRCF insight of spreading out survey requests to obtain higher temporal resolution at only a small cost of accuracy.

The student ratings observed in a single week of HRCF survey are samples *without replacement* from all student ratings for that week. For a finite population $\{x_1, ..., x_N\}$ of size N with standard deviation σ, the mean $\overline{X}$ of a sample $\{X_1, ..., X_n\}$ of size n has standard error [14, Theorem A.2.13][4]

$$\sigma_{\overline{X}} \approx \sigma \sqrt{\frac{N-n}{n(N-1)}}. \tag{2}$$

With this property, we can characterize the standard error of HRCF's weekly course mood estimates if we know the population size N, number of observed ratings R, and the population standard deviation σ of the true ratings. Although we do not have access to the ground-truth population standard deviation σ, we can instead compute the sample standard deviation and use it to approximate the standard error of the class mood estimates for each week of a course.

To illustrate what the standard of error of the HRCF estimates for a "typical" course would look like, we selected the course with the median average standard deviation for all weeks[5] and assumed that the standard deviation of the ratings would be the same as the chosen course. We then calculated the weekly standard error as we varied the ratio $\rho = R/N$ of the number of observed responses to the total class size.[6] (See Equation 1.)

Figure 4 plots the average standard error[7] of the weekly class mood estimates for varying values of N and ρ. From this we immediately notice that spreading the samples achieves high sample efficiency at the cost of only 0.12 and 0.17 error in estimation for

[4]This assumes that $n/N \to C \in (0, 1)$. Note the difference between the standard error for i.i.d. samples.
[5]We limited our search to courses that had at least 10 ratings for all weeks to avoid degenerate cases.
[6]To get a rough sense of the magnitude of ρ, a week with a 50% response rate in a 10-week course where students are surveyed in 2 uniformly random weeks yields an average of $\rho = 0.1$.
[7]A standard error of σ roughly means that the estimates are within 2σ of the true mood around 95% of the time.

$N = 400$ and $N = 200$, and the standard error continues to decrease as more students respond.

4.3 Improving Class Mood Estimation When the Number of Observations Is Small

When the number of observed responses is small for a particular week, the standard error of HRCF's weekly class mood estimate becomes large. How can we reduce estimation error when we have only a small number of observed ratings for a given week?

Bayesian inference is a useful paradigm for achieving small-sample robustness in parameter estimation. In this paradigm, we treat the parameter to be estimated θ as a random variable drawn from the *prior* probability distribution $p(\theta)$, and θ gives rise to the observations X through $p(X|\theta)$. Once $p(\theta)$ and $p(X|\theta)$ have been chosen, estimating θ based on observation X corresponds to computing the *posterior* distribution $p(\theta|X)$ given by Bayes' Theorem:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta}. \tag{3}$$

We study two Bayesian approaches to estimating the weekly class mood, which we call `Indep` and `Markov`. In both approaches, we make the (coarse and approximating) assumption that each of the observed ratings $X_{w,i}$ for week w are independent and identically distributed (i.i.d.) samples from a Gaussian distribution with mean θ_w and fixed standard deviation σ_X:

$$X_{w,i}|\theta_w \sim \mathcal{N}(\theta_w, \sigma_X).$$

`Indep` and `Markov` each make different choices for the prior distribution $p(\theta)$. `Indep` assumes that θ_w's are i.i.d. samples from a Gaussian with fixed mean and standard deviation:

$$\theta_1, ..., \theta_W \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_\theta, \sigma_\theta). \tag{Indep}$$

`Markov`, on the other hand, assumes that the θ_w's are drawn from a Markovian random walk that depends on the previous week's mean θ_{w-1}, which embodies the assumption that class mood doesn't change dramatically across weeks:

$$\theta_w|\theta_{w-1} \sim \mathcal{N}(\theta_{w-1}, \sigma_\theta), \quad \theta_1 \sim \mathcal{N}(\mu_\theta, \sigma_\theta). \tag{Markov}$$

Given all observed ratings up to the current week, both of these assumptions nicely yield closed-form Gaussian posterior distributions for the weekly class mood

$$\theta_w|X_{1,1}, ..., X_{w,N} \sim \mathcal{N}(\widetilde{\mu}_w, \widetilde{\sigma}^2_w),$$

whose parameters are given as

$$\begin{cases} \widetilde{\mu}_w = \dfrac{\sigma_X\mu_\theta + \sigma_\theta \sum_i X_{w,i}}{\sigma_X + N\sigma_\theta} \\ \widetilde{\sigma}^2_w = \dfrac{\sigma^2_\theta\sigma^2_X}{\sigma^2_X + N\sigma^2_\theta} \end{cases} \tag{Indep}$$

for `Indep`, and as

$$\begin{cases} \widetilde{\mu}_w = \dfrac{\sigma^2_X\mu_{w-1} + (\sigma^2_{w-1} + \sigma^2_\theta)\sum_i X_{w,i}}{\sigma^2_X + N(\sigma^2_{w-1} + \sigma^2_\theta)} \\ \widetilde{\sigma}^2_w = \dfrac{\sigma^2_X(\sigma^2_{w-1} + \sigma^2_\theta)}{\sigma^2_X + N(\sigma^2_{w-1} + \sigma^2_\theta)} \end{cases} \tag{Markov}$$

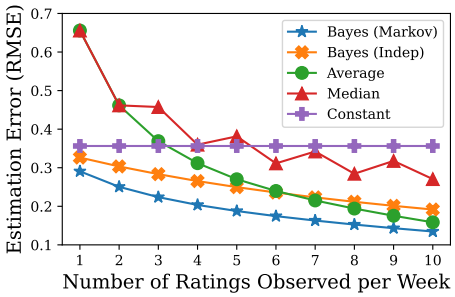


Figure 5: Evaluating different methods of class mood estimation for small sample sizes.

for `Markov`. We take the mean of these Gaussian posterior distributions as our point estimate of class mood.

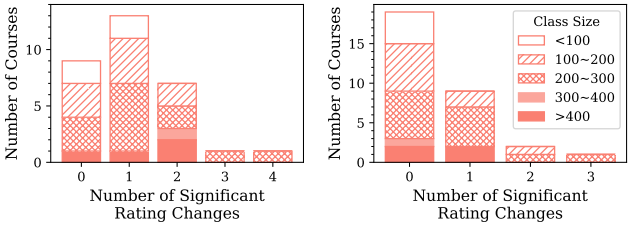
Estimation error with Bayesian inference. We compared the estimation error of `Markov` and `Indep` against 3 baseline methods of estimation: `Average`, `Median`, and `Constant`. `Average` and `Median` each compute the average and median of the observed ratings for each week, and `Constant` takes the average over all observed ratings across all courses and makes a constant estimate throughout the course. Similar to our study on standard error, the exact error of the estimates are unobtainable since the we do not have access to the true class mood. In this experiment, we used the average of the observed mood ratings for a given week as an approximation to the unknown true mood, and limited our experiment to 16 courses that had at least 10 responses for all weeks to ensure the quality of approximation.

For each course, we varied the number of ratings to be observed each week ($k = 1, ..., 10$) and compared the class mood estimated by each method against the approximated ground-truth mood. For each value of k, we simulated 5,000 uniform samples of k responses for each course and computed the root-mean-squared error (RMSE) of the estimates across all courses, weeks, and samples. Figure 5 shows the RMSE for each method of estimation. Note that `Markov` achieves the lowest RMSE for all values of k, with up to 0.1 improvement over `Indep` for $k = 5, 6$. `Indep` also has lower RMSE than `Average` for $k \leq 5$, which provides evidence for the robustness of Bayesian methods for small samples.

4.4 Capturing Significant Changes in Class Mood Throughout the Course

Are samples from a random subset of students enough for HRCF to detect significant changes in class mood throughout the course? From a statistical perspective, a pair of weeks in a given course has a significant difference in mood if the result of a two-sample Student's t-Test with finite population correction[8] for the equality

[8]We applied finite population correction since the observed ratings are samples with replacement from a finite population, not an i.i.d. sample.



(a) Across Multiple Weeks (b) Across Consecutive Weeks

Figure 6: Histogram of the number of statistically significant changes in class mood detected throughout 9 weeks of HRCF survey (a) across spans of multiple weeks and (b) across pairs of consecutive weeks.

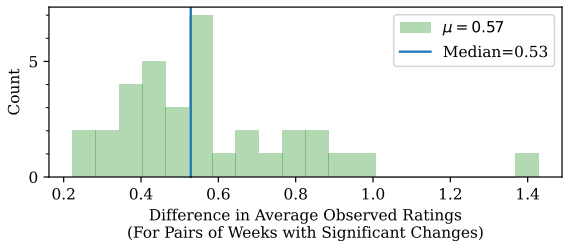


Figure 7: Histogram of the magnitude of the observed average mood rating differences for the 34 significant multi-week class mood changes in Figure 6a.

of means is statistically significant ($p < 0.05$).[9] For each of the participating courses,[10] we counted (a) the number of non-overlapping week-spans with significant class mood changes (multi-week mood changes), and (b) the number of consecutive pairs of weeks with significant class mood changes (weekly mood changes).

Figure 6 plots the histogram of the number of significant mood changes detected in each course, categorized by the size of the course. HRCF was able to detect significant class mood changes across courses of all sizes including ones that had fewer than 100 students, although changes in mood ratings were witnessed more frequently in larger courses. More than 2/3 of the courses witnessed at least one significant change in mood throughout the course, and nearly 1/3 of the courses had more than 2. For weekly mood changes, more than 30% of the courses had a pair of consecutive weeks where the class mood changed significantly.

Figure 7 further plots the distribution of the magnitude of the differences in average observed ratings for weeks with significant multi-week mood changes (Left of Figure 6a). Notice that more than half of the significant changes had an observed difference of over one half of a rating, which we take as a positive signal that large magnitude of differences in true class mood could be present. While this is a noisy estimate of the difference in true class mood, an interesting and important future research lies in characterizing the dynamics of the average course ratings with greater precision.

[9]Here we are making a simplifying assumption that the choices of which students' ratings are observed in 2 different weeks are statistically independent. In reality, this independence assumption can be broken due to several factors. For instance, this assumption may not hold for a pair of consecutive weeks as we avoid requesting surveys from the same student twice in a row. Nevertheless, we believe that this is assumption is a reasonable approximation for decently-sized classes.

[10]Most courses had 9 weeks of survey, with the exception of 3 courses that were shorter.

ID	Role	Course Size	Type
I1	Principal Instructor	200 ~ 300	Oral
I2	Head Teaching Assistant	400 <	Oral
I3	Principal Instructor	400 <	Oral
I4	Head Teaching Assistant	200 ~ 300	Oral
I5	Principal Instructor	400 <	Oral
I6	Principal Instructor	200 ~ 300	Oral
I7	Principal Instructor	100 ~ 200	Written
I8	Principal Instructor	100 ~ 200	Written
I9	Head Teaching Assistant	200 ~ 300	Written

Table 3: Information of the participants in the interview

5 INSTRUCTOR EXPERIENCE WITH HIGH-RESOLUTION COURSE FEEDBACK

Having analyzed the characteristics of student responses in HRCF, we now turn to perhaps the most important question from a practical point of view: what concrete impact can HRCF have in everyday classrooms?

To answer this question, we interviewed 9 instructors who used HRCF (Table 3) and asked about their experiences with the tool and how it impacted their courses in ways that other feedback mechanisms could not. The interview consisted of 6 questions which focused on (1) what concrete changes the instructors made in response to the feedback and the estimated class mood, (2) what differences they noticed between HRCF and other student feedback mechanisms they used in the past, (3) what new insights HRCF helped them learn about the class, and (4) their process of reading and addressing the feedback. 6 of the interviews were conducted orally, and 3 of the interviews occurred over email. In both oral and written interviews, we followed-up with more questions in response to the instructor's answers if we needed a clarification or a more in-depth analysis of their experience.

5.1 Early and Actionable Feedback

The instructors noted that HRCF helped them understand the frequent pain points and opinions of the students early on, allowing them to evaluate in a timely manner whether adjustments need to be made to their course or teaching behaviors. Even when students wrote about events that had already concluded (such as an exam or a recent assignment), the instructors found those feedback noteworthy to act on for the next similar event or future offerings of the course. The instructors also found HRCF useful in quickly and proactively probing student opinions.

5.1.1 Early Detection of Ongoing Student Pain Points. The most prominent impact that HRCF had on the instructors was to help them detect and potentially mitigate student pain points at an early stage. The following list highlights the areas where HRCF feedback

was considered in making adjustments and the specific types of adjustments that were made:

Lecture Styles (5 out of 9 instructors): Adjusting the pace of the lecture. Adding a short recap in the beginning of lecture. Having short breaks during long lectures. Providing additional clarification to prior week's lectures. Introducing more worked examples of most confusing contents. Improving the readability of whiteboard handwriting.

Course Content (6 out of 9): Addressing in lectures or sections the contents that students found most confusing in the previous week. Providing additional reading materials and worked examples during classes or around exam periods. Adjusting course load for the rest of the quarter depending on how much time students spent per week.

Course Structure (5 out of 9): Adjusting office hour structures. Introducing virtual office hours or additional office hours during high-demand periods. Requesting more TA support to the department.

Assignments (5 out of 9): Assessing levels and topics of student interest in assignments. Making adjustments to future homework based on reactions to on past homework. Determining whether to provide a class-wide deadline extension.

Several instructors (5 out of 9) noted that when multiple students repeatedly raised a particular issue, those comments were considered more heavily when determining the need to address the issue. For instance, the following account made by I9 was representative of and similar to several other instructor experiences:

> *We saw that students found the first homework confusing. In later homeworks, we added some homework-specific OH, where we highlighted parts of lecture that would be relevant and worked through similar problems. We were also extra careful in future homeworks to make sure our language was clear and TAs were on the same page when answering questions.* - I9

5.1.2 Quick and Inexpensive Probing of Student Opinions. The instructors also found HRCF useful in proactively inspecting the general student opinions to help gauge their teaching moves (2 out of 9). The modifiable custom questions of the week were used to inexpensively and actively probe diverse aspects of the course:

> *Some students told me offline that my lectures were too fast, but I wasn't sure if most students felt that way. So I asked this question on another survey, and students were mixed evenly between "too fast" and "too slow" so it was helpful to know I'm actually balancing well!* - I7

This was particularly useful when experimenting with new course activities or assignments and making consistent changes to them. For instance, I5 noted:

> *I'm pretty regularly making updates to the assignments, trying to get them to converge on something that works really well that get students interested and excited. [...] When I would ask people, like, how are the assignments going, I'd get a huge amount of feedback [...] And a lot of it is also just like, every time there's something new, "okay, let's see how this is going" [...] and getting that sample is very helpful there.* - I5

5.2 Increased Participation

Response rates are critical to the quality of student feedback surveys. When only few students provide feedback, the conclusions drawn from the responses may not generalize to the population of students in question [3, 39]. The problem of low response rates has been shown to be more prevalent in online surveys, which makes web-based surveys even more susceptible to errors [5, 40, 46].

In this regard, many instructors (5 out of 9) noted a much higher rate of participation and constructive feedback in HRCF compared to other methods of formative feedback surveys that they have used before at a course level. This included public discussion forums (such as EdStem or Piazza), anonymous surveys administered during the midterm and end-of-term (e.g., using Google Forms),[11] and anonymous weekly surveys that were open throughout the term. The higher rate of participation, along with the fact that the surveys targeted all students an equal number of times, gave several instructors more confidence in how representative the HRCF comments are of the general student opinion (3 out of 9):

> *It was also useful to get feedback from a larger sample of the class. That helped us understand which concerns were general and affecting many students.* - I9

5.2.1 Factors that can promote participation. While a rigorous study would need to be carried out to analyze the patterns of student participation in HRCF, one potential factor that can affect participation is transparency. 4 out of 9 instructors said that they brought up the comments during class and openly discussed which comments could and would be addressed and which comments could not (e.g., due to irreversible structures of the course). Providing students with the confidence that their feedback will be taken seriously and ensuring that their opinion is heard have been known to result in a greater voluntary participation of students in providing feedback [6, 26, 44, 45]. We hypothesize that a weekly cadence of such transparency will result in a more positive impact on participation.

Additionally, some instructors (2 out of 9) hypothesized that the act of directly soliciting feedback at different points of the term led to higher participation, which would be an interesting hypothesis to test in future research:

> *I think just putting something in their inbox makes it more likely for them to do it instead of posting a link and saying "please fill out the survey". So, it's just that little layer of personalization, like just putting it in their inbox that maybe they are prone to submit it more, also because not everyone's surveyed at the same time.* - I2

5.3 Weekly Mood Ratings as a Preliminary Screening

The instructors found the weekly mood ratings to be a useful measure of the general sentiment of the class and said that they reacted to them in some way (8 out of 9). 5 instructors said that they used the ratings for preliminary screening, to be followed by an in-depth analysis of the root causes of the changes in the mood if a significant change was present. For instance, I2 noted:

[11] Our institution didn't offer official midterm evaluations, so we could not compare HRCF against a systematized, institution-wide midterm evaluation system.

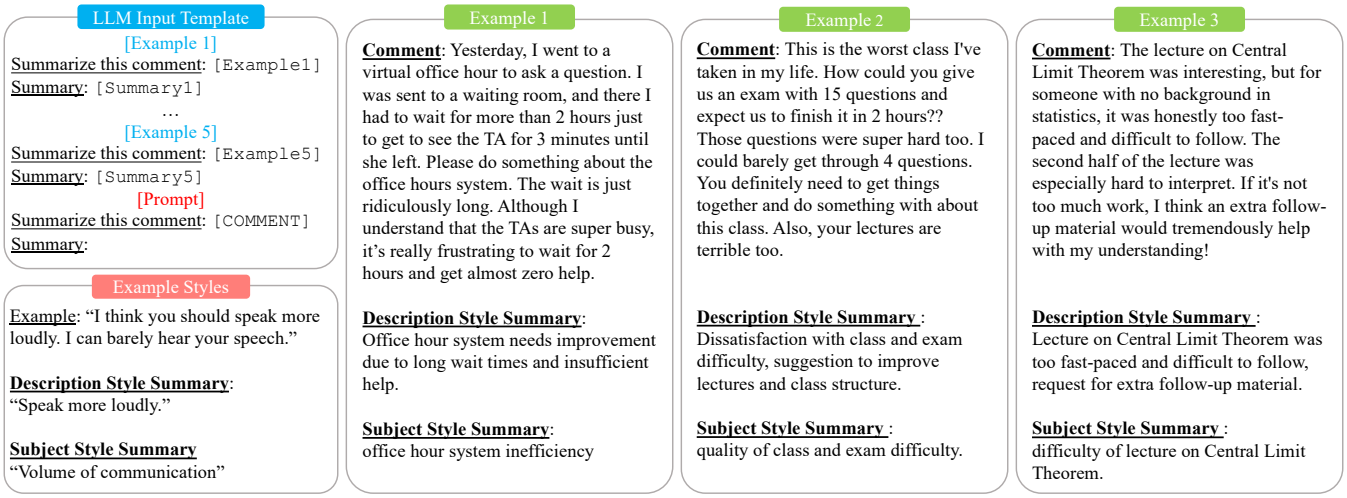


Figure 8: Example of using LLM's in-context learning to generate summary outputs for 3 author-constructed comments in 2 different styles. Examples included in the input prompts can be used to control the style of the resulting summary.

> *It gives us like a general idea of how the students are feeling and also kind of a preliminary gateway. Like, if the rating was really low in that particular week, we would spend a lot more time reading through every single feedback and just kind of see what we can do. If it's in general higher, we also noticed that the comments were more positive.* - I2

Similarly, 2 instructors said that the mood ratings were used to "validate their teaching moves" to quickly diagnose whether the students in the class faced any difficulties.

Ratings were also useful (2 out of 9) as a way to prioritize and triage which written comments to focus attention to:

> *The only time when I would be concerned is when I would see a student that would rate the class "poor," in which case, my first indication would be well, maybe let's talk about that same person's other response as to see if there's like a really big [class-level] pain point that we've just been missing.* - I4

Yet, concern was also raised about potentially over-fitting to the weekly ratings in ways that does not enhance student learning, which is a common issue in student ratings of teaching [50] that could be amplified by the increased frequency of HRCF feedback:

> *"My one concern is that we may have become too fixated on the scores, which turned our goal into trying to please students. [...] Overly fixating on scores made simple solutions like making assignments easier appealing"* - I9

6 TOWARDS AUTO-SUMMARIZING COMMENTS WITH LARGE LANGUAGE MODELS

To effectively address student feedback in massive classes that scale well beyond the size of a typical university classroom, it becomes critical to organize and present student comments with an intuitive and accessible structure. In fact, even in a typical university setting, most instructors we interviewed (6 out of 9) mentioned that they read the feedback comments by grouping them by topics, and that they would like to see a feature that organizes semantically similar comments together. In this section, we explore the possibilities of using Large Language Models (LLMs) to automate the organization and presentation of student comments according to their semantic relevance.

Earlier studies [19, 21, 38] have developed methods for organizing and visualizing written student feedback using sentiment analysis, topic modeling, and aspect extraction. Topic models and aspect extractors can be used to group comments that discuss the same aspects of the course, either for a pre-defined set of aspects such as lectures, assignments, or office hours, or for an unsupervised set of topics that would later be annotated by hand post-hoc.

While these methods could provide a coarse first-order summary and semantic grouping of feedback, what instructors need could often be more than just the aspect or the topic of the feedback comments, especially when the course is massive and feedback arrives in bulks. In addition to clustering together feedback on lectures, for instance, we might be interested in feedback that talk about the *pace* of the lectures. Among the comments on lecture pace, we might further consider grouping together comments that say that the lectures were too fast and compare them against those that say they were too slow. For such a wide range of abstraction levels, we practically cannot obtain a fine-grained enough training set to train an aspect extractor or use topic models that work for our targeted level of granularity.

Can we be more flexible in how we summarize and organize comments, perhaps by using a manageable number of demonstrations of our desired output style? State-of-the-art Large Language Models such as the latest GPT series [7, 33] are known to be "few-shot" learners that can generalize to many tasks after seeing just a few examples. These models achieve rapid generalization through "in-context learning," which uses the text of the input prompt as

a specification of a task. By providing few demonstrations of the style of summary that we desire, LLMs could generalize to provide summaries in the right level of abstraction, and we would be able to more easily and effectively organize comments based on them.

Figure 8 demonstrates the possibility of using in-context learning to generate different styles of summary for 3 author-constructed example comments. While a "subject" style summary extracts the fine-grained topic of the comment, the "description" style summary is designed to also retain the opinion expressed in the comments. The template of the input to the LLM (which was GPT-3 `text-davinci-003` in this demonstration) contains 5 (constructed) example comment-summary pairs followed by a prompt, where the provided summaries were in the desired style of summary. The resulting "subject" style output has enough detail to potentially be easily clustered together, and the "description" style output concisely delivers the main opinion.

Moreover, LLMs could also be used to control the tone of the output summary. Notice in all examples of Figure 8 (especially in Example 2) that emotional and potentially offensive language has been toned down to neutral. We may also explicitly state in our input prompt a description of the tone for the summary. A promising direction for future research is to find the best input template for generating the most useful abstractions of the comments, and a mechanism for semantically organizing the comments thereof.

7 DISCUSSIONS AND LIMITATIONS

7.1 Frequency of Survey

For our deployment, both the survey and the feedback digest report were all conducted with the same cadence of once per week. For typical university courses, a weekly cadence may be appropriate considering the workload of each student and the frequency of lectures. For much larger courses at scale such massive open online courses that take place every day for a short span of weeks, however, a different, potentially much shorter cadence of once every 2 or 3 days could be more suitable. Yet, as we discussed in Section 5.3, feedback administered too frequently may cause instructors to feel pressured to over-fit to the feedback or ratings, as changes made to the course are reflected almost instantaneously in the next round of feedback. An important direction for future research is to explore which cadence of surveys works best under which settings.

7.2 Potential Downsides of Anonymous Feedback

Although the anonymity provided by HRCF encourages candid and constructive feedback, it can also lead to comments that are hurtful and emotionally challenging to read as many anonymous feedback mechanisms do. Instructors are known to have very different emotional reactions to confronting anonymous student feedback [16]. Many instructors feel nervous and anxious about reading student comments [49], and it is not uncommon for instructors to take student feedback personally and feel devastated by the most extreme comments [4]. Several (4 out of 9) instructors also reported having had some level of emotional reaction to the student comments, ranging from feeling nervous to read the comments to feeling "crummy" to face a week with negative feedback or "targeted by the comments" as they read them.

Unlike typical student feedback systems, however, HRCF occurs on a weekly basis. An important topic for future research is to understand in what ways a high dosage of feedback from a subset students can emotionally affect the instructors and how it differs from feedback given during midterm or end-of-term surveys.

A possible strategy to minimize emotional impact is to have a 3rd person (or possibly even a machine learning agent, as discussed in Section 6) paraphrase unnecessarily hurtful feedback into less harmful language. 4 out of 9 instructors we interviewed said that they had a teaching assistant initially read the comments and summarize the main points for them, and while there were many reasons for doing this (e.g., for the efficiency of time), it had a buffer effect of focusing on constructive criticism. An important direction for future research is to automate the process of converting an emotionally hurtful comment to constructive criticism.

7.3 Student Perspectives on HRCF

The primary focus of our study was on the instructors' perspectives on HRCF when using it to make adjustments to their courses. Since students are the ones who provide these feedback, it is very important to explore the patterns in which students interact with HRCF and how they feel about it. Moreover, as the goal of using student feedback is ultimately to enhance student learning, it is also important to explore what downstream effects HRCF has on the quality of student learning experiences and their learning outcomes.

8 CONCLUSION

We proposed High-Resolution Course Feedback (HRCF), an open-source framework for obtaining timely student feedback with no extra burden on the students compared to the common practice of midterm and end-of-term feedback surveys. HRCF works by requesting feedback from a random subset of the students each week while keeping the number of surveys per student constant, typically at twice per term. This reduces the delay between when an issue comes up in a course and when the students get a chance to provide feedback about it, and allows instructors to make changes early enough to enhance student learning experience.

Based on our deployment of HRCF to 31 courses with a cumulative total of 6,835 students, we analyzed both the qualitative and statistical merits of HRCF in obtaining timely and representative feedback from courses. User interviews showed that HRCF collected constructive and actionable feedback from a wider pool of students than other methods they used, and that the average weekly course experience ratings were good summaries of the class mood. We further studied the possibilities of using Large Language Models (LLMs) to automatically provide fine-grained and controllable summary and organization of large volumes of feedback at scale. Understanding the student perceptions of HRCF and studying the emotional impacts these surveys have on instructors remain as important directions for future research.

ACKNOWLEDGMENTS

The authors would like to thank Stanford Institute for Human-Centered AI (Hoffman-Yee Research Grant) and Kwanjeong Educational Foundation for their generous support.

REFERENCES

[1] Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A., and Hess, C. W. Satisfaction with processes of collecting student opinions about instruction: The student perspective. *Journal of Educational Psychology 82*, 2 (1990), 201.

[2] Abrami, P. C., d'Apollonia, S., and Cohen, P. A. Validity of student ratings of instruction: What we know and what we do not. *Journal of educational psychology 82*, 2 (1990), 219.

[3] Adams, M. J., and Umbach, P. D. Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education 53*, 5 (2012), 576–591.

[4] Arthur, L. From performativity to professionalism: Lecturers' responses to student feedback. *Teaching in Higher Education 14*, 4 (2009), 441–454.

[5] Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., and Bell, D. Electronic course evaluations: does an online delivery system influence student evaluations? *The Journal of Economic Education 37*, 1 (2006), 21–37.

[6] Brown, M. J. Student perceptions of teaching evaluations. *Journal of Instructional Psychology 35*, 2 (2008), 177–182.

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems 33* (2020), 1877–1901.

[8] Centra, J. A. Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in higher education 44*, 5 (2003), 495–518.

[9] Chen, Y., and Hoshower, L. B. Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & evaluation in higher education 28*, 1 (2003), 71–88.

[10] Cohen, P. A. Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of educational Research 51*, 3 (1981), 281–309.

[11] Cross, K. P., and Angelo, T. A. Classroom assessment techniques. a handbook for faculty.

[12] Diamond, M. R. The usefulness of structured mid-term feedback as a catalyst for change in higher education classes. *Active Learning in Higher Education 5*, 3 (2004), 217–231.

[13] Driscoll, L. A., and Goodwin, W. L. The effects of varying information about use and disposition of results on university students' evaluations of faculty and courses. *American Educational Research Journal 16*, 1 (1979), 25–37.

[14] Ethier, S. N. *The doctrine of chances: probabilistic aspects of gambling*. Springer.

[15] Feldman, K. A. College students' views of male and female college teachers: Part i—evidence from the social laboratory and experiments. *Research in Higher Education 33* (1992), 317–375.

[16] Flodén, J. The impact of student feedback on teaching in higher education. *Assessment & Evaluation in Higher Education 42*, 7 (2017), 1054–1068.

[17] Gravestock, P., and Gregor-Greenleaf, E. *Student course evaluations: Research, models and trends*. Higher Education Quality Council of Ontario Toronto, 2008.

[18] Greenwald, A. G., and Gillmore, G. M. Grading leniency is a removable contaminant of student ratings. *American psychologist 52*, 11 (1997), 1209.

[19] Grönberg, N., Knutas, A., Hynninen, T., and Hujala, M. Palaute: An online text mining tool for analyzing written student course feedback. *IEEE Access 9* (2021), 134518–134529.

[20] Groves, R. M., and Peytcheva, E. The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly 72*, 2 (2008), 167–189.

[21] Hu, Y., Zhang, S., Sathy, V., Panter, A., and Bansal, M. Setsum: Summarization and visualization of student evaluations of teaching. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations* (2022), pp. 71–89.

[22] Iarossi, G. *The power of survey design: A user's guide for managing surveys, interpreting results, and influencing respondents*. World Bank Publications, 2006.

[23] Irons, A., and Elkington, S. *Enhancing learning through formative assessment and feedback*. Routledge, 2021.

[24] Keutzer, C. S. Midterm evaluation of teaching provides helpful feedback to instructors. *Teaching of psychology 20*, 4 (1993), 238–240.

[25] Kulik, J. A. Student ratings: Validity, utility, and controversy. *New directions for institutional research 2001*, 109 (2001), 9–25.

[26] Lewis, K. G. Using midsemester student feedback and responding to it. *New Directions for Teaching and Learning 2001*, 87 (2001), 33–44.

[27] Marsh, H. W. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. *The scholarship of teaching and learning in higher education: An evidence-based perspective* (2007), 319–383.

[28] Marsh, H. W., and Roche, L. The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American educational research journal 30*, 1 (1993), 217–251.

[29] Marsh, H. W., and Roche, L. A. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American psychologist 52*, 11 (1997), 1187.

[30] McLaughlin, M. W., and Pfeifer, R. S. *Teacher Evaluation: Improvement, Accountability, and Effective Learning*. Teachers College Press, 1988.

[31] Murray, H. G. Does evaluation of teaching lead to improvement of teaching? *The International Journal for Academic Development 2*, 1 (1997), 8–23.

[32] Narasimhan, K. Improving the climate of teaching sessions: the use of evaluations by students and instructors. *Quality in Higher Education 7*, 3 (2001), 179–190.

[33] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[34] Overall, J., and Marsh, H. W. Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of educational psychology 71*, 6 (1979), 856.

[35] Penny, A. R. Changing the agenda for research into students' views about university teaching: Four shortcomings of srt research. *Teaching in higher education 8*, 3 (2003), 399–411.

[36] Penny, A. R., and Coe, R. Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of educational research 74*, 2 (2004), 215–253.

[37] Porter, S. R. Raising response rates: What works? *New directions for institutional research 2004*, 121 (2004), 5–21.

[38] Pyasi, S., Gottipati, S., and Shankararaman, V. Sufat-an analytics tool for gaining insights from student feedback comments. In *2018 IEEE Frontiers in Education Conference (FIE)* (2018), IEEE, pp. 1–9.

[39] Richardson, J. T. Instruments for obtaining student feedback: A review of the literature. *Assessment & evaluation in higher education 30*, 4 (2005), 387–415.

[40] Sax, L. J., Gilmartin, S. K., and Bryant, A. N. Assessing response rates and nonresponse bias in web and paper surveys. *Research in higher education 44*, 4 (2003), 409–432.

[41] Schmelkin, L. P., Spencer, K. J., and Gellman, E. S. Faculty perspectives on course and teacher evaluations. *Research in Higher Education 38* (1997), 575–592.

[42] Simpson, R. D. Uses and misuses of student evaluations of teaching effectiveness. *Innovative Higher Education 20*, 1 (1995), 3–5.

[43] Spencer, K. J., and Schmelkin, L. P. Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education 27*, 5 (2002), 397–409.

[44] Spooren, P., Brockx, B., and Mortelmans, D. On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research 83*, 4 (2013), 598–642.

[45] Svinicki, M. D. Encouraging your students to give feedback. *New Directions for Teaching and Learning 2001*, 87 (2001), 17–24.

[46] Thorpe, S. W. Online student evaluation of instruction: An investigation of non-response bias. In *42nd Annual Forum of the Association for Institutional Research, Toronto, Ontario, Canada. ERIC Document No. ED472469* (2002).

[47] Tourangeau, R. Remembering what happened: Memory errors and survey reports. In *The science of self-report*. Psychology Press, 1999, pp. 41–60.

[48] Veeck, A., O'Reilly, K., MacMillan, A., and Yu, H. The use of collaborative midterm student evaluations to provide actionable results. *Journal of Marketing Education 38*, 3 (2016), 157–169.

[49] Yao, Y., and Grady, M. L. How do faculty make formative use of student evaluation feedback?: A multiple case study. *Journal of Personnel Evaluation in Education 18* (2005), 107–126.

[50] Zabaleta, F. The use and misuse of student evaluations of teaching. *Teaching in higher education 12*, 1 (2007), 55–76.