Rethinking Teaching Evaluation Reports: Designing AI-transformed Student Feedback for Instructor Engagement

ANONYMOUS AUTHOR(S)

Many instructors minimally engage with or avoid student evaluations of teaching (SETs) due to the significant time, cognitive, and emotional cost associated with effective usage. Nevertheless, SETs can contain feedback about students' learning experiences that instructors can use to improve instructional and educational delivery. In this work, we explore how to redesign SET reports to increase instructor engagement with this feedback. We explore the use of language models (LMs) to process and filter students' feedback to highlight recurring or important ideas, to identify actionable changes for instructors, and to de-emphasize demotivating aspects of this feedback. We explored a 4×4 strategy-presentation design space, generating six representative mock-ups that combine different strategies with various presentation formats. Through interviews with 16 post-secondary instructors, we learned how and when they engage with current SETs, and how they would perceive and use the LM-powered redesigned SET mock-ups. We found that instructors valued different kinds of presentation strategies depending on their needs, be it to actually improve their teaching, to get a one-time gestalt impression of their teaching performance, or to provide summative reports about their teaching performance. These findings shed light on new opportunities for designers to design dynamic SET reports, customized to instructors needs.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; *HCI design and evaluation methods.*

Additional Key Words and Phrases: Student Evaluations of Teaching, SET, Feedback, AI-powered Design

ACM Reference Format:
Anonymous Author(s). 2018. Rethinking Teaching Evaluation Reports: Designing AI-transformed Student Feedback for Instructor Engagement. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 36 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Student Evaluation of Teaching (SET) – also referred to as Teaching Evaluations or Course Evaluations – is one of the most common methods used for evaluating teaching and courses in higher education [21, 37]. These evaluations, typically conducted at the end of a course, allow students to share their opinions on the quality of instruction, course materials, and their overall learning experience. SETs are intended to "safeguard and improve the quality of instruction received by students" [12] and conceptualized to give students a "voice" [104].

However, a growing body of research highlights the alarming emotional and psychological toll that SETs can take on educators [47, 60, 66]. Studies have shown that SETs often contain non-constructive, abusive, or potentially harmful comments, with some students using them as a tool to bully and inflict harm on teachers [23, 67]. The impact on educators' wellbeing is substantial, leading to stress, mental and physical health issues, and potentially job dissatisfaction and burnout [6, 66].

Despite these recognized potential harms, most institutions do not implement screening measures for SET comments, perhaps due to resource constraints, technological limitations, or the perception that offensive comments are relatively

Manuscript submitted to ACM

rare [23, 111]. While previous research has explored automated approaches to analyze SETs using text analytics and LMs [24, 52, 88], there is still room to explore how these efforts align with instructors' specific needs and goals [92].

Our research is motivated by the research question: **How can we support instructors in their goals to gather practical and useful information from their SETs while minimizing the impact of distracting and unhelpful commentary?** Current SET reports typically present responses according to a set of standardized questions posed to students (See example in Appendix A). This static, unfiltered format often requires instructors to search through all responses to find related comments across different questions.

Building upon broader CSCW research on supporting feedback in a collaborative setting adhering to the traditional structure dictated by administrator-posed *questions*, we begin our design process by considering: *what do instructors want to think about and understand from their SET reports*? Moreover, the increasing capabilities of specialized language models (LMs) offer opportunities to perform natural language processing tasks such as topic modeling, text summarization, information extraction, and ideation [16] based on student feedback. Recently, the generative capabilities of pre-trained large language models (LLMs) also offer opportunities to identify latent meaning and nuance in student comments [13]. By harnessing the power of these NLP techniques, our work explores novel ways to provide structure to the vast amounts of unstructured text typically found in SETs.

We employed a multi-stage approach to explore redesigned SETs in this work. We first generated a set of mock designs to present feedback in different ways and explored several theory-informed strategies to cope with negative feedback. To demonstrate these could be created based on existing SET reports, we designed and built a system that would transform our institutions' SET reports into our mock presentation designs using leveraging LMs (e.g., SiEBERT, GPT-3.5-turbo) for sentiment analysis and zero-shot classification. To understand how these mock designs could be improved, we conducted an interview study with 16 instructors. We presented instructors with LM-generated mock designs based on the actual SET reports they've received. This approach helped us gauge their perceptions of various AI-powered design interventions and identify potential refinements to these mock-ups to better address their needs.

Our findings show that instructors engage with SETs for a number of fundamentally different reasons. Yet, these are often outweighed by the cost of the engagement—for instance, when they encounter negative feedback that is damaging rather than negative feedback that is actionable. These reinforce the motivations for exploring SET report design. Instructors helped us to identify even more strategies to improve SET presentation and design, including providing multiple views of the data, offering actionable insights and suggestions for improvement, balancing positive and negative feedback presentation, and enabling interactive exploration of the feedback.

Our work contributes to both higher education and HCI through the following:

- A empirically-derived typology of student negative feedback derived from instructors' lived experiences and perceptions. This categorization reflects interpretations of different feedback types, providing a foundation for designing systems that align with instructor perspectives.
- Empirical insights from a user study using design mock-ups based on real SET report data and theoretically-grounded, LM-powered intervention strategies and presentations, uncovering how instructors interact with and perceive AI-enhanced SET redesigns and factors influencing their effective usage.
- Design implications that illuminate promising avenues for future work in reimagining teaching evaluations. These include exploring the role of AI in feedback processing, developing hybrid and dynamic interaction modalities, and creating systems that support longitudinal engagement with feedback.

2 BACKGROUND AND RELATED WORK

2.1 Student Evaluation of Teaching (SET): Purposes, Uses, and Challenges

Student Evaluation of Teaching (SET) has become a standard practice in post-secondary institutions worldwide [18, 21]. These evaluations serve multiple stakeholders and purposes: students voice their opinions on teaching quality and learning experiences [119]; administrators use SETs to track teaching performance for certifications and rankings [82]; and instructors use them to reflect on and improve their teaching practice [120]. Additionally, SETs provide input for appraisal exercises (e.g., tenure/promotion decisions) and offer evidence for institutional accountability [100].

While formal SETs were initially introduced in the 1970s primarily for formative purposes, they have evolved to serve both formative and summative roles [8, 36, 50, 100]. Formative use of SETs aims to understand how teaching is received by students and to make improvements [120]. Instructors can use SETs to identify student misconceptions, struggles, and learning gaps, and to assess how to address those gaps. On the other hand, summative use factors into administrative decision-making and performance evaluations [107, 117]. However, this dual-purpose creates tension, often leading to "fear, damaged relationships, and self-doubt" [61], particularly among junior faculty who may lack the experience to critically assess student feedback [127]. Prior work has found the summative use of SETs to be problematic as it may not truly reflect the effectiveness of teaching [42, 50, 58, 100, 117]. These concerns include misalignment between student and instructor perceptions of effective teaching [5, 28], students' tendency to report negative experiences more readily [119], and the impact of poorly designed questionnaires on data reliability [100, 101]. Critics also point to issues of timing, consistency across courses, and unclear metrics [114], leading many to question the validity of SETs as a sole measure of teaching effectiveness [58].

Furthermore, a growing body of research highlights the alarming emotional and psychological toll that SETs can take on educators. Instructors report that SETs contain non-constructive, abusive, or potentially harmful comments [23, 67]. It is widely acknowledged that some students use SETs as a tool to bully, wound, and inflict harm on teachers [67]. This abuse can be particularly severe for women and marginalized academics, who receive lower ratings and abusive comments at higher rates [47, 77, 79]. The impact of these negative evaluations on educators' wellbeing is substantial. A survey with 810 instructors found that a vast majority (81%) of respondents reported receiving anonymous feedback that caused personal stress, with significant negative impacts on mental health (64%) and physical health (56%) [66]. The experience of receiving such comments has been identified to be similar to cyberbullying [66], which has been defined as an "aggressive, intentional act carried out by a group or individual using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend himself or herself" [99]. SETs have been identified as contributing to educator stress through "mischievous and untrue criticisms that damage the morale of teachers" [60], with particularly devastating effects on precariously employed female educators [98].

The emotional costs can lead to concrete negative consequences. The anticipation and repeated exposure to negative and critical evaluations can further lead to job dissatisfaction and even burnout [6]. The anonymity of SETs may depersonalize student-instructor relationships or even lead to abusive responses [14, 119]. These issues have led many instructors to disengage from SETs or focus solely on quantitative scores for career purposes, rather than using the feedback to improve teaching [96]. Some instructors avoid reading student survey comments altogether due to fear of encountering abusive or unacceptable remarks, preventing them from engaging with constructive feedback [24]. This disengagement can be counterproductive, potentially leading to changes that don't actually benefit student learning.

2.2 Coping with negative feedback

Despite the concerns raised about SETs, these evaluations remain a valuable tool for improving teaching quality when used appropriately [114]. Prior work indicates that student feedback can contribute to the development of lecturers' professionalism, while others also note that feedback is crucial for lecturers' reflective practices [6, 56]. This underscores the importance of not discarding SETs entirely, but rather focusing on how to effectively cope with and utilize the feedback they provide, particularly when it is negative. The impact of SET feedback on instructors is largely determined by their interpretation and response to it. As Gaertner argues, student feedback can assist lecturers in developing their teaching only if it is constructive and if lecturers understand, interpret, and cope with it properly [35]. This interpretive stance on feedback highlights the need for effective coping strategies, especially when dealing with negative comments.

The ways instructors cope with feedback can be broadly categorized into problem-based and emotion-based approaches [9, 34]. Problem-based strategies focus on addressing issues directly, while emotion-based strategies deal with managing the psychological impact of feedback. Arthur's typology [6] provides a useful framework, identifying four common reactions to student feedback: shame, blame, tame (the students), and reframe (seeing negatives as opportunities for growth). Building on this understanding, researchers have identified several strategies to help instructors cope more effectively with SET feedback. Reflective practices, such as keeping teaching diaries, allow instructors to contextualize student comments [112]. Collaborative approaches, like peer mentoring [55], provide external perspectives and support. Developing feedback literacy skills [26] and using visualization tools can enhance instructors' ability to process and act on SETs constructively. The emotional aspect of receiving feedback is particularly important. Värlander [113] suggests a novel approach where instructors provide feedback on the feedback they receive, addressing questions like "How did you perceive the feedback?" and "How did you feel when receiving it?" This process not only allows for emotional release but also helps instructors better understand and adapt their own feedback practices [27].

These coping strategies align with positive psychology perspectives, particularly Fredrickson's broaden-and-build theory [33]. This theory posits that cultivating positive emotions, even in the face of negative feedback, can build resources for future challenges. Built on this theory, research has found that how instructors interpret and respond to SET feedback can lead to either upward or downward emotional spirals [76]. However, despite some lecturers managing student feedback well, the authors found that others continue to struggle, even after pedagogical training. The paper suggests that existing support structures are often incidental rather than intentionally designed to help lecturers manage feedback, and more purposeful cultivation of positive coping strategies is needed.

2.3 Automatic text analytics with NLP

Although the emotional toll and potential harm induced by negative feedback in SETs are well-recognized, most institutions don't implement screening measures. Heffernan [47] found that only 21% of surveyed academics reported their institutions filtering or censoring comments before release. This lack of intervention is often attributed to resource constraints, technological limitations, and the perception that offensive comments are relatively rare [23, 111].

In response to these challenges, academic researchers have explored automated approaches to analyze and mitigate harmful content in SETs. These efforts have demonstrated clear benefits of using text analytics and NLP techniques to process free-text comments written by students [24]. Researchers have produced tools that can provide visual summary reports and suggestions [88], or summaries and visualizations of the underlying SETs [52], while others analyze the feedback using topic modeling and emotion analysis [40]. In a recent work, Cunningham et al. [23] applied machine learning techniques to screen and remove abusive or harmful comments in SETs, drawing inspiration from similar work in online communities, such as the automatic detection of misogynistic tweets on Twitter [7]. The application of NLP to SET analysis extends beyond screening for harmful content. For instance, Hum et al. [54] discussed how their approach to text analysis of SET surveys revealed "critical issues that merited or required immediate intervention".

While potentially useful, these works have rarely reported on whether the tools were ultimately useful for instructors, or even if they were what instructors were seeking in their SETs. Moreover, language models trained on SET data may inadvertently perpetuate existing biases. For instance, Okoye et al. [86] found correlations between the prevalence of negative sentiments and instructor gender, as well as confidence in teaching. Similarly, Rybinski et al. [90] demonstrated that while student evaluation text could predict quantitative ratings to some extent, such models exhibited gendered biases. Some researchers have begun to address the practical application of these tools in institutional contexts. Santhanam et al. [92] also concluded that while there is growing interest in text analysis of qualitative SET data and agreement on its value for quality improvement, many of the approaches are resource-intensive. They also noted a lack of consideration for how these methods can be feasibly integrated into institutional reporting and quality assurance processes. In our work, we take a user-centered approach, aiming to identify new design opportunities that address instructors' needs that are provided by capabilities of LMs.

2.4 Tools Supporting Feedback Processing

The HCI community has long recognized the importance of effective feedback processing, particularly in educational and design contexts. Sadler [91] argued that good feedback must be specific, goal-oriented, and actionable, providing a foundation for much of the subsequent work in this area. HCI Researchers have explored various approaches to support feedback processing, predominantly through two avenues of research. One has concentrated on structuring feedback during elicitation to improve its quality and usefulness [38, 64, 64, 109, 126, 131]. For instance, CritViz [109] supported peer critique in college courses, while Voyant [126] employed visualizations such as word clouds and histograms to aggregate crowd feedback. However, in the context of SETs, unlike crowd workers, students are the direct recipient of the teaching experience. Intervention to ensure the quality of feedback may compromise the authenticity of students' experiences.

Therefore, our work has more overlap with other line of approach, which is to support feedback recipients in engaging with and interpreting the feedback they receive [130]. Prior research has shown that for feedback to be effective, recipients must interpret, learn from, and act on it [62, 121]. Various strategies have been explored to facilitate this process, including reflection [4, 129], coping activities [124], and action planning [59]. One significant challenge in feedback processing is the cognitive demand imposed by conflicting perspectives within the feedback [89]. To address this, researchers have investigated ways to add structure to feedback content [32]. Visualization techniques have emerged as a promising approach to facilitate feedback interpretation and decision-making. For example, ConsensUs [74] supported multi-criteria group decisions by visualizing points of disagreement, while Unakite [73] scaffolded developers' decision-making using web-based information. In the broader context of text visualization, researchers have developed techniques to extract and visualize attributes such as topic, sentiment, and term frequencies [53, 72, 128].

The sentiment and tone of feedback have also been shown to significantly impact its perceived usefulness and the recipient's ability to engage with it constructively. Studies have found that positively framed feedback tends to be rated higher [131] and can lead to better overall work quality [85]. However, the relationship between sentiment and usefulness is complex, with some research suggesting that mildly negative feedback can be particularly effective [65]. The order in which feedback of different sentiments is presented can also influence its reception [123]. Importantly, negative feedback can evoke strong emotional responses, especially when it conflicts with the recipient's self-perception

[93]. To mitigate these effects, researchers have explored strategies such as balancing positive and negative feedback [124] and facilitating reflection to enhance feedback acceptance [94].

Our work builds upon these findings and approaches, leveraging the enhanced capabilities of Language Models (LMs) to process and present feedback in novel ways. This approach allows us to scale the benefits of structured feedback and visualization techniques to the large volumes of unstructured text typically found in SETs, while also incorporating strategies to mitigate the potential negative emotional impact of critical feedback.

3 EXPLORING SET DESIGNS

We focused on feedback *strategies* and *presentations* in our design process. For *strategies*, we drew from prior literature to identify approaches that address barriers to engaging with negative feedback, are feasible to implement using NLP techniques, and are compatible with the existing format of anonymous, textual feedback. For *presentations*, we developed a conceptual framework based on two fundamental dimensions: the degree of structure and the balance between analytical and narrative approaches (see Figure 1). This allowed us to systematically identify and explore different parts of the design space, ensuring a diverse range of approaches. The specific rationales and design details will be elaborated in the following sections.

We identified a final set of four strategies to encourage engagement with students' feedback, as well as four presentation designs to enhance instructors' ability to discern and identify important information. Visual instances of these strategies are illustrated in Appendix A, Figure 3, and Figure 4. We realized these first as a set of visual mock-ups, refining these through discussion and iterations. While our selected set of strategies and presentations is grounded in the prior literature and the conceptual framework, they are not intended to be exhaustive or prescriptive, but rather to serve as probes and have variations to elicit a broader spectrum of needs and issues.

3.1 Strategies to encourage engagement

We selected feedback strategies based on three criteria: 1) addressing the primary barriers to engaging with SETs (harmful and unconstructive negative feedback), 2) feasibility of facilitation through NLP techniques, and 3) compatibility with the existing format of anonymous, textual, short, qualitative feedback. Our feedback strategies were informed by prior literature [63, 83]. We also drew insights from online content moderation research [102], as coping with anonymous negative feedback from a group of students shares similarities with mitigating online hate speech from a group of users.

Remove [1] This strategy removes negative feedback, serving as a baseline. We took inspiration from the removal of online hate speech [115], which is one of the most direct and effective content moderation strategy to reduce harm caused by hate speech. This mimics moderation strategies in online spaces, where messages are removed when they do not adhere to a community's guidelines or rules [102].

Sandwich This is one of the most widely recognized feedback methods [69], involves strategically placing negative comments between positive ones [29, 30, 48]. By cushioning criticism with positive feedback, this technique aims to enhance receptivity to areas of improvement [95, 103]. The Sandwich method leverages the psychological importance of framing and sequence to create a balanced and supportive feedback experience.

Paraphrased This strategy reframes negative feedback more positively and succinctly, without adding new content. Drawn from the use of mitigating language, which is a common techinque used by reviewers [57, 84] and also a

[1] Throughout this paper, we use purple text to highlight specific strategies and presentations in our design space, helping readers easily identify these key elements in our discussion.

type of affective language [83] that has been shown to enhance writing performance [110]. Mitigating language can improve the reviewer's perceived likability, increasing the likelihood of feedback implementation [84].

Constructive This approach goes beyond paraphrasing negative feedback by adding new, actionable content in the form of explicit solutions. Grounded in another desirable feedback characteristic of "offering a solution" [83], this strategy involves providing concrete suggestions to address identified problems [11, 105].

3.2 Presentation Designs to enhance feedback processing

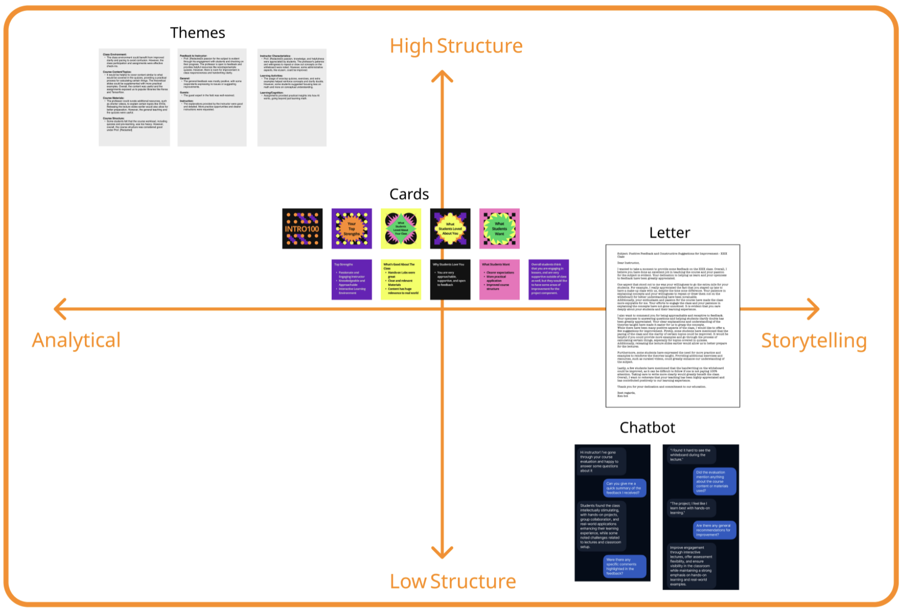


Fig. 1. Illustration of the design space for feedback presentation, with the four presentation designs for SET data mapped onto the two dimensions. The vertical axis represents the degree of structure, ranging from high (top) to low (bottom). The horizontal axis represents the approach to information presentation, ranging from analytical (left) to storytelling (right). The four designs - Themes, Cards, Letter, and Chatbot - are positioned according to their characteristics within this conceptual framework.

To move beyond the traditional static report formats, we also explored four presentation designs to scaffold feedback processing. This exploration is grounded in two fundamental dimensions of information presentation: the degree of structure and the balance between analytical and narrative approaches, as shown in Figure 1 as the two axes. The degree of structure axis aims to bring organization to feedback, built on prior work providing structure to crowd feedback [32, 130]. In addition, we added the second axis, which is grounded in foundational work in cognitive psychology, particularly the ideas of primary modes of thought: argumentation (propositional thinking) and storytelling (narrative thinking) proposed by Bruner [17]. Each mode offers a distinct means of organizing experience and has its own criteria for effectiveness. While propositional thinking aim to convince through truth and characterized by its deductive nature, storytelling seek to persuade through lifelikeness and leverages imagination [49].

Themes This approach groups student comments into categories, inspired by thematic analysis techniques [22]. Each category is represented by a summary generated from its grouped comments. Drawing from our experiences with experienced teaching consultants who categorize instructor feedback for improved interpretability, this format offers the highest degree of structure among our explored designs.

Cards This "bite-sized" approach presents hyper-summarized information. Each card is designed to be quickly digestible, readable within 10 seconds. This format is inspired by card components used in graphic design and web apps (e.g., [51, 122]). It also leverages the proven benefits of cards in supporting ideation [43] and fostering creativity [75], which are useful for feedback processing.

Letter This long-form narrative version of SETs imagines feedback presented as if written by a student representative or trusted colleague. Mimicking the style of an appreciation letter (e.g., [25]), it includes a greeting, body, and a closing signed by students. This approach is inspired by research showing that narrative formats can scaffold information processing [19, 31] and tend to elicit stronger positive affect and emotional responses.

Chatbot This design envisions a chatbot trained on SET remarks, allowing instructors to interact with an AI-based understanding of student feedback (e.g., [87]). Chatbots offer versatility in supporting various purposes, including informational and companionship roles [68]. In this context, the chatbot could reframe or rephrase ideas from the SETs and provide practice improvement suggestions based on its interpretation of student comments, offering flexibility to meet diverse instructor needs when engaging with feedback.

3.3 Automated Generation of New SET Artefacts

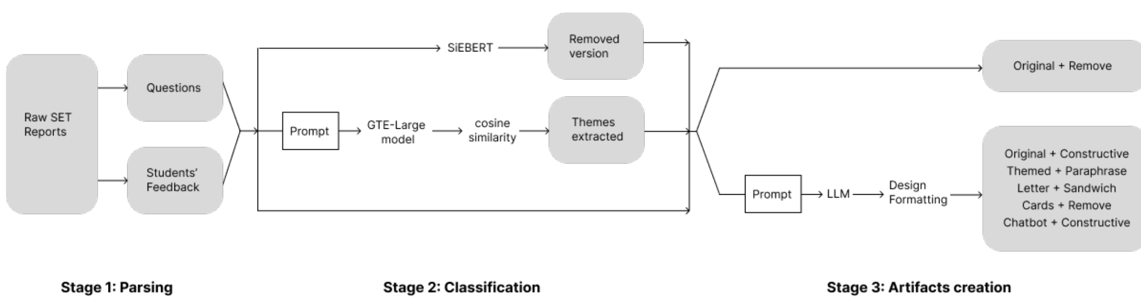


Fig. 2. Illustration of automated generation pipeline as described in Section 3.3

To ensure that our design ideas were viable, we designed a tool to generate the static artefacts (described above) based on real-world SET reports from two institutions. This tool would take, as input, raw SET reports from the authors' respective institutions, and could transform this data in to several distinct static SET report types (themes, cards, letter), along with different strategies (remove, paraphrased, sandwich, and constructive). This tool used several NLP techniques, such as text classification and summarization, and was built on Streamlit. Figure 2 illustrates the workflow for generating these mock-ups. First, the tool extracts the qualitative feedback along with the corresponding questions by parsing the original SET document, then filters the document based on known document structures of each University's Student Evaluation of Teachers (SET).

We used two different techniques for approaches that required classification. To remove negative feedback, we used the SiEBERT model [44] to perform sentiment analysis for each line of feedback and removed those that were classified to have negative sentiment. To classify the feedback against themes, we used a zero-shot approach, and generated text embeddings for descriptions of each theme, and and each line of feedback using the GTE-Large model [71]. We then used the cosine similarity of the text embeddings of each line of feedback against the respective themes, and chose the theme that had the highest cosine similarity as the class assigned to the line of feedback. To create the artefacts themselves, we used the pre-processed data combined with a specialized prompts for each of the report types using a general-purpose large language model (OpenAI gpt-3.5-turbo and gpt-3.5-turbo-16k). The code for this system is available[2].

4 INTERVIEW STUDY

To understand how SETs can be redesigned to support instructors' information and emotional needs, we conducted an interview study with 16 instructors. While there have been several recent attempts to build tools to help analyze SETs (e.g. [40, 52, 88]), their design was not fundamentally informed by instructors' practices. Under our main research question (RQ): **How can we redesign SETs to support instructors in their goals to gather practical and useful information from their SETs while minimizing the impact of distracting and unhelpful commentary**, we designed our study with two primary sub-RQs:

- **RQ1**: How, when, and why do instructors currently engage with their SET Reports?
- **RQ2**: How do the functional and interactive design space (elaborated in section 3) resonate with instructors' needs?

Our goal here was not to derive a "final correct design" for such tools, but rather to understand the requirements for such tools—what are the foundational needs of instructors, and how can these needs be addressed through algorithmic (i.e. extraction, summarization, etc.) or interactive design.

4.1 Participants

We recruited 16 post-secondary instructors (11 men, 5 women) from four universities through word of mouth and social media (Table 1). Of these participants, two were student instructors (graduate students who were Instruct of Record for a course), two were guest instructors (having a full-time job in industry), six were teaching professors (primary role is teaching), four were pre-tenure professors, and two were tenured professors. Participants are primarily teaching in STEM and social science fields (e.g., chemistry, engineering, computer science, design).

4.2 Method

We conducted 60 minute interviews either in-person or over Zoom. Prior to meeting participants provided us with a recent SET that we used to generate customized mock-ups of the the SET redesigns illustrated in Section 3. In the first part (~10 minutes), we explored participants' current practices and impressions of SET Reports, focusing on RQ1. The second part focused on participants' reactions and impressions of the redesigned SET mock-ups, corresponding to RQ2.

Part I. In exploring participants' practices and impressions with SETs, we focused our questions on how and when participants engaged with their SETs. We explored how the designs aided or hindered them in finding information, and what kinds of information they liked to see, as what kinds of information they did not want to see. In particular, we elicited how they currently dealt with and processed negative comments.

Part II. The majority of the interview was dedicated to exploring the various SET redesign concepts with our participants. For each design that we presented, we asked about their immediate reactions, and probed the ways that the

[2] Anonymized

Participant	Rank	Gender
p1	teaching professor	m
p2	teaching professor	f
p3	student instructor	m
p4	guest instructor	f
p5	tenure-track professor (tenured)	m
p6	teaching professor	m
p7	student instructor	m
p8	tenure-track professor (pre)	f
p9	tenure-track professor (tenured)	m
p10	teaching professor	m
p11	teaching professor	f
p12	tenure-track professor (pre)	m
p13	tenure-track professor (pre)	m
p14	teaching professor	m
p15	tenure-track professor (pre)	m
p16	instructor	f

Table 1. Participant Information. Information on their current occupation and gender.

		Strategies				
		Control	Remove	Paraphrased	Sandwich	Constructive
Presentations	Original	x	x			x
	Themes			x		
	Letter				x	
	Cards		x			
	Chatbot	x		x		x

Table 2. We characterize our design exploration along two design dimensions: *Presentation* and *Strategies*. We mark with an *x* locations in this design space that we generated a mockup that was shown to participants in our study.

designs fit or did not fit with their practices and information needs. Finally, we asked participants to rank the different designs, in part to provide a summative, comparative assessment of the different designs and strategies.

4.3 Materials

We generated customized mock-ups of our designs on a per participant basis based on the SET that they provided us prior to the interview. We presented a subset of the possible combinations of strategies and presentations in our design space (illustrated in Table 2) to provide participants with a broad sampling of the potential design space. Our intention was to not be exhaustive, but rather deliberate so that participants could experience the different ideas in multiple ways.

We illustrate two examples of these in Figure 3 and Figure 4 (see the other generated mock-ups in Appendix B). Figure 3 illustrates the Cards + Remove condition. Here, we presented this image to our participants with the contents tailored with the inputs from their evaluations. We presented these as cards and asked them to look through it, where the top left card as the initial and the bottom right card would be the final card. In the middle, the top card is representative of the main idea , and is further elaborated in the back, which is shown through the bottom card. In this specific design, we utilized the removal of negative feedback, which is why there are no cards dedicated to “Top Weaknesses” or “What students disliked.”

Fig. 3. Cards + Remove. This is a sample mock-up for the presentation condition Cards, with the strategy of removing negative feedback. It illustrates a short "slideshow" of six slides, where the front slide flips to show the back side of the card before moving to the next card.

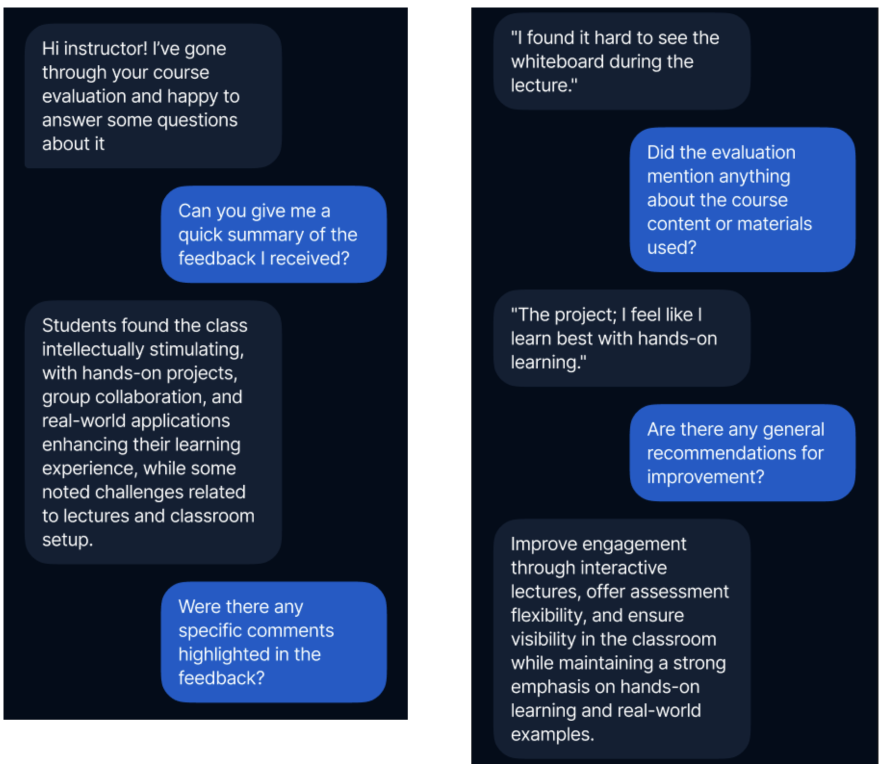


Fig. 4. Chatbot + Constructive. This is a sample mock-up for the chatbot presentation

Figure 4 illustrates the Chatbot + Constructive condition. Here participants were shown an example interaction with the system (participants did not interact with the system directly). The specific interaction in this screenshot highlights three features of the chatbot: (i) its ability to summarize and paraphrase the comments in the SET; (ii) extraction of quotes from the raw data, and (iii) providing constructive, actionable recommendations for teaching improvement.

4.4 Analyses

We used reflexive thematic analysis (RTA) to guide our data analysis. Braun & Clarke describe RTA as a theoretically flexible method for analyzing and interpreting patterns across a qualitative dataset [22]. This approach acknowledges that the researcher's position and contribution is a necessary and important part of the process, emphasizing the term 'reflexive': as researchers, we draw from our own experiences, pre-existing knowledge, and social position to critically interrogate how these aspects influence and contribute to the research process and potential insights into qualitative data [22].

Three of our co-authors have had experience teaching in post-secondary institutions, and have received and read SET reports. Two of these co-authors are tenured professors, collectively with 26 years of teaching experience. All four co-authors have had experience preparing remarks and comments as students for SET reports. As researchers, we operate at the intersection of HCI and NLP: thus, we are well-versed with HCI techniques and take a user-centered design orientation to the problem. We are informed by our working understanding of NLP techniques (both in terms of practical know-how, as well as near future capabilities of NLP tools). These experiences inform and shape how we conceptualized this work, and therefore how we analyzed our data.

The interviews were transcribed by otterai [1] with the authors correcting any misspellings or misunderstandings of the system. We then open coded interview transcripts using Google Sheets [2], and developed potential themes through an iterative process of clustering and grouping codes on Miro [3]. Through iterative discussion of codes, participant quotes, and potential themes, we developed our candidate sets of themes. As we wrote this paper, the candidate themes evolved to final themes, and we report on salient themes that reflect our position as HCI researchers and instructors.

To complement our thematic analysis and provide an overview of participants' experiences with feedback and perceptions of our design mock-up, we synthesized the data into visual representations. Figure 5a illustrates reactions to design mock-ups and rankings of presentations and strategies. Additionally, we conducted a post-hoc analysis of participants' experiences with different types of student feedback, resulting in the typology presented in Table 3.

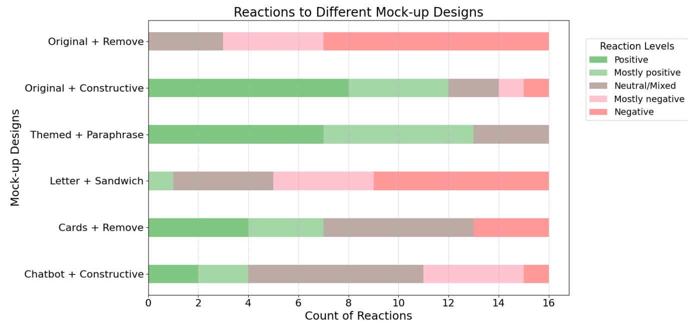
5 FINDINGS

Our analysis of participant reactions and rankings reveals the complexity of preferences for SET redesigns. Figure 5a shows the Original + Constructive and Themes + Paraphrased design received very positive reactions, while Original + Remove and Letter + Sandwich were viewed less favorably. Table 5b shows similar patterns. Interestingly, several design options received both the highest (1) and lowest (5) rankings, suggesting polarized opinions among participants. Moreover, most presentations and strategies were ranked first by at least one participant, indicating that each resonated strongly with some instructors, despite variations in overall rankings. While these preliminary analyses provide a high-level overview of participants' preferences, our thematic analysis of the findings reveal the underlying reasons and fundamental needs beneath these summaries.

Through our analysis, we identified four key challenges instructors face when engaging with student feedback and the design space: forming actions, emotional impact, trust in AI-assisted processing, and longitudinal engagement. By probing participants with design mock-ups of strategies (Remove, Constructive, Paraphrased, Sandwich) and presentation (Themes, Letter, Cards, Chatbot), we identified opportunities for addressing these challenges. These findings address our main RQ on how to redesign SETs to support instructors processing feedback.

Type of feedback	Elaboration	Participants	Associated keywords	Specific examples
Actionable and constructive	Specific feedback that can be directly addressed or implemented	P1, P2, P5, P6, P15	Helpful, actionable, specific, easy to fix, constructive	"The work was too much weighted towards the last part of the quarter" "The readings really didn't relate to the assignments" "Deadlines are confusing" "Not enough lecture, not enough coverage of this technique or that technique"
Insightful but challenging to implement	Feedback that identifies real issues but may be difficult to address	P2, P4, P5, P7, P9, P15	Critical, heavy, challenging	Requests for more exercises or content when the course is already full Comments about heavy workload that might be necessary for the course "I didn't enjoy the social justice oriented readings" "Be more confident" or "Speak louder"
Contradictory or inconsistent	Feedback that conflicts with other comments or itself	P1, P4, P6	Incongruent	"There was too much freeform time to work on projects" vs. "There wasn't enough time" Some students praising hybrid format while others wanting more in-person classes
Emotionally charged	Highly emotional feedback that may obscure the actual issue	P11, P12, P14	Negative, sad, pissy, roasting	Long paragraphs of extremely negative feedback on all aspects of the course Feedback from students who got into conflicts with the instructor over grades or policies
Vague or non-specific	General complaints without clear suggestions for improvement	P4, P6, P11	Weird, whining	"This class totally sucks"
Feedback on factors beyond instructor control	Comments on aspects the instructor can't directly change	P5, P15	Not within control	Complaints about the amount of content in standardized courses Comments about the classroom or technology issues
Factually incorrect	Feedback based on misconceptions or false information	P11	Blatantly untrue	"No one in the real world actually writes code anymore. They just write apps and use extensions"
Biased or discriminatory	Comments reflecting prejudices (e.g., gender bias)	P4	Harsh, underappreciation	Underappreciation of expertise, particularly for women teaching in technical fields
Personal or ad hominem	Comments targeting the instructor's personal characteristics rather than teaching	P1, P2, P4, P7, P8	Useless, cheeky, flattering, harmful, rude, unrealistic	Comments about the instructor's appearance or clothing and language skills "You have no business writing in English because your English is so broken"

Table 3. Typology of Student Feedback Based on Instructors' Experiences. Column descriptions: (1) Type of feedback, (2) Elaboration on the feedback type, (3) Participants who mentioned this, (4) Associated keywords reflecting participants' subjective characterizations and emotional responses to these feedback types, and (5) Specific examples provided by participants. This typology is classified based on the nature and perceived impact of the feedback, offering insight into their personal experiences and perceptions.



(a) **Comparison of participants' reactions to 6 design mock-ups presented in the study.** Each horizontal bar represents the distribution of reactions to a design mock-up, with the length of colored segments indicating the number of instructors (out of 16) whose reactions were categorized on a scale from positive to negative.

Item	Mean	Max	Min
Presentations			
Themes	1.63	3	1
Cards	2.63	5	1
Chatbot	3.44	5	1
Original	3.19	4	1
Letter	4.25	5	2
Strategies			
Constructive	2.00	3	1
Paraphrase	2.19	4	1
None	2.81	5	1
Sandwich	4.06	5	3
Remove	4.13	5	3

(b) Mean, Maximum, and Minimum Rankings for Presentations and Strategies

Fig. 5. Comparison of design mock-up reactions and rankings for presentations and strategies

5.1 Beyond Summarization: Forming actions from feedback

We found that the quantitative portion of evaluations often fails to provide clear guidance for improvement, as participants struggle to interpret the meaning behind scores and translate numerical ratings into actionable changes (P3, P6, P10). In contrast, the qualitative component offers more valuable insights that can lead to concrete actions, aligning with the primary goal of many participants in engaging with student feedback. As P3 encapsulated, "*My primary goal for reading course evals is to just see what is actionable.*"

However, several issues hinder the effective utilization of open-ended question responses. Firstly, participants describe the standardized questions as restrictive, irrelevant, or ill-suited to their specific courses (P2, P7, P10). P10 illustrated this point: "*Sometimes, the question was ·[Was] the course intellectually stimulating or stretch your thinking,' [but] sometimes the basis of a course is not intellectual stimulation. It's practical skills.*" Furthermore, as instructors gain experience and develop a clearer understanding of students' perspectives, the value of certain questions diminishes, leading to a saturation of insights over time (P6).

Moreover, the nature of unmediated, anonymous student feedback presents additional challenges. Table 3 presents an empirically-derived typology to illustrate the diverse nature of students' feedback. While some categories, such as "*Factually incorrect*" or "*Feedback on factors beyond instructor control,*" are readily identified as less useful, many others require careful consideration. This categorization reflects instructors' own definitions of usefulness and challenges in processing student feedback. Instructors often start with skimming through feedback and identify the negative feedback for potential issues (P10, P8, P16). However, this approach can lead to difficulties in distinguishing sincere concerns from disgruntled students' remarks (P5). Participants also encounter unconstructive comments and feedback on factors beyond their control (P11, P3, P4, P5). In larger classes, the sheer volume of feedback can be overwhelming (P1, P5), while conflicting student opinions complicate interpretation and decision-making regarding necessary changes (P1, P4, P6, P7, P10, P14).

Given the complexity of raw feedback, instructors must carefully sift through responses to identify substantive concerns that warrant changes in teaching approach or style (P10). Some employ personal annotation strategies, such as underlining key points, marking noteworthy comments, and tallying recurring issues (P11). Yet, the unguided process remains complicated, unguided and demands significant manual effort, compounded by the standardized nature of evaluation questions and the inherent variability of student feedback, hindering the efficient translation of feedback into actionable improvements in course design and delivery.

5.1.1 *Focused view and prioritization.*

The Themes presentation, which groups similar feedback based on predefined topical themes (introduced in Section 3.2) informed by practices, was well-received among participants. Themes affords a focused view on certain issues and help instructors prioritize issues that they want to work on. It allows instructors to tease apart different types of feedback (P1) and help to see the big picture to avoid "*getting caught in the weeds*" and overemphasis on the individual comments that might only apply to one specific course (P8). Moreover, this approach also assists in dealing with conflicting feedback by grouping different opinions under the same type of problems and synthesizing the divergence towards a more generalized solution instead of over focusing on individual opinions (P3). In addition to the topical themes we've provided in the mockups, some participants also expressed interest in seeing a breakdown and analysis of sentiment and attitudes (P2, P4, P5). P5 was concerned of changing the problems might affect things that they've already doing well, so knowing to what extent they liked certain aspects of the class would help with the actions.

In contrast, participants generally reacted negatively (Figure 5a) to the Remove strategy for two main reasons. First, they feared that removing all negative feedback might inadvertently filter out important concerns that could serve as a valuable source of action. Second, they felt that the lack of a more granular classification of what constitutes "negative" feedback could lead to the removal of potentially useful information. In addition, Letter revealed that participants prefer more structured and concise presentation when forming actions. This suggests that while instructors value the insights provided by negative feedback, they also appreciate having the information organized in a way to efficiently facilitates actionable next steps.

Furthermore, participants' reactions to Chatbot emphasized the importance of guided interaction to maintain focus and reduce cognitive load, allowing instructors to concentrate more fully on processing the feedback itself. Participants expressed concern that generating questions independently would be cognitively demanding (P4, P3) and might lead to overlooking crucial issues (P7). They viewed predefined questions as a means to ensure consistent information access across instructors (P6). The Chatbot should offer both general questions applicable to all instructors (P7, P1, P3) and context-specific queries tailored to individual evaluation reports (P5, P8). Examples of general questions included asking about changes in course evaluations over time (P1), the clarity of lectures (P5), or summarize key points in past SET reports from previous teaching (P14). Context-dependent questions could involve locating specific student feedback (P8) or identifying particularly problematic assignments or readings (P2, P3).

5.1.2 *Provide additional perspectives to encourage divergent thinking.*

We found that the strategies and presentations afford perspective shift and divergent thinking. Design strategies like Paraphrased and Constructive that directly modifies the feedback content offered a more distanced perspective that enable instructors to break free from their established patterns of thinking (P2, P8, P11). P2 noted its value "*for instructors who have taught the class for a long time by providing a fresh perspective and making them 'see the forest from the trees'*". In addition to the affordances brought by direct content manipulation, the Chatbot presentation affords divergent thinking through targeted questioning and assisted-ideation, encouraging active solution-seeking. P5 envisioned asking "*very specific questions about what could I do differently or how could I improve learning in the classroom,*" while P3 saw the chatbot as a tool for exploring various ways to enhance the course by asking "*'what if' type of questions to brainstorm ideas for upcoming classes.*" Moreover, our design probes revealed the potential for interventions to foster curiosity-driven exploration, complementing instructors' judgement. While P9 envisioned their use of Chatbot: "*Here's what I know from reading the evals but what does the system think?*" This approach could encourage consideration of diverse perspectives while not replacing instructors' personal insights.

5.1.3 Tailoring Feedback Specificity to Instructors' Information Processing Needs. The desired level of specificity in the feedback depended on whether instructors were at the stage to get implications or form concrete action plans. On the one hand, when the Paraphrased feedback was less specific, participants benefit from getting inspirations from it. P7 appreciated how the feedback "*gives at least a start of where to go,*" helping them to contemplate the next steps themselves. Too detailed actionable item would make it appear "*prescriptive*"(P8). P8 explained, "*I want to have something to get me thinking as I move forward with my planning. Doesn't have to be specific.*" On the other hand, the more direct and detailed suggestions, like those provided by the Constructive feedback strategy, are particularly effective in bridging problems and solutions (P1, P2, P5). P5 appreciated the specificity and directness of the constructive feedback, stating, "*I appreciate when things are just very to the point... it's helpful that it's in a more positive light... it makes it a little easier to think about what I could be doing differently.*" P1 concurred, highlighting the practical value of the suggestions in facilitating iteration based on the feedback received.

5.2 Emotional benefits and celebration

Engaging with student feedback evokes a range of emotions for instructors. While some comments are affirming and motivating, others can be personally hurtful, biased, or emotionally taxing, acting as a significant barrier to processing evaluations (P8, P12, P13, P14, P15). The "Emotionally charged" and "Personal or ad hominem" types of negative feedback in Table 3 are particularly challenging to handle. Personal negative comments directed at the instructor rather than the course are especially challenging to handle (P8). Even without overtly harmful comments, the prospect of reading critical feedback induces anxiety and stress, particularly for inexperienced instructors (P6, P10, P14). This emotional toll can linger, affecting future interactions with students (P12) and be amplified in environments emphasizing teaching excellence (P15). The impact of negative feedback is disproportionate. As P11 pointed out, "*even if 90% of the comments are really nice, and like everything was working great. The ones that sting will sting a lot.*" This highlights the negative bias in feedback processing, where negative comments tend to carry more weight and emotional impact than positive ones, regardless of their relative frequency.

Awareness of potential harm discourages some from engaging with evaluations altogether, creating a tension - while they may find some value in the feedback, the emotional cost of sifting through negative comments often exceeds the perceived benefit of extracting new information (P2). However, experienced instructors develop resilience to criticism over time (P6, P10, P14), with some adopting a mental model of feedback as informative for improvement (P10). Collaborative approaches, such as having peers review evaluations together (P2, P6, P8, P11), provide emotional support and a "buffer" for harsh comments (P8). Despite the challenges, instructors also derive emotional benefits from feeling reconnected to their students' voices and experiences through the raw feedback (P9, P14), appreciating the personal connection it fosters (P9).

The emotional experience is not solely negative. P14 mentioned, "*the thing that interests me the most on course feedback...I like to read nice things once in a while.*" Beyond complimentary comments, P9 feels a "*personal connection*" to their students as individuals through reading the evaluations and appreciates maintaining "*as close a relationship as I can to the learners in my classroom*" through the raw student feedback. These positive sentiments reveal how instructors can derive an emotional benefit from feeling reconnected to their students' voices and experiences when reviewing evaluations, not just focusing on areas for improvement.

5.2.1 Reducing negativity. Some of our designs aimed to highlight the positive aspects of the teaching feedback. From participants' reactions, we found value in having the design not only reframe the evaluation processing experience by reducing negativity and stress, but also help function as means of celebration and sharing. Design strategies could encourage engagment through visual appeal and reduced stress for negativity. Cards stood out as being visually more engaging than the traditional SET reports format (P1, P3, P6, P7). P1 found it "*a lot more visually interesting than the traditional format*", and P7 also noted that visuals can potential stick along longer than the actual phrases. On the other hand, P5's less positive response to the colorful themes showed that visual preferences vary among instructors.

In addition to visual appeal, several strategies addressing negative feedback showed potential to make it more acceptable and less stressful. The Remove strategy increased P2's willingness to engage with feedback overall, "*I would be more likely to engage, or at least open it. Right now, ..., it's just their space to rant.*" These designs also addressed the initial exposure to feedback, which some participants (P8, P14) identified as the most stressful moment. P14 appreciated the Remove for offering a choice to view the full report later, contrasting it with the "*apprehension*" they typically felt when first viewing traditional SET reports . The Chatbot design similarly appealed to P14 for enabling gradual engagement with feedback, particularly when anticipating negative comments. The Sandwich approach also showed promise in reducing stress. P3 found it potentially less stressful to read, while P1 noted its ability to balance negative perceptions of the class. However, participants emphasized the importance of user control in these designs. P4 suggested an opt-in feature for removing negative comments, similar to Twitter's sensitive content warnings, allowing users to reveal them if desired.

5.2.2 Reframing as celebration and encouraging sharing. In addition to visual appeal and emotional benefits that could encourage engagement, some participants also call out the benefits of celebration and sharing. While participants perceive certain presentations as less effective as Themes for thinking through the feedback and forming actions (P7, P8, P12), they still acknowledge the celebratory benefits of them to serve as a complementary form of positive reinforcement. Specifically, P8 and P12 commented that Cards makes them feel good about teaching. Similarly, P7 explained the Letter that "*the appeal of it coming from my students collectively that there's I wrote it is a nice touch. It feels a little bit more warm.*" Moreover, Cards + Remove combination was called out by multiple people to be shareable (P1, P11). The current evaluations contain negative comments that require hedging when sharing, whereas a celebratory design "*would result in more sharing between colleagues or other instructors*" compared to the original format where "*it's a lot easier to share this kind of stuff than all the then having to like caveat with like, hey, number three, pretty sure I know who this is. They were just really pissy*" (P11).

5.3 Building trust in AI

While many participants appreciated the potential benefits of leveraging AI to process student evaluations, some expressed hesitation and skepticism towards certain AI-driven approaches (P2, P6, P12). This reluctance stemmed from various factors, including preconceived negativity based on personal experiences. P6, reflecting on their own prior encounters with AI, noted, "*We were so skeptical of chatbots and understanding of how they're constructed and how they work and how unreliable they are. So definitely a negative reaction immediately.*"

Another main concern was about mischaracterization of the original comments. Participants worried that AI summaries could fail to capture the true voice and intent of what students wrote. As P16 stated, there was discomfort with the summaries feeling "*too distant from the text that the students wrote...It's still not the student's voice.*" This unease was rooted in skepticism about the capabilities of generative AI, with P12 expressing wariness about "*the susceptibility to hallucinations*" and models "*conjuring up connections where those connections don't really exist.*" However, some instructors like P14 were less troubled by obvious hallucinations they could easily identify as nonsensical, like suggesting to teach non-existent classes or completely irrelevant topics. If an AI made a surprising suggestion, P14 reported that they would be alert and fact-check the sentiment against the original student comments.

Some participants expressed concerns that AI-powered tools may not provide sufficiently high quality or creative suggestions, as how P10 questioned whether there was "*real intelligence*" behind the AI's outputs. Similarly, P4 and P6 are concerned about some of the AI-generated action items from feedback being too generic and vague to be useful. Instructors value the time and effort they invest in carefully considering student feedback and crafting innovative solution—a process some fear could be short-circuited by an over-reliance on AI. As P4 noted, if instructors defaulted to AI-generated solutions without pushing themselves to come up with additional creative ideas, it could result in less thoughtful engagement and changes.

5.3.1 Retain access to context. Some of the design probes further reveal the underlying fear of losing the original context and how the use of AI should be complemented with a source of truth. For example, P1 expressed concern that the Remove strategy might overhype the positive reactions, while P4 felt that the Letter format was too "*proper*" and didn't feel like it was actually coming from the students. Moreover, almost all participants explicitly stated their need for access to the original, raw feedback. As P12 explained, "*When you just remove the data, I think that it removes potentially useful information.*" P4 worried that nuances may be lost, and P10 was concerned that emotional and personal expressions would not be retained in the paraphrased version. P3 articulated this sentiment, saying, "*I guess it's less about potentially missing out on information, but more on like the feeling of potentially missing out information.*" P11 noted that "*having the individual data points (original feedback) can help with validating the paraphrased feedback.*" Similarly, when the Chatbot made recommendations, P9 wanted to cross-reference them with the raw feedback, stating "*I would look at that and then I would go back to the [reviews] myself, and I'd say okay, is this actually accurate or not?*"

5.3.2 Transparency in the "who" and the "how". As our study design did not explicitly specify the type of AI models used, participants raised questions about the transparency of the AI's inner workings and its capacity to understand the subtle meanings of student feedback. Some participants expressed concerns and curiosity about how certain feedback is chosen to be removed (P4, Remove), whether the Letter format consolidates input from all students (P5), and who generates the constructive content (P6, Constructive). When presented with the artifacts, P4 questioned the removal process, while P7 wondered, "*Where that's coming from who's doing that? Is it human moderated, is it algorithmically moderated?*" P11 also wondered, "*who is paraphrasing? are they doing it accurately?*"

Participants' skepticism extended to the AI's ability to fully grasp the subtleties and severity of feedback without the additional context that human instructors possess. P6 doubted the Chatbot's capacity to distinguish between serious critiques and one-off complaints, stating, "*I think it's unlikely to know when a student says this was the worst class I've ever taken, should I take that really seriously... versus this was the worst class I've ever taken, but their other comments suggest that maybe it was just a one off bad experience.*"

In contrast, participants placed greater trust in experienced teaching consultants, who have "*the same type of experience and proven track record of success*" (P15) and possess the relevant knowledge and experience to interpret feedback holistically (P7, P10). P2 described her trusted consultant's practice as "*magic*", while P12 expressed "*high,* *almost blind trust*" in a consultant's expertise, stating, "*if [xx] told me to go march into the ocean, i would probably consider it.*"

5.4 Longitudinal use

Our findings reveal that instructors' use of SET reports often extends beyond the initial reception of feedback. They revisit SET reports for various purposes, such as extracting quotes for annual reviews (P14), identifying trends, or determining common issues to inform course changes (P1, P12, P16). However, the current form of SET reports is a static document provided at the end of the quarter without any support for long-term usage.

Access to comprehensive historical SET reports data could provide valuable insights by enabling instructors to track trends over time and assess the criticality of issues (P1, P12, P16). P1 noted that it takes teaching a class multiple times helps identify trending feedback and inform changes. P16's experience further exemplifies this value: after receiving feedback about mumbling for two consecutive terms, they implemented a proactive strategy based on a colleague's advice, effectively resolving the issue in subsequent evaluations. To address the challenges of long-term recall and implementation, participants have developed personal systems for longitudinal reference. P4 maintains a "*master document*" with a running list of desired tweaks for the class, incorporating SET feedback into this ongoing record. Similarly, P6 refers to a reflection form resulted from their departmental annual review process that contains actionable points for future improvements.

Instructors emphasized the importance of collaborative efforts and iterative processes in driving structural changes to courses and curricula. P1 noted the value of discussing open-ended feedback with TAs in their current practices. Moreover, P2 highlighted the value of using feedback from multiple cohorts of students to inform significant changes, sharing an example of how multi-year feedback led to a major curriculum revision: "*It's only through really digging in with multiple cohorts of students that we have, we've come to this. ... It was like a collective process.*" The importance of longitudinal analysis extends to administrative purposes as well. Deans and department chairs review faculty members' course evaluations during annual performance reviews (P6), suggesting potential for processed SET reports data to streamline this task. Moreover, there's interest in comparing SET reports data across similar institutions over time to gain broader insights (P11). The need for multi-stakeholder involvement extends the purpose of sharing beyond celebration as described in Section 5.2.2 to encompass collaborative improvement efforts at various levels of academic organization.

5.4.1 Contextualization and recontextualization. Our design probes revealed how processed versions of SET reports can facilitate both contextualization for stakeholders and recontextualization for instructors over time.

Recontextualization for instructors emerged as a key benefit. P2 expressed interest in revisiting evaluations through Themes + Paraphrased formats when teaching the course again, potentially a year later. This suggests these formats preserve context more effectively than raw data. P14 noted that accurately processed information with actionable insights could eliminate the need to refer back to raw data, indicating a shift in how instructors might interact with feedback over time. P4 highlighted the challenge of recall, expressing interest in a system that reminds them of key takeaways from previous evaluations and facilitates historical data-based brainstorming.

In addition to self-referential usage, participants also reported needs in material to help other stakeholders more easily contextualize the feedback. The anonymity and structure enhanced by Paraphrased or Themes strategies could result in greater comfort with sharing (P3, P8). P3 speculated that with access to such processed data, "*maybe more people, hypothetically, might be willing to share course. Share and have like brainstorming sessions for how to address* *things. We don't necessarily have that happening now."* In addition, presentations and strategies that reformat and extract the information at a higher level of abstraction were found to be easier to be shared with new instructors teaching the same course (P2, Themes + Paraphrased), shared at the departmental level (P8, Themes), share as a reference letter on behalf of students from that class (P1, Letter). This increased shareability stems from the processed formats' ability to distill and synthesize key insights from all the qualitative data while preserving privacy.

6 DISCUSSION

Despite the value of student evaluations of teaching (SETs), instructors face barriers in engagement that leads to underutilization [96]. While advanced NLP techniques, such as large language models (LLMs), show promise in addressing some limitations, we argue that simply applying them can fall short due to the intricate nature of feedback, the complex relationship between instructors and students, and the ways in which feedback is utilized. These challenges echo observations in other contexts where users interpret LLM-generated outputs, even for tasks that are more structured and rule-based [41, 132]. Effective solutions require thoughtful design and a deep understanding of instructors' needs and usage patterns. Our experimentation with AI-enhanced SETs, using various presentations (Themes, Letter, Cards, Chatbot) and strategies (Remove, Paraphrased, Constructive, Sandwich) uncovered key insights around supporting action formation, mitigating emotional burdens, reframing feedback positively, fostering trust, and considering temporal elements. In the following discussion, we explore the implications of these findings for system designers from a broader perspective.

6.1 AI as a first pass through the feedback

Aligned with prior literature [6, 14, 23, 50, 61, 66, 67, 119, 127], our participants also reported significant emotional and cognitive costs when dealing with processing SET reports. Importantly, our findings reveal how different forms of redesigns provide various affordances to mitigate these burden. We further discuss implications of these insights and opportunities provided by AI.

6.1.1 Reduce emotional cost and bring emotional benefits.

AI as an emotional buffer. One key insight from our study is that instructors desire a way to reduce exposure to overtly negative or harmful comments, especially upon first receiving the feedback. Hence, we could leverage AI's enhanced capabilities in text classification to detect these instances, particularly sentiment analysis [80], which has shown promise in parallel context of online hate speech detection [20, 46, 70]. Beyond overtly harmful feedback, other types of negative feedback can still induce emotional cost and hinder further actions. Studies suggest that the extent to which people value and follow feedback depend on how it is expressed [83, 84], with positive affective language increasing positive emotions and work quality compared to critiques without it [85]. Our findings around the Paraphrased and Sandwich strategies demonstrate that AI can be leveraged by retaining the essence of feedback while making it more palatable. Future work can explore more specific types of paraphrasing and tonal adjustments, as LLMs offers novel use cases in switching tones [106, 132].

It is important to note that albeit the similarities with content moderation, the instructor-student relationship differs significantly from that of online content creators and commenters. Unlike impersonal, one-time exchanges in online communities, instructors and students develop familiarity over an entire semester. This extended interaction makes student feedback inherently more personal and impactful. A dichotomy emerges: harsh comments are more emotionally challenging for instructors due to this personal connection, yet they may contain valuable insights for teaching improvements. Our design probes also revealed this tension, with participants expressing concerns about direct AI-driven comment removal despite desiring emotional buffering. This finding underscores the need for a more nuanced approach. Rather than direct removal, system designers should first use AI-powered classifiers to flag and hide the potentially harmful comments from intial view. Then, LLMs can be leveraged to provide local explanations directly in natural language [97], even expressing nuances like uncertainties about its prediction [108, 125] to help instructors make informed decisions about how to engage with challenging feedback without confronting the unfiltered negativity.

6.1.2 *Reduce cognitive cost and support action formation.*

Categorization and quantification of feedback. Our findings highlight the challenge of identifying actionable items within unstructured feedback. Strategies like Themes, which provide clear internal structure and pattern visibility, prove particularly useful. Participants have individual conceptualization for useful feedback, and our categorization in Table 3 offers a starting point for assessing usefulness and actionability in feedback that captures instructors' nuanced mental models. LMs can significantly reduce manual work and cognitive effort through initial feedback clustering and grouping. The flexibility offered by few-shot learning [116] further avoids cost in tuning or training the model and enables instructors to create personalized AI classifiers with minimal examples (1-5 per class). Moreover, we have found the a need for quantitative insights from qualitative data to understand criticality. AI can assist by categorizing and quantifying feedback distribution, efficiently summarizing recurring themes and sentiments. This capability answers questions like "How many students found me unclear?" or "What percentage liked the materials?" The system can also generate on-demand visualizations [126], facilitating easier comprehension of overall sentiment and areas needing attention.

Balancing flexibility and best-practices. While many instructors criticize the current one-size-fits-all approach to question design and desire more tailored methods, they often find crafting their own questions burdensome. Participants voiced concerns about "having to think of questions to ask the chatbot," contrasting this with their experiences with human experts who guide them through the reflection process and highlight important aspects. This underscores the need for both context-dependent flexibility and predefined best practices to reduce interaction costs. Mirroring contemporary chatbot designs, these questions can be either text-dependent (as in many LLM-based chatbots) or standardized (common in rule-based chatbots). Offering a predefined list of questions based on student feedback best practices can guide instructors during their initial chatbot interactions. Simultaneously, the system can surface context-dependent concerns (e.g., outliers, recurring issues), prompting instructors to ask targeted, self-defined questions.

6.2 Fluid transition between different usages and purposes

Our study reveals how AI-powered redesigns can enable the integration and seamless movement between the summative and formative purposes of SETs for self-referential use, extending prior work that acknowledged these uses are not mutually exclusive [10, 15, 81]. Specifically, while both goals has the ultimate purpose of improving teaching, formative use refers to understanding the areas that need improvement and identifying actions, whereas summative use means examining feedback from an overview look to understand overall students' reactions and experiences. The use of AI techniques like unsupervised clustering (Themes) and query-answering (Chatbot) was crucial in creating dynamic, interactive interfaces supporting different levels of feedback analysis, from high-level summaries to targeted deep dives, enabling fluid transitions that traditional SETs do not support.

This fluid transition aligns with principles from crowdsourced design critique systems like Voyant [126], which uses coordinated views to provide a summarized visual overview while enabling inspection of specific explanations behind ratings. Similarly, teaching evals tools should provide a global data view for summative overviews that seamlessly linked to the underlying raw feedback comments, enabling smooth transitions between the overall summative understanding and the detailed individual feedback for formative exploration. The use of AI enables fluidity: For example, the use of AI in Themes provides high-level summaries and identifies the source quotes to allow drilling-down into specific remarks. For interpersonal summative purposes, which informs administrative decisions, our findings reveal a disconnect between the scoring and the complexity of actual students' experience. AI can be leveraged to contextualize the scores and provide a more holistic review of instructors' relationship with students and areas for growth. For example, score-comment alignment techniques can identify which aspects of qualitative comments correlate with specific ratings, while sentiment-score reconciliation compares sentiment in comments with numeric scores to highlight any discrepancies.

Also, embedding advanced language models into chatbots can handle different types of natural language queries seamlessly, letting instructors investigate feedback at different granularities and shift between summative and formative lenses, starting with overview queries like "*What were the most common issues?*" and drilling down with targeted follow-ups like "*What suggestions did students have for improving readings?*". Based on this, we envision SETs to be transformed into living documents that evolve over time, building a feedback database for AI to retrieve information upon queries for tasks like trend detection (e.g., Does this new assignment lead to better final grades?), surfaces persistent issues (e.g., Is there something I haven't fixed yet?), and allows for historical queries (e.g., What did students complain about last time I taught this course?).

6.3 Foster upward and iterative long-term mindset

While prior work emphasized on the stress and mental burden, our findings revealed that the participants still value the positive experience. It's not just removing the negatives, but also how to highlight the positives. Prior work adopting the Positive Psychology Framework [33] has identified upward and negative spirals among instructors, where proactive coping strategies and rational feedback processing lead to problem-solving, while blaming students leads to negativity [76]. Redesigning feedback can indeed support the formation of this upward spiral by highlighting the positives that can lead to more actions and changes. Our findings specifically suggest that design elements like color and visual design in the Cards presentation can promote a celebratory orientation. Inspired from the design of "Spotify Wrapped" that invites users to share their annual music listening habits [118], redesigning feedback into bite-sized formats could encourage sharing the positives. The goal of highlighting positive feedback align with the concept of celebratory technology, which assumes user competency and advocates for augmenting current practices by providing new ways to engage [39]. While some feedback demands immediate changes, much of it addresses non-binary aspects, and a positive mindset could encourage instructors to innovate and experiment with their teaching based on feedback. This celebratory perspective extends beyond traditional formative and summative uses of SETs. Future work should explore balancing celebratory and corrective uses of SETs, acknowledging the importance of both informing areas for improvement and encouraging a positive orientation that promotes experimentation and innovation in teaching.

Building on prior literature's notion that successful students view assessment as part of their larger development [15], we advocate for designs that facilitate instructors' engagement with SETs for long-term developmental use rather than treating feedback as a one-time event. Our redesigns, like Themes and Chatbot , suggest that structured archiving and retrieval can support feedback recontextualization. Easy access to these presentations enables quicker recall of key points, reducing memorization friction and facilitating action formation. Moroever, our study provided only a one-time exposure for participants to these redesign ideas. The successful implementation of such systems requires careful consideration of deployment strategies and user acceptance over time. Drawing from work on longitudinal trust formation and technology acceptance [45, 78], we recognize that repeated exposure and a step-by-step approach are crucial. Affirmation time and strategies to encourage users to try and understand new systems are essential for long-term adoption and effectiveness.

7 LIMITATIONS AND FUTURE WORK

Our study is subject to several deliberate scoping choices, which also open up various avenues for future research. First of all, our decision to use mock-ups and conduct single-session interviews enabled us to gather rich initial reactions and insights, rather than feedback contingent on specific design details. However, we acknowledge that longitudinal studies with functional prototypes could reveal how perceptions and usage patterns evolve over time. Moreover, we recruited instructors primarily from STEM and social science fields. While our findings center on fundamental experiences and needs that transcend specific domains, it's important to acknowledge that educators from certain backgrounds may have unique characteristics. For instance, STEM field instructors may particularly value succinct formats and efficiency. Future work could explore how these design guidelines may adapt to across different academic disciplines and teaching styles. Additionally, while we have some observations on instructors' teaching experiences and class sizes, future research could dive deeper into investigating how various SET designs might differentially support junior versus senior faculty, or those teaching small versus large classes. At the same time, as AI continues to advance, further exploration around factors like potential privacy and ethical considerations will be crucial for scaled deployment and usage. Lastly, while we focus on post-secondary education teaching feedback, there's potential to explore how these findings might apply in other feedback exchange contexts and educational settings, including K-12 education, professional training, or even peer-to-peer feedback systems.

8 CONCLUSION

The goal of our work is straightforward: to increase the benefit instructors receive when engaging with their SETs, and to reduce the cost of engaging with their SETs. In our explorations, we designed and implemented a system to create SETs that presented SET information differently and used different techniques to hide/filter/mask negative feedback. Based on our study with 16 instructors, we found that because instructors use SETs in different ways, it is important to provide this information in ways that effectively support their needs—whether it be to affirm their teaching practices and approach, or to collect formative feedback on their approaches to understand how to improve their practice. We found that there are exciting opportunities for applying NLP techniques to provide this type of feedback, and look forward to the day that we can also look at our SETs without a twinge of anxiety.

REFERENCES

[1] [n. d.]. https://otter.ai/
[2] [n. d.]. https://www.google.com/sheets/about/
[3] [n. d.]. https://miro.com/
[4] Frederik Anseel, Filip Lievens, and Eveline Schollaert. 2009. Reflection as a strategy to enhance task performance after feedback. *Organizational Behavior and Human Decision Processes* 110, 1 (2009), 23–35.
[5] Margarite A Arrighi and Judith C Young. 1987. Teacher perceptions about effective and successful teaching. *Journal of Teaching in Physical Education* 6, 2 (1987), 122–135.
[6] Linet Arthur. 2009. From performativity to professionalism: Lecturers' responses to student feedback. *Teaching in Higher Education* 14, 4 (2009), 441–454.

[7] Md Abul Bashar, Richi Nayak, and Nicolas Suzor. 2020. Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set. *Knowledge and Information Systems* 62, 10 (2020), 4029–4054.

[8] Ronald A Berk. 2005. Survey of 12 strategies to measure teaching effectiveness. *International journal of teaching and learning in higher education* 17, 1 (2005), 48–62.

[9] Amanda Biggs, Paula Brough, and Suzie Drummond. 2017. Lazarus and Folkman's psychological stress and coping theory. *The handbook of stress and health: A guide to research and practice* (2017), 349–364.

[10] John Biggs. 1998. Assessment and classroom learning: A role for summative assessment? *Assessment in Education: Principles, policy & practice* 5, 1 (1998), 103–110.

[11] John Bitchener, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of second language writing* 14, 3 (2005), 191–205.

[12] Dale L Bolton. 1973. Selection and evaluation of teachers. *(No Title)* (1973).

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[14] Guy A Boysen. 2008. Revenge and student evaluations of teaching. *Teaching of Psychology* 35, 3 (2008), 218–222.

[15] Susan M Brookhart. 2001. Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice* 8, 2 (2001), 153–169.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[17] Jerome S Bruner. 2009. *Actual minds, possible worlds.* Harvard university press.

[18] Judith Prugh Campbell and William C Bozeman. 2007. The value of student ratings: Perceptions of students, teachers, and administrators. *Community College Journal of Research and Practice* 32, 1 (2007), 13–24.

[19] Jeanette Carlsson Hauff, Anders Carlander, Amelie Gamble, Tommy Gärling, and Martin Holmen. 2014. Storytelling as a means to increase consumers' processing of financial information. *International Journal of Bank Marketing* 32, 6 (2014), 494–514.

[20] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472* (2020).

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

[22] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[23] Samuel Cunningham, Melinda Laundon, Abby Cathcart, Md Abul Bashar, and Richi Nayak. 2023. First, do no harm: automated detection of abusive comments in student evaluation of teaching surveys. *Assessment & Evaluation in Higher Education* 48, 3 (2023), 377–389.

[24] Samuel Cunningham-Nelson, Mahsa Baktashmotlagh, and Wageeh Boles. 2019. Visualizing student opinion through text analysis. *IEEE Transactions on Education* 62, 4 (2019), 305–311.

[25] Matthew Davis. 2012. How to Thank a Teacher. *George Lucas Educational Foundation* (2012).

[26] Renske AM de Kleijn. 2023. Supporting student and teacher feedback literacy: an instructional model for student feedback processes. *Assessment & Evaluation in Higher Education* 48, 2 (2023), 186–200.

[27] Renske AM de Kleijn, Larike H Bronkhorst, Paulien C Meijer, Albert Pilot, and Mieke Brekelmans. 2016. Understanding the up, back, and forward-component in master's thesis supervision with adaptivity. *Studies in Higher Education* 41, 8 (2016), 1463–1479.

[28] Jerome G Delaney, Albert Johnson, Trudi Dale Johnson, and Dennis Treslan. 2010. Students' perceptions of effective teaching in higher education. (2010).

[29] Dennis M Docheff. 1990. The feedback sandwich. *Journal of Physical Education, Recreation & Dance* 61, 9 (1990), 17–18.

[30] Anne Dohrenwend. 2002. Serving up the feedback sandwich. *Family practice management* 9, 10 (2002), 43–46.

[31] Katharina Emde, Christoph Klimmt, and Daniela M Schluetz. 2016. Does storytelling help adolescents to process the news? A comparison of narrative news and the inverted pyramid. *Journalism studies* 17, 5 (2016), 608–627.

[32] Eureka Foong, Steven P Dow, Brian P Bailey, and Elizabeth M Gerber. 2017. Online feedback exchange: A framework for understanding the socio-psychological factors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* 4454–4467.

[33] Barbara L Fredrickson. 2001. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American psychologist* 56, 3 (2001), 218.

[34] Erica Frydenberg and Ramon Lewis. 1993. Boys play sport and girls turn to others: Age, gender and ethnicity as determinants of coping. *Journal of adolescence* 16, 3 (1993), 253–266.

[35] Holger Gaertner. 2014. Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation* 42 (2014), 91–99.

[36] Craig S Galbraith, Gregory B Merrill, and Doug M Kline. 2012. Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and bayesian analyses. *Research in Higher Education* 53 (2012), 353–374.

[37] Marcel L Goldschmid. 1978. The evaluation and improvement of teaching in higher education. *Higher education* 7, 2 (1978), 221–245.

[38] Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 235–244.

[39] Andrea Grimes and Richard Harper. 2008. Celebratory technology: new directions for food research in HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 467–476.

[40] Niku Grönberg, Antti Knutas, Timo Hynninen, and Maija Hujala. 2021. Palaute: An online text mining tool for analyzing written student course feedback. *IEEE Access* 9 (2021), 134518–134529.

[41] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M Drucker. 2023. How Do Analysts Understand and Verify AI-Assisted Data Analyses? *arXiv preprint arXiv:2309.10947* (2023).

[42] Norman Hackerman and Marye Anne Fox. 2003. *Evaluating and improving undergraduate teaching in science, technology, engineering, and mathematics*. National Academies Press.

[43] Lalita Haritaipan, Miki Saijo, Celine Mougenot, et al. 2018. Leveraging creativity of design students with a magic-based inspiration tool. In *DS 93: Proceedings of the 20th International Conference on Engineering and Product Design Education (E&PDE 2018), Dyson School of Engineering, Imperial College, London. 6th-7th September 2018*. 265–270.

[44] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing* 40, 1 (2023), 75–87.

[45] Abhinav Hasija and Terry L Esper. 2022. In artificial intelligence (AI) we trust: A qualitative investigation of AI technology acceptance. *Journal of Business Logistics* 43, 3 (2022), 388–412.

[46] Liam Hebert, Gaurav Sahu, Yuxuan Guo, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. 2024. Multi-modal discussion transformer: Integrating text, images and graph transformers to detect hate speech on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22096–22104.

[47] Troy Heffernan. 2022. Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education* 47, 1 (2022), 144–154.

[48] Amy J Henley and Florence D DiGennaro Reed. 2015. Should you order the feedback sandwich? Efficacy of feedback sequence and timing. *Journal of Organizational Behavior Management* 35, 3-4 (2015), 321–335.

[49] Hubert JM Hermans. 1996. Voicing the self: From information processing to dialogical interchange. *Psychological bulletin* 119, 1 (1996), 31.

[50] Henry A Hornstein. 2017. Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education* 4, 1 (2017), 1304016.

[51] Gary Hsieh, Brett A Halperin, Evan Schmitz, Yen Nee Chew, and Yuan-Chi Tseng. 2023. What is in the Cards: Exploring Uses, Patterns, and Trends in Design Cards. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[52] Yinuo Hu, Shiyue Zhang, Viji Sathy, AT Panter, and Mohit Bansal. 2022. Setsum: Summarization and visualization of student evaluations of teaching. *arXiv preprint arXiv:2207.03640* (2022).

[53] Jeff Huang, Oren Etzioni, Luke Zettlemoyer, Kevin Clark, and Christian Lee. 2012. Revminer: An extractive interface for navigating reviews on a smartphone. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 3–12.

[54] Gregory Hum, Brad Wuetherick, and Yeona Jang. 2021. Supporting practical use and understanding of student evaluations of teaching through text analytics design, policies, and practices. In *Analysing Student Feedback in Higher Education*. Routledge, 180–191.

[55] Therese Huston and Carol L Weaver. 2008. Peer coaching: Professional development for experienced faculty. *Innovative higher education* 33 (2008), 5–20.

[56] Mark Huxham, Phyllis Laybourn, Sandra Cairncross, Morag Gray, Norrie Brown, Judy Goldfinch, and Shirley Earl. 2008. Collecting student feedback: a comparison of questionnaire and other methods. *Assessment & Evaluation in Higher Education* 33, 6 (2008), 675–686.

[57] Fiona Hyland and Ken Hyland. 2001. Sugaring the pill: Praise and criticism in written feedback. *Journal of second language writing* 10, 3 (2001), 185–212.

[58] Isabeau Iqbal. 2013. Academics' resistance to summative peer review of teaching: questionable rewards and the importance of student evaluations. *Teaching in Higher Education* 18, 5 (2013), 557–569.

[59] Maria Jackson and Leah Marks. 2016. Improving the effectiveness of feedback by use of assessed reflections and withholding of grades. *Assessment & Evaluation in Higher Education* 41, 4 (2016), 532–547.

[60] Colin James, Caroline Strevens, Rachael Field, and Clare Wilson. 2019. Student wellbeing through teacher wellbeing: A study with law teachers in the UK and Australia. *Student Success* 10, 3 (2019), 76–83.

[61] Rachel Johnson. 2000. The authority of the student evaluation questionnaire. *Teaching in Higher Education* 5, 4 (2000), 419–434.

[62] Anders Jonsson. 2013. Facilitating productive use of feedback in higher education. *Active learning in higher education* 14, 1 (2013), 63–76.

[63] Rachel Jug, Xiaoyin "Sara" Jiang, and Sarah M Bean. 2019. Giving and receiving effective feedback: A review article and how-to guide. *Archives of pathology & laboratory medicine* 143, 2 (2019), 244–250.

[64] Hyeonsu B Kang, Gabriel Amoako, Neil Sengupta, and Steven P Dow. 2018. Paragon: An online gallery for enhancing design feedback with visual examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[65] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M Gerber, Brian P Bailey, and Steven P Dow. 2017. Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 4627–4639.

[66] Richard Lakeman, Rosanne Coutts, Marie Hutchinson, Debbie Massey, Dima Nasrawi, Jann Fielden, and Megan Lee. 2022. Stress, distress, disorder and coping: the impact of anonymous student evaluation of teaching on the health of higher education teachers. *Assessment & Evaluation in Higher Education* 47, 8 (2022), 1489–1500.

[67] Richard Lakeman, Rosanne Coutts, Marie Hutchinson, Debbie Massey, Dima Nasrawi, Jann Fielden, and Megan Lee. 2023. Playing the SET game: how teachers view the impact of student evaluation on the experience of teaching and learning. *Assessment & Evaluation in Higher Education* 48, 6 (2023), 749–759.

[68] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[69] SW LeBaron and Jay Jernick. 2000. Evaluation as a dynamic process. *Family Medicine* 32, 1 (2000), 13–14.

[70] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web* 18, 2 (2024), 1–36.

[71] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281* (2023).

[72] Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*. 342–351.

[73] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding developers' decision-making using the web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.

[74] Weichen Liu, Sijia Xiao, Jacob T Browne, Ming Yang, and Steven P Dow. 2018. ConsensUs: Supporting multi-criteria group decisions by visualizing points of disagreement. , 26 pages.

[75] Andrés Lucero, Peter Dalsgaard, Kim Halskov, and Jacob Buur. 2016. Designing with cards. *Collaboration in creative design: Methods and tools* (2016), 75–95.

[76] Sonja Lutovac, Raimo Kaasila, Jyrki Komulainen, and Merja Maikkola. 2017. University lecturers' emotional responses to and coping with student feedback: a Finnish case study. *European journal of psychology of education* 32 (2017), 235–250.

[77] Lillian MacNell, Adam Driscoll, and Andrea N Hunt. 2015. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education* 40 (2015), 291–303.

[78] Nikola Marangunić and Andrina Granić. 2015. Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society* 14 (2015), 81–95.

[79] Elaine Martin. 1984. Power and authority in the classroom: Sexist stereotypes in teaching evaluations. *Signs: Journal of Women in Culture and Society* 9, 3 (1984), 482–492.

[80] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.

[81] Harry G Murray. 1984. The impact of formative and summative evaluation of teaching in North American universities. *Assessment and evaluation in Higher Education* 9, 2 (1984), 117–132.

[82] Harry G. Murray. 1997. Does evaluation of teaching lead to improvement of teaching? *International Journal for Academic Development* 2, 1 (1997), 8–23. https://doi.org/10.1080/1360144970020102 arXiv:https://doi.org/10.1080/1360144970020102

[83] Melissa M Nelson and Christian D Schunn. 2009. The nature of feedback: How different types of peer feedback affect writing performance. *Instructional science* 37 (2009), 375–401.

[84] Christine M Neuwirth, Ravinder Chandhok, David Charney, Patricia Wojahn, and Loel Kim. 1994. Distributed collaborative writing: A comparison of spoken and written modalities for reviewing and revising documents. In *Proceedings of the sigchi conference on human factors in computing systems*. 51–57.

[85] Thi Thao Duyen T Nguyen, Thomas Garncarz, Felicia Ng, Laura A Dabbish, and Steven P Dow. 2017. Fruitful Feedback: Positive affective language and source anonymity improve critique reception and work outcomes. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1024–1034.

[86] Kingsley Okoye, Arturo Arrona-Palacios, Claudia Camacho-Zuñiga, Nisrine Hammout, Emilia Luttmann Nakamura, Jose Escamilla, and Samira Hosseini. 2020. Impact of students evaluation of teaching: A text analysis of the teachers qualities by gender. *International Journal of Educational Technology in Higher Education* 17, 1 (2020), 1–27.

[87] José Quiroga Pérez, Thanasis Daradoumis, and Joan Manuel Marquès Puig. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education* 28, 6 (2020), 1549–1565.

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

[89] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.

[90] Krzysztof Rybinski and Elzbieta Kopciuszewska. 2021. Will artificial intelligence revolutionise the student evaluation of teaching? A big data study of 1.6 million student reviews. *Assessment & Evaluation in Higher Education* 46, 7 (2021), 1127–1139.

[91] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional science* 18, 2 (1989), 119–144.

[92] Elizabeth Santhanam, Bernardine Lynch, Jeffrey Jones, and Justin Davis. 2021. From anonymous student feedback to impactful strategies for institutional direction. In *Analysing Student Feedback in Higher Education*. Routledge, 167–179.

[93] Joan Sargeant, Karen Mann, Douglas Sinclair, Cees Van der Vleuten, and Job Metsemakers. 2008. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Advances in Health Sciences Education* 13 (2008), 275–288.

[94] Joan M Sargeant, Karen V Mann, Cees P Van der Vleuten, and Job F Metsemakers. 2009. Reflection: a link between receiving and using assessment feedback. *Advances in health sciences education* 14 (2009), 399–410.

[95] Roger Schwarz. 2013. The "sandwich approach" undermines your feedback. *Harvard Business Review* 19 (2013).

[96] Penny M Simpson and Judy A Siguaw. 2000. Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education* 22, 3 (2000), 199–213.

[97] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761* (2024).

[98] Sandra Smele, Andrea Quinlan, and Emerson Lacroix. 2021. Engendering inequities: precariously employed academic women's experiences of student evaluations of teaching. *Gender and education* 33, 8 (2021), 966–982.

[99] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry* 49, 4 (2008), 376–385.

[100] Pieter Spooren, Bert Brockx, and Dimitri Mortelmans. 2013. On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research* 83, 4 (2013), 598–642.

[101] Pieter Spooren, Dimitri Mortelmans, and Joke Denekens. 2007. Student evaluation of teaching quality in higher education: development of an instrument based on 10 Likert-scales. *Assessment & Evaluation in Higher Education* 32, 6 (2007), 667–679.

[102] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[103] Claude M Steele. 1988. The psychology of self-affirmation: Sustaining the integrity of the self. In *Advances in experimental social psychology*. Vol. 21. Elsevier, 261–302.

[104] Sarah J Stein, Allen Goodchild, Adon Moskal, Stuart Terry, and Jenny McDonald. 2021. Student perceptions of student evaluations: enabling student voice and meaningful engagement. *Assessment & Evaluation in Higher Education* 46, 6 (2021), 837–851.

[105] Yoshihito Sugita. 2006. The impact of teachers' comment types on students' revision. *ELT journal* 60, 1 (2006), 34–41.

[106] Leif Sundberg and Jonny Holmström. 2024. Innovating by prompting: How to facilitate innovation in the age of generative AI. *Business Horizons* (2024).

[107] Paul WG Surgenor. 2013. Obstacles and opportunities: addressing the growing pains of summative student evaluation of teaching. *Assessment & Evaluation in Higher Education* 38, 3 (2013), 363–376.

[108] Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1072–1080.

[109] David Tinapple, Loren Olson, and John Sadauskas. 2013. CritViz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15, 1 (2013), 29.

[110] Sheng-Chau Tseng and Chin-Chung Tsai. 2007. On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education* 49, 4 (2007), 1161–1174.

[111] Beatrice Tucker. 2014. Student evaluation surveys: Anonymous comments that offend or are unprofessional. *Higher education* 68 (2014), 347–358.

[112] Martin Ukrop, Valdemar Švábenský, and Jan Nehyba. 2019. Reflective diary for professional development of novice teachers. In *Proceedings of the 50th ACM technical symposium on computer science education*. 1088–1094.

[113] Sara Värlander. 2008. The role of students' emotions in formal feedback situations. *Teaching in higher education* 13, 2 (2008), 145–156.

[114] Keisha A Villanueva, Shane A Brown, Nicole P Pitterson, David S Hurwitz, and Ann Sitomer. 2017. Teaching evaluation practices in engineering programs: current approaches and usefulness. *International Journal of Engineering Education* 33, 4 (2017), 1317–1334.

[115] Sai Wang and Ki Joon Kim. 2023. Content moderation on social media: does it matter who and why moderates hate speech? *Cyberpsychology, Behavior, and Social Networking* 26, 7 (2023), 527–534.

[116] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.

[117] Phillip C Wankat and Frank S Oreovicz. 2015. *Teaching engineering*. Purdue University Press.

[118] Haley Weiss. 2018. Why People Love Spotify's Annual Wrap-Ups. *The Atlantic* (13 December 2018). https://www.theatlantic.com/technology/archive/2018/12/spotifywrapped-and-data-collection/577930/

[119] Maxwell K. Winchester and Tiffany M. Winchester. 2012. If you build it will they come?; Exploring the student perspective of weekly student evaluations of teaching. *Assessment & Evaluation in Higher Education* 37, 6 (2012), 671–682. https://doi.org/10.1080/02602938.2011.563278 arXiv:https://doi.org/10.1080/02602938.2011.563278

[120] Tiffany M Winchester and Maxwell K Winchester. 2014. A longitudinal investigation of the impact of faculty reflective practices on students' evaluations of teaching. *British Journal of Educational Technology* 45, 1 (2014), 112–124.

[121] Naomi E Winstone, Robert A Nash, James Rowntree, and Michael Parker. 2017. 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience. *Studies in Higher Education* 42, 11 (2017), 2026–2041.

[122] Christiane Wölfel and Timothy Merritt. 2013. Method card design dimensions: A survey of card-based design tools. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part I 14*. Springer, 479–486.

[123] Y Wayne Wu and Brian P Bailey. 2017. Bitter sweet or sweet bitter? How valence order and source identity influence feedback acceptance. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 137–147.

[124] Y Wayne Wu and Brian P Bailey. 2018. Soften the pain, increase the gain: Enhancing users' resilience to negative valence feedback. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–20.

[125] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063* (2023).

[126] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1433–1444.

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

[128] Koji Yatani, Michael Novati, Andrew Trusty, and Khai N Truong. 2011. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1541–1550.

[129] Yu-Chun Grace Yen, Steven P Dow, Elizabeth Gerber, and Brian P Bailey. 2017. Listen to others, listen to yourself: Combining feedback review and reflection to improve iterative design. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 158–170.

[130] Yu-Chun Grace Yen, Joy O Kim, and Brian P Bailey. 2020. Decipher: an interactive visualization tool for interpreting unstructured design feedback from multiple providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[131] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1005–1017.

[132] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

A EXAMPLES OF CURRENT SET REPORTS

IASystem
The Course Evaluation Standard

COURSE SUMMARY REPORT
Numeric Responses

Evaluation Delivery: Online
Evaluation Form: A
Responses: 21/22 (95% very high)

Course type: Face-to-Face

Taught by:
Instructor Evaluated:

Overall Summative Rating represents the combined responses of students to the four global summative items and is presented to provide an overall index of the class's quality:

Combined Median	Adjusted Combined Median
3.5	3.5
(0=lowest; 5=highest)	

Challenge and Engagement Index (CEI) combines student responses to several *IASystem* items relating to how academically challenging students found the course to be and how engaged they were:

CEI: 4.9
(1=lowest; 7=highest)

SUMMATIVE ITEMS

	N	Excellent (5)	Very Good (4)	Good (3)	Fair (2)	Poor (1)	Very Poor (0)	Median	Adjusted Median
The course as a whole was:	21	5%	33%	52%	5%	5%		3.3	3.2
The course content was:	21	24%	24%	38%	10%	5%		3.4	3.4
The instructor's contribution to the course was:	21	48%	5%	33%	10%		5%	4.0	4.0
The instructor's effectiveness in teaching the subject matter was:	21	33%	29%	24%	5%	10%		3.9	3.9

STUDENT ENGAGEMENT

Relative to other college courses you have taken:	N	Much Higher (7)	(6)	(5)	Average (4)	(3)	(2)	Much Lower (1)	Median
Do you expect your grade in this course to be:	21		33%	29%	29%	5%		5%	4.9
The intellectual challenge presented was:	20	5%	40%	20%	25%		5%	5%	5.2
The amount of effort you put into this course was:	21	29%	24%	19%	19%		10%		5.6
The amount of effort to succeed in this course was:	21	19%	38%	29%	10%		5%		5.7
Your involvement in course (doing assignments, attending classes, etc.) was:	21	29%	43%	10%	10%	10%			6.0

On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, writing papers and any other course related work? **Class median: 11.2 Hours per credit: 2.8 (N=21)**

Under 2	2-3	4-5	6-7	8-9	10-11	12-13	14-15	16-17	18-19	20-21	22 or more
		24%	10%	5%	14%	24%	14%	5%		5%	

From the total average hours above, how many do you consider were valuable in advancing your education? **Class median: 6.5 Hours per credit: 1.6 (N=21)**

Under 2	2-3	4-5	6-7	8-9	10-11	12-13	14-15	16-17	18-19	20-21	22 or more
14%	10%	10%	33%	14%	14%					5%	

What grade do you expect in this course? **Class median: 3.5 (N=21)**

A (3.9-4.0)	A- (3.5-3.8)	B+ (3.2-3.4)	B (2.9-3.1)	B- (2.5-2.8)	C+ (2.2-2.4)	C (1.9-2.1)	C- (1.5-1.8)	D+ (1.2-1.4)	D (0.9-1.1)	D- (0.7-0.8)	F (0.0)	Pass	Credit	No Credit
5%	52%	24%	19%											

In regard to your academic program, is this course best described as: **(N=21)**

In your major	A core/distribution requirement	An elective	In your minor	A program requirement	Other
57%	10%	29%		5%	

Fig. 6. An example of the front page of the SET report, showing the quantitative metrics of teaching evaluation.

IASystem
The Course Evaluation Standard

COURSE SUMMARY REPORT
Student Comments

Evaluation Delivery: Online
Evaluation Form: A
Responses: 21/22 (95% very high)

Course type: Face-to-Face

Taught by:
Instructor Evaluated:

STANDARD OPEN-ENDED QUESTIONS

Was this class intellectually stimulating? Did it stretch your thinking? Why or why not?

1. YES. Forced me think about environmental ui.
2. kind of
3. I imagine that this course would have been difficult to teach since the audience was both people really new to the work and others who already have industry experience. For me personally, it was a breeze since I already work in interaction design in my own job. A whole project dedicated to creating a log in flow seemed reeeeally elementary.
4. Yes. This is probably the best class I have taken in the HCDE program. It made me think deeper and broader on the topic of interaction design. I feel like I am more knowledgeable and practiced on the subject matter after taking this class
5. Yes, learning new material always stretches your thinking. Led me to think about interactions in a new way, and how to best create them for certain scenarios.
6. Yes, especially with accessibility.
7. Yes. This was a good boost in terms of thinking about interactions in specific.
8. The amount of little and big things required to think about interaction design was far greater than I expected. Part of me thought this class would be fairly easy since I have a graphic design background. However, I quickly found that this level of thinking was more work and effort, but in a good way.
9. This class was intellectually stimulating, in the sense that is a very granular thinker and makes us consider edge cases we might not have thought of before. However, the final project was not... intellectually stimulating. I understand the importance and need of a video sketch, but i do not see how this would be helpful to us in terms of having a portfolio piece (compared to Project 1 or 2). When presenting a video sketch to an interviewer in the industry, i do not see how this would be helpful. Or serve as a powerful piece to showcase interaction design. I wish we spent more time exploring iOT, machine learning and how interaction design can be implemented to those areas. A video sketch could be a helpful supplement, but spending 4 weeks on this was intellectually draining.
10. This class was intellectually stimulating by challenging my current and newly learned abilities. The articles, reading, and projects helped me further gain insights into how to make personas relatable through meaningful connections.
11. Yes, it was intellectually stimulating
12. In some ways it was intellectually stimulating because it forced me to learn new UX design technologies and thinking. On the other hand, it felt very tactical but without any real assistance in tactics. For example, we had a tutorial on Adobe XD but we didn't really go in depth on it and then it played a huge part in how we were graded. We had to do an augmented video but I don't think we had any in-depth guidance on how to make it look and feel augmented.
13. No, I dont think the class content is compatible and suitable with the class curriculum.
15. The class was intellectually stimulating. I made me think of different way to design and focus more on the personas.
16. Yes. The problem statement forced us to think beyond conventional design processes.

What aspects of this class contributed most to your learning?

1. first two projects and critique
2. design critiques
3. I enjoyed all of the time devoted to critique, it helped to get extra 1:1 instructor feedback.
4. Design critique. Specifically the group ones. It's very helpful to see other people's ideas and suggestions on your own design
5. bringing drafts of assignments to class for critique before turning in the final product
7. The chapters from ABout Face 2.0
8. This class made me think about all the little moving pieces that make interaction good, great, and easy. It is very challenging to continually put all these little pieces together and think about interactions that do not yet exist or may be easy to learn and then become the norm in the future. This class has made me think about the future of computing in our daily lives and thinking beyond desktop and mobile design. This class has made me think about all these little moving pieces and how I can go back and redo many of my other web and mobile design projects.
9. knowledge, course content and readings were great!
10. Aspects of this class contributed most to my learning through guest lectures, assigned reading, project 01, project 02, and direct feedback from both Dr. and instructor . I personally believe that the recommended books, reading, speakers and overall assignments helped me further grasp the idea making meaningful connections.
12. I like the assignments and some of the critique.
14. Having time to work with my team and having one on one time with the instructors
15. Creating projects and getting feedback

Fig. 7. An example of the open-ended question page of the SET report, showing the anonymized and randomly-ordered students' qualitative feedback.

B GENERATED MOCKUPS

Original with Mocked Negative Feedback

Question 18. Please give responsible feedback regarding the instructor: b. What suggestions do you have to improve the instructor's teaching?

The course material was not helpful in my learning at all.

I didn't find any value in the course content.

At times, the professor spends a bit too long going through particular topics in class. There were instances where some concepts were quite confusing. One such example was going through SVDs. It would be very helpful if prof curated certain resources (preferably shorter videos) that would help explain it. A youtube channel I found that was quite helpful at the initial stages of the course was Statquest.

more clarity and pacing of class, sometimes will get confusing

nothing

prof can try to explain things a bit slowly especially the math portions of the module as some students might not be well versed with the math side especially if they come from IS and not CS.

I would like to have more examples of how the theories work since such example help me understand better. I would also like to have the slides for the lecture to be released earlier so I can go through them and be more prepared for the lecture.

no

I think the prof thinks too highly of the students. I personally struggled a lot with the topic, not sure about the other students.

I think everything is great with prof's teaching :)

None.

I think it would be good to cover content similar to what would be covered in the quizzes e.g. go through the process of how to calculate certain things, because the slides are very theoretical.

The theory are hard using real numbers and question when teaching along will help increase in understanding

None from me! If I have any friends wanting to take CS420 in future I will definitely recommend Prof Ledent.

Handwriting could be better as it could be hard to follow what was written on the board if one is not paying 100% attention.

Lessen all the talks about the maths and bring up more about the concepts. As some people don't really understand it if the lessons started from math

Nil.

More practices

More confidence!

Can be more clear

Perhaps he is too enthusiastic

teaching could be clearer

Fig. 8. Original:Mock

Original w/o Negative

Question 18. Please give responsible feedback regarding the instructor: b. What suggestions do you have to improve the instructor's teaching?

prof can try to explain things a bit slowly especially the math portions of the module as some students might not be well versed with the math side especially if they come from IS and not CS.

I would like to have more examples of how the theories work since such example help me understand better. I would also like to have the slides for the lecture to be released earlier so I can go through them and be more prepared for the lecture.

no

I think everything is great with prof's teaching :)

The theory are hard using real numbers and question when teaching along will help increase in understanding

None from me! If I have any friends wanting to take CS420 in future I will definitely recommend Prof [redacted].

Lessen all the talks about the maths and bring up more about the concepts. As some people don't really understand it if the lessons started from math

More practices

More confidence!

Can be more clear

Question 19. Please give responsible feedback regarding the course: a. What elements of the course most contributed to your learning?

assignments help me understand the content abit more, quizzes also give a better understanding

I think the general teaching of this course was good and contributed well to my learning. The two quizzes was also useful.

The slides are the main source of learning but the examples and questions really help

Assignments, Wooclap quizzes, in-class exercises

The content itself was very useful. Assignments are quite good as they expose us to popular libraries like Keras and Tensorflow.

Exercises & assignments

The assignment and quiz help me to consolidate what I have learn and apply them, thus allowing me to understand the theory better.

Slides

Assignments.

Projects and in-class exercises were helpful.

chatgpt and google

assignments gave me an insight of how ai works instead of just learning math

quizzes

Fig. 9. Original:Remove

Original with Actionable Feedback

Question 18. Please give responsible feedback regarding the instructor: b. What suggestions do you have to improve the instructor's teaching?

For Question 18b, the actionable feedback is to suggest improvements such as curating additional resources for complex topics, providing more clarity and pacing in class, and offering more examples to enhance understanding

At times, the professor spends a bit too long going through particular topics in class. There were instances where some concepts were quite confusing. One such example was going through SVDs. It would be very helpful if prof curated certain resources (preferably shorter videos) that would help explain it. A youtube channel I found that was quite helpful at the initial stages of the course was Statquest.

more clarity and pacing of class, sometimes will get confusing

nothing

prof can try to explain things a bit slowly especially the math portions of the module as some students might not be well versed with the math side especially if they come from IS and not CS.

I would like to have more examples of how the theories work since such example help me understand better. I would also like to have the slides for the lecture to be released earlier so I can go through them and be more prepared for the lecture.

no

I think the prof thinks too highly of the students. I personally struggled a lot with the topic, not sure about the other students.

I think everything is great with prof's teaching :)

None.

I think it would be good to cover content similar to what would be covered in the quizzes e.g. go through the process of how to calculate certain things, because the slides are very theoretical.

The theory are hard using real numbers and question when teaching along will help increase in understanding

None from me! If I have any friends wanting to take CS420 in future I will definitely recommend Prof [redacted].

Handwriting could be better as it could be hard to follow what was written on the board if one is not paying 100% attention.

Lessen all the talks about the maths and bring up more about the concepts. As some people don't really understand it if the lessons started from math

More practices

More confidence!

Can be more clear

Perhaps he is too enthusiastic

teaching could be clearer

Very finnicky with whiteboard and his laptop. Feels unprepared not with technical aspects, but with administrative things like eLearn.

Fig. 10. Original:Constructive

tables and figures/mockups/Themes-Par aphrase.png

Fig. 11. Themes:Paraphrase

Letter Format Sandwich

Subject: Positive Feedback and Constructive Suggestions for Improvement - XXX Class

Dear Instructor,

I wanted to take a moment to provide some feedback on the XXX class. Overall, I believe you have done an excellent job in teaching the course and your passion for the subject is evident. Your dedication to helping us learn and your openness to feedback have been greatly appreciated.

One aspect that stood out to me was your willingness to go the extra mile for your students. For example, I really appreciated the fact that you stayed up late to have a make-up class with us, despite the time zone difference. Your patience in explaining concepts and your willingness to repeat or draw them out on the whiteboard for better understanding have been invaluable.

Additionally, your enthusiasm and passion for the course have made the class more enjoyable for me. Your efforts to engage the class and your patience in explaining the concepts have not gone unnoticed. It is evident that you care deeply about your students and their learning experience.

I also want to commend you for being approachable and receptive to feedback. Your openness to answering questions and helping students clarify doubts has been greatly appreciated. Your clear explanations and understanding of the theories taught have made it easier for us to grasp the concepts.

While there have been many positive aspects of the class, I would like to offer a few suggestions for improvement. Firstly, some students have mentioned that the pacing of the class and the clarity of certain topics could be improved. It would be helpful if you could provide more examples and go through the process of calculating certain things, especially for topics covered in quizzes. Additionally, releasing the lecture slides earlier would allow us to better prepare for the lectures.

Furthermore, some students have expressed the need for more practice and examples to reinforce the theories taught. Providing additional exercises and resources, such as curated videos, could greatly enhance our understanding of the subject.

Lastly, a few students have mentioned that the handwriting on the whiteboard could be improved, as it can be difficult to follow if one is not paying 100% attention. Taking care to write more clearly would greatly benefit the class.

Overall, I want to reiterate that your teaching has been highly appreciated and has contributed positively to our learning experience. Thank you for your dedication and commitment to our education.

Best regards,

Ken-bot

Fig. 12. Letter:Sandwich