The Canadian Journal for the Scholarship of Teaching and Learning

Volume 14 | Issue 1 Article 10

Spring 05-31-2023

Evaluating and Improving the Formative Use of Student Evaluations of Teaching

Kamel Omer
Western University, komer@uwo.ca
Shoshanah Jacobs
University of Guelph, sjacob04@uoguelph.ca
Karl Cottenie
University of Guelph, cottenie@uoguelph.ca
Bill Bettger
University of Guelph, wbettger@uoguelph.ca
John Dawson
University of Guelph, jdawso01@uoguelph.ca
Steffen Graether
University of Guelph, graether@uoguelph.ca
Coral Murrant
University of Guelph, cmurrant@uoguelph.ca
John Zettel
University of Guelph, jzettel@uoguelph.ca
Genevieve Newton
University of Guelph, newton@uoguelph.ca

https://doi.org/10.5206/cjsotlrcacea.2023.1.10960

Recommended Citation
Omer, K., Jacobs, S., Cottenie, K., Bettger, B., Dawson, J., Graether, S., Murrant, C., Zettel, J., & Newton, G. (2023). Evaluating and improving the formative use of student evaluations of teaching. *The Canadian Journal for the Scholarship of Teaching and Learning, 14*(1). https://doi.org/10.5206/cjsotlrcacea.2023.1.10960

Evaluating and Improving the Formative Use of Student Evaluations of Teaching

Abstract
Student Evaluations of Teaching (SETs) are a ubiquitous tool in higher education. Though they are not effective means of evaluating teaching ability, they are useful in formative teaching development of a teaching career. We characterise the current formative use of and attitudes towards SETs by instructors across all disciplines (STEM and non-STEM). We found that tenured instructors used SETs for formative development more than untenured, and that non-STEM instructors had more negative associations with SETs than STEM instructors. Based upon these data, we make recommendations to redesign the SET instrument and change the way in which the data are used to support formative teaching development.

Les évaluations de l'enseignement par les étudiants et les étudiantes sont un outil omniprésent dans l'enseignement supérieur. Bien qu'elles ne soient pas un moyen efficace d'évaluer l'aptitude à enseigner, elles sont utiles pour le développement de l'enseignement formatif au sein d'une carrière d'enseignant. Nous caractérisons l'emploi formatif actuel des évaluations de l'enseignement par les étudiants et les étudiantes, ainsi que les attitudes envers ces évaluations par les instructeurs et les instructrices, dans toutes les disciplines (STEM et autres que STEM). Nous avons constaté que les instructeurs et les instructrices qui avaient la permanence utilisaient les évaluations de l'enseignement par les étudiants et les étudiantes pour le développement formatif davantage que ne le faisaient les instructeurs et les instructrices n'ayant pas la permanence, et que les instructeurs et les instructrices qui enseignaient des cours autres que STEM avaient davantage d'associations négatives par rapport à ces évaluations que les instructeurs et les instructrices de cours STEM. Sur la base de ces données, nous recommandons de redéfinir les évaluations de l'enseignement par les étudiants et les étudiantes et de changer la manière dont ces données sont employées pour soutenir le développement d'un enseignement formatif.

Keywords
student evaluations of teaching, formative teaching development; évaluations de l'enseignement par les étudiants et les étudiantes, développement de l'enseignement formatif

Cover Page Footnote
The University of Guelph resides in the ancestral and treaty lands of several Indigenous peoples, including the Attawandaron people and the Mississaugas of the Credit, and we recognize and honour our Anishinaabe, Haudenosaunee, and Métis neighbours. We acknowledge that the work presented here occurred on their traditional lands so that we might work to build lasting partnerships that respect, honour, and value the culture, traditions, and wisdom of those who have lived here since time immemorial.

This work was supported by the University of Guelph Learning Enhancement Fund (Grant JE S2403).

This research paper/Rapport de recherche is available in The Canadian Journal for the Scholarship of Teaching and Learning: https://doi.org/10.5206/cjsotlrcacea.2023.1.10960

"The emphasis must be formative in order for faculty members to set clearly defined goals, describe specific procedures for working toward those goals, and gathering useful feedback on their progress toward the accomplishment of those goals." - Shannon et al 1996

Student evaluations of teaching (SETs) are widespread (Linse, 2017; Stein et al., 2013) and likely here to stay. While their execution reflects primarily a summative intention for their use (Linse, 2017; Smith, 2008), SETs have been identified in previous studies as important tools for instructor professional development, (Gupta & Bajaj, 2018; Penny & Coe, 2004; Dresel & Rindermann, 2011; Marsh, 1987; Marsh & Roche, 1993; Murray, 1997; Stein et al., 2013; Cain et al., 2017; Moore & Kuol, 2005; Newton et al., 2019); or formative use. For the purposes of this study, we refer to 'formative' as the long-term career development of instructors, where instructors use a series of student evaluations of their teaching to reflect upon repeated themes within the feedback and evolve their teaching practice throughout their career. It has been proposed that some instructors have negative perceptions of SETs because they are not accurate measures of teaching effectiveness (Golding & Adam, 2016; MacNell et al., 2015; Uttl et al., 2017) and because of this finding, their use for summative purposes will likely be phased out in Canada (see Farr, 2018). This may further affect instructor perception of the value of student evaluations and deter them from using the feedback in a formative manner. A restructuring of the evaluation questions, the way in which they are administered, and the way in which the data are managed, will support a more formative use of the SETs to support student learning.

University instructors' perceptions and attitudes result in a lack of consistent or systematic use of SETs (review Golding & Adam, 2016, and see Edström, 2008; Smith, 2008). Both academic rank and teaching load influence instructor attitudes towards the summative use of SETs, with junior tenure-track and contingent teaching instructors expressing the greatest support for their usefulness for administrative purposes such as tenure and promotion review (Avi-Itzhak & Kramer, 1985). In addition to the evidence that students are not skilled evaluators of teaching effectiveness, factors such as institutional expectations (Nasser & Fresko, 2002), quality of the evaluation instrument (Ballantyne et al, 2000; Penny & Coe, 2007), execution and use by the institution (Edström, 2008; Nasser & Fresko, 2002; Spiller and Ferguson, 2011), and instructor teaching philosophy (Hendry et al., 2007) can affect instructor use of SETs.

However, less focus is placed upon the use of SETs for improving teaching practice within the literature. Shannon et al. (1996) found that over 90% of surveyed instructors reported using SETs for teaching feedback, and that instructors from science and mathematics departments, and those with no formal training in pedagogy, were less likely to use SETs in a formative way. In addition, Kember et al. (2007) found no evidence that student feedback contributed to course changes. Although some strategies have been explored to improve instructor engagement with student feedback, such as in peer-groups and with expert consultations (van Lierop et al., 2018; Spooren et al., 2013), there is a lack of understanding about what factors improve faculty perception of SETs and how this feedback is being used to improve teaching (see Ulker, 2021).

Though SETs could create a channel of communication between instructor and students (Surgenor, 2013), the summative execution of SETs may also be problematic (Sozer et al., 2019) because it does not facilitate their formative use by instructors. Typically, students complete SETs near the end of the semester, before final exams, and instructors are not given access to them until all grades are finalized, leaving no time for instructors to respond and adjust teaching practice within the semester from which feedback was received (Newton et al., 2019; Gravestock & Greenleaf, 2008; Groen & Herry, 2017). In addition, this execution does not encourage the

instructor to reflect or respond to the feedback received. Administering SETs near the end of a semester may explain why most instructors, when asked about the purpose of SETs, agree that there are opportunities for formative development but also state that the data gathered by SETs are of limited utility (Surgenor, 2013).

Given that SETs will continue to be administered in post-secondary institutions and that their summative use is being phased out (see, e.g., Lederman, 2020; Ryerson University vs Ryerson Faculty Association, 2018), we seek to retrofit the SET system such that they can become more valuable in supporting instructors in their formative development. In this study we 1) assess the variability in SET procedures within our university, 2) assess the degree to which SETs are currently being used by faculty for formative purposes, 3) identify the factors that influence their potential formative use, including SET questions and procedures, and 4) suggest revisions of the way in which SETs are administered to support formative use. We conduct our research at the University of Guelph where the SET administrative model described above as 'typical' is followed (Newton et al 2019). Here SETs are made available to students at the end of the semester, usually electronically, where completing the evaluation is optional. The evaluation tool includes open ended and closed questions with Likert Scale response options. Students can choose to remain anonymous or can sign their evaluations. Completed evaluations are made available to instructors and department administration.

Method

Participants

This study was granted ethics approval by the University of Guelph's Research Ethics Board (REB number: 17-10-024). It was conducted according to the University of Guelph's policies regarding research involving human participants. All academic departments at the University of Guelph (n=32) were requested to distribute an email inviting instructors to participate in the study by completing a survey. The email was targeted to all instructors, including sessional lecturers, tenured/tenure-track faculty, and contractually limited faculty.

SET Instruments

Departmental representatives were contacted to request a copy of their SET instrument. Characteristics of each SET instrument, such as number of evaluative questions, timing of SETs, and method of distribution were analyzed to compare SET instruments between departments. Of 32 departments, SET instruments for 28 were received. Appendix A is a comparative table that outlines SET instrument differences between departments.

Survey

The survey (Appendix B) was built upon by a previous study that examined the formative use of SETs instructors in the College of Biological Sciences (CBS) at the University of Guelph (Newton et al., 2019). This initial survey was developed based on responses from focus groups and interviews involving CBS instructors, which collected information on SET perception and the usefulness of SETs to improve teaching practice. Based on the results of the initial survey distributed to CBS instructors, the present survey was modified in anticipation of a wider target

population to include departmental questions. Participants were asked to identify their respective departments.

The survey was created, distributed and exported using the online *Qualtrics* platform. The survey was open for a period of two and a half weeks and was distributed to departments by email on two occasions. As an incentive for participation, participants were informed that upon completing the 15-minute survey, they could voluntarily enter a draw to win an electronic tablet. Surveys with less than 50% completion of the questions were discarded (n=14).

Data Analysis

Building upon the findings of Kwan (2006) that academic discipline has an effect on student ratings of their instructors, we predicted that SET administration and usage by instructors would vary across disciplines. Respondent disciplines were categorized into one of two categories for data analysis: (1) Science, Technology, Engineering, Math (STEM) disciplines (n=111), and (2) Non-STEM (n=107). We included academic rank as a second predictor of SET usage by instructors. For this factor, we categorized Professor, Associate Professor into "Tenured," and Assistant Professor, Sessional Instructor into "Untenured," because more Assistant Professors and Sessional Lectures reported having used SETs in a formative way than either the Professors or Associate Professors (Figure 1); however, we did not have enough participants in all discipline and academic rank combinations for statistical analyses.

Differences based on questions involving ordinal values, such as Likert-scale questions (e.g., "How satisfied are you with SET evaluations?"), were analyzed using Kruskal-Wallis one-way ANOVA, a non-parametric test to identify significant differences between two independent groups. For analysis of differences based on nominal data (e.g., "Which of the following statements do you most agree with?"), chi-squared tests were conducted to identify where SET perceptions significantly diverged between STEM and non-STEM instructors. The outcome to these tests were then summarized into a table to visualize potential differences. Composite categories included STEM Tenured, STEM Untenured, Non-STEM Tenured, and Non-STEM Untenured.

Strategies outlined in Maguire and Delahunt's (2017) guide for research on teaching were used to investigate and qualitatively analyze open-ended responses. Recurring responses identified in comments and responses to open-ended questions were synthesized into themes central to improving formative perception of SETs.

Results

SET Instruments

SET instruments for 28 University of Guelph departments were collected. Our qualitative analysis of this sample of SET instruments revealed four notable differences between STEM and Non-STEM disciplines (Appendix A). Non-STEM SETs had:

1. More student-specific demographic questions.
2. More questions overall.
3. More freeform text responses than Likert-scale questions.
4. Greater focus on the course, rather than the instructor.

Survey responses and demographics

Thirty-one of the 32 departments are represented in the survey responses (the Department of Clinical Studies is not represented). Two hundred and twenty-five responses were collected from instructors at the University of Guelph; 789 full-time instructors are employed at the University of Guelph, in addition to part-time instructors. Each department is represented by, on average, approximately eight participants. More female-identifying faculty and sessional instructors responded to our survey from the Non-STEM than STEM, likely reflecting the higher proportion of these groups within the Non-STEM. Overall, tenured instructors represent the largest group of participants across both the Non-STEM and STEM disciplines (Table 1).

Table 1.

Demographic Characteristics of the Participants in this Study, by Discipline.

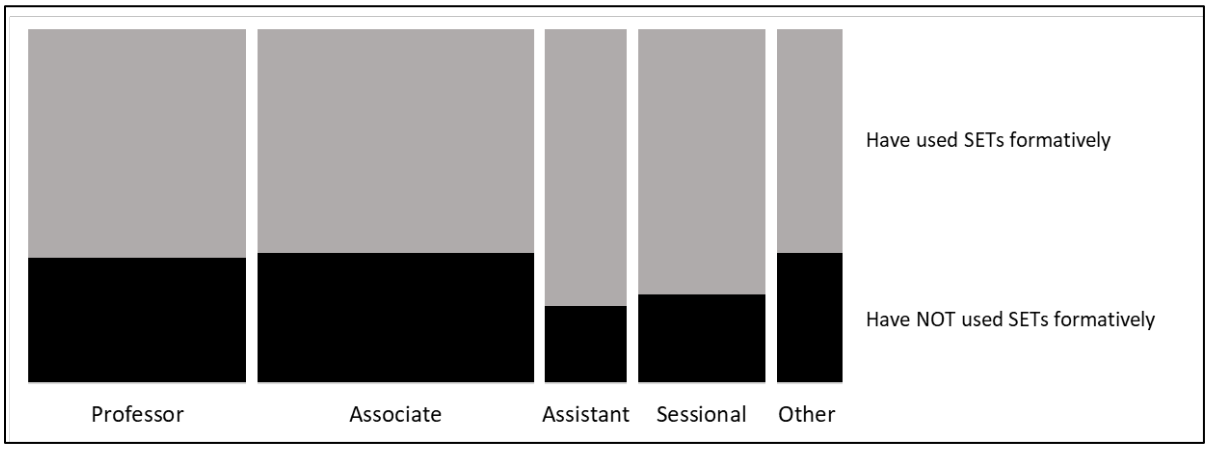
Variable	STEM (*n*=111) % Total	Non-STEM (*n*=107) % Total
Sex (*n*=215)		
Female	45	60
Male	69	41
Other/prefer not to say	6	6
Academic Rank (*n*=216)		
Professor (Tenured)	31	28
Associate Professor (Tenured)	42	36
Assistant Professor (Untenured)	12	11
Sessional Instructor (Untenured)	8	28
Other	16	4
Age (*n*=203)		
24-30	10	8
31-35	7	12
36-40	11	13
41-45	16	14
46-50	19	12
51-55	18	16
56-60	13	14
61-65	6	9
66+	1	4
Received SETs (*n*=216)		
Yes	104	102
No	6	4

Note. Numbers in brackets beside variables represent number of participants from whom the data was provided. Response rates varied between questions, leading to different number of responses included in demographic analyses.

Tenured instructors (i.e., professors and associate professors) displayed significantly different proportions of instructor use of SETs compared to Untenured instructors, with the latter having higher proportion of instructors that have used SETs to inform teaching practice (Figure 1). No significant differences were detected between STEM and non-STEM groups (data not shown).

Figure 1

Mosaic Plot representation of Formative Use of SETs Reception Among Participants, Categorized by Rank.



Note. Width of bars represent relative proportion of total survey participants from that category.

Composite analysis of both academic rank and discipline revealed that discipline is the most important influence in instructor perception of SET effectiveness to inform practice as a means of formative feedback (Tables 2-4). While this analysis displays most of the significant relationships in individual discipline-dependent and rank-dependent analyses, some mutual differences (e.g., SET feedback is not clear; Table 3) are not statistically significant. STEM categories (both Tenured and Untenured) had overall more favorable attitudes towards SETs than Non-STEM. Non-STEM instructors reported higher levels of dissatisfaction with SET instruments (Table 4), and higher levels of agreement with negative perceptions of SETs (e.g., SET is biased; Table 3) compared to STEM instructors. Perceptions regarding whether SET feedback is biased or polarized showed differences between female and male instructors where male instructors were less likely to agree that SET feedback are biased or polarized (Table 3). In all other questions, there were no differences between male or female participants.

Table 2.

Kruskal-Wallis (Measuring Between-Group Differences) of Instructor Agreement with Various Statement Variables, Stratified by Composite Categories of Academic Rank and Discipline.

Statement Variable	Academic Discipline	Academic rank[1]	Mean Likert Score	Mean rank [2]	p-value (two-tailed)
I am able to make course changes following SETs	STEM	Tenured (*n*=78)	2.18	101.28	0.414
		Untenured (*n*=36)	2.42	117.78	
	Non-STEM	Tenured (*n*=62)	2.37	112.62	
		Untenured (*n*=43)	2.37	115.52	
I am able to make changes to my teaching practice following SETs	STEM	Tenured (*n*=74)	1.84	91.1	**0.027***
		Untenured (*n*=36)	2.22	112.8	
	Non-STEM	Tenured (*n*=62)	2.40	122.2	
		Untenured (*n*=43)	2.19	112.7	
I value continuous improvement of teaching	STEM	Tenured (*n*=78)	1.34	115.01	0.423
		Untenured (*n*=36)	1.22	101.94	
	Non-STEM	Tenured (*n*=63)	1.41	114.48	
		Untenured (*n*=43)	1.28	103.64	

Note. * indicates statistical significance.

[1] n represents number of participants from whom the data could be collected and analyzed. Tenured instructors encompass associate professors and professors, while Untenured instructors include assistant professors, sessionals and others.

[2] Ranks were derived (via Kruskal-Wallis H test) from Likert-scale responses ranging from "strongly agree" to "strongly disagree" assigned to ordinal values. Extremes ranged from "strongly agree" (assigned value of 1) to "strongly disagree" (assigned value of 5). Subsequently, higher levels of mean rank correspond to a collectively higher level of disagreement with the statement variable.

Table 3.

Kruskal-Wallis H Test (Measuring Between-Group Differences) of Instructor SET Perception, Stratified by Composite Categories of Academic Rank and Discipline.

Statement Variable	Academic Discipline/ Sex	Academic rank [1]	Mean rank[2]	p-value (two-tailed)
SET feedback can be unclear	STEM	Tenured (*n*=68)	95.3	
		Untenured (*n*=26)	103.7	0.123
	Non-STEM	Tenured (*n*=54)	77.3	
		Untenured (*n*=31)	89.2	
SET feedback can be polarized	STEM	Tenured (*n*=71)	91.26	
		Untenured (*n*=24)	100.90	0.296
	Non-STEM	Tenured (*n*=54)	80.98	
		Untenured (*n*=31)	97.29	
	Female (n = 88)		78.63	**0.043***
	Male (*n*=84)		94.01	
SET is not constructive	STEM	Tenured (*n*=69)	92.22	
		Untenured (*n*=26)	115.06	**0.035***
	Non-STEM	Tenured (*n*=54)	80.81	
		Untenured (*n*=31)	82.94	
SET is biased	STEM	Tenured (*n*=71)	95.40	
		Untenured (*n*=26)	104.29	0.125
	Non-STEM	Tenured (*n*=54)	78.25	

		Untenured (*n*=31)	94.92	
	Female		79.6	**0.043***
	Male		95.02	
SET is not constructed properly	STEM	Tenured (*n*=71)	92.58	**0.013***
		Untenured (*n*=25)	119.68	
	Non-STEM	Tenured (*n*=55)	79.19	
		Untenured (*n*=31)	88.15	

Note. * indicates statistical significance.

[1] n represents number of participants from whom the data could be collected and analyzed. Tenured instructors encompass associate professors and professors, while Untenured instructors include assistant professors, sessionals and others.

[2] Ranks were derived (via Kruskal-Wallis H test) from Likert-scale responses ranging from "strongly agree" to "strongly disagree" assigned to ordinal values. Extremes ranged from "strongly agree" (assigned value of 1) to "strongly disagree" (assigned value of 5). Subsequently, higher levels of mean rank correspond to a collectively higher level of disagreement with statement variable.

Table 4.

Kruskal-Wallis H Test (Measuring Between-Group Differences) of Instructor SET Satisfaction in their Department, Stratified by Composite Categories of Academic Rank and Discipline.

Statement Variable	Academic Discipline	Academic rank[1]	Mean rank[2]	p-value (two-tailed)
I am satisfied with SET questions with respect to their formative value	STEM	Tenured (*n*=70)	89.01	
		Untenured (*n*=26)	68.63	**0.042***
	Non-STEM	Tenured (*n*=55)	100.25	
		Untenured (*n*=31)	100.76	
I am satisfied with procedural administration of SETs with respect to their formative value	STEM	Tenured (*n*=71)	91.69	
		Untenured (*n*=26)	72.42	0.051
	Non-STEM	Tenured (*n*=55)	105.15	
		Untenured (*n*=31)	85.81	
I am satisfied with my department's SETs for formative purposes overall	STEM	Tenured (*n*=71)	90.98	
		Untenured (*n*=26)	62.19	**0.003***
	Non-STEM	Tenured (*n*=55)	107.03	
		Untenured (*n*=32)	95.53	

Note. * indicates statistical significance.

[1] n represents number of participants from whom the data could be collected and analyzed. Tenured instructors encompass associate professors and professors, while Untenured instructors include assistant professors, sessional instructors and others.

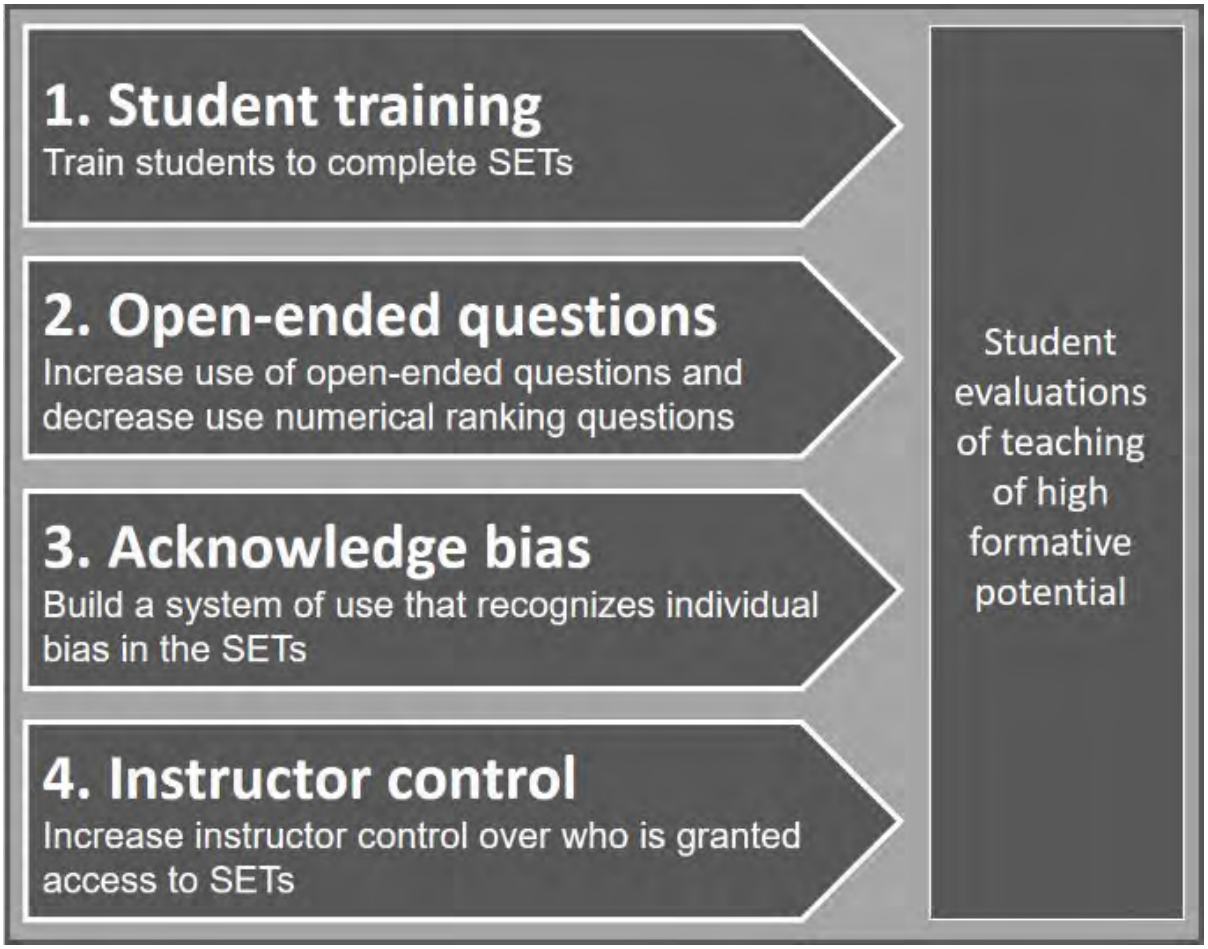
[2] Ranks were derived (via Kruskal-Wallis H test) from Likert-scale responses ranging from "strongly agree" to "strongly disagree" assigned to ordinal values. Extremes ranged from "strongly agree" (assigned value of 1) to "strongly disagree" (assigned value of 5). Subsequently, higher levels of mean rank correspond to a collectively higher level of disagreement with statement variable.

Qualitative analysis

Thematic analysis of comments and responses to open-ended feedback led to the identification of four themes universal to all participant groups (Figure 4). Comments were generally made with respect to improving formative SETs. Therefore, the themes converge at creating departmental SETs that are formatively valuable to instructors.

Figure 4.

Schematic Summarizing Collective Themes in Qualitative Analysis of Both Non-STEM and STEM Instructor Responses.



Themes were identified and are central to identifying SET suggestions that may improve formative value.

Table 5.

Verbatim Responses Supporting Four Themes Identified in Qualitative Analysis that are Central to Improving Formative Value of SETs.

Theme	Supporting statements by participants	
	STEM	**Non-STEM**
Training students for SET completion	"Provide basic implicit bias training to all students." "Ask students to first reflect on their learning before assessing an instructor." "Students (especially younger ones) are not qualified to assess University teaching quality." "There needs to be something noted about bias. Students need to understand the impacts of their comments and while negative comments can be helpful, I think if they understood the importance of these evaluations, they would respond properly. You can still provide constructive feedback without being rude."	"Educate students to take the SET as a constructive way to support teaching." "Have someone (not the instructor) facilitate the feedback gathering- providing explicit direction in how to provide constructive feedback, underlining value of both learning to provide change-enhancing feedback and of their contribution to future instruction in the course/program." "I would encourage greater focus on the classroom culture and environment that was created."
Increased open-ended, decreased numerical questions	"Numerical scores are worse than useless. [They] distract from the meaningful comments, and they provide students an erroneous sense that they can provide some kind of relevant feedback this way."	"A SET question like 'Overall, the instructor was an effective teacher' with a 5-point Likert is meaningless since effective teaching is a multifaceted construct."

	"The numerical scores are not really useful, in part because the questions are quite vague." "Have students write actual comments instead of a numerical ranking." "Keep the freeform text feedback. It is the most useful." "Instead of a general comment box, I would prefer two boxes where students could write things they liked and areas for improvement."	"… the numerical scores should probably be eliminated altogether. They're demonstrably problematic (even though I tend to do well in them)." "I look much more closely at student comments than numerical averages. It's my experience that comments don't usually align with numerical averages well. I.e., students can have glowing positive comments and no negative comments and give a 4/5 on a particular measure with no apparent reason or identifiable area for improvement." "I receive excellent numerical scores, but the comments are usually reserved for how the course or the professor makes the students feel." "I would like to concentrate on the numerical scores, but since we have gone to an online system, the sample of students submitting evaluations are too low to be helpful." "…the comments are really needed to fully understand the meaning behind the numbers."

Acknowledging inherent bias of SETs	“Student opinion is going to be biased by how well they are doing in the course” “The online SETs are very biased. Usually only a fraction of the class fills them out, and I expect that these are mostly the students who were unhappy with the course of the prof.” “[SETs] only provide a biased account of likely a narrow aspect of an instructor's approach to teaching.” “I find students who are required to take a course from a discipline outside their program are biased from the outset.” “It could be linked to accessing grades, but this does not prevent students from providing outrageous or biased comments” “The online SETs are very biased. Usually only a fraction of the class fills them out, and I expect that these are mostly the students who were unhappy with the course of the prof.”	“No matter what, evaluations are going to be biased and polarized.” “Responses will likely still be biased based on the professor's age, gender, race, etc.” “I find that students who don't like the course or think their grade isn't high enough tend to give negative feedback” “There is too much bias available based on students' grades so we need questions that don't give those prominence” “[I] am concerned about the bias in teaching evaluations that negatively affect women and racialized instructors, so I take everything reported in the SETs with a large grain of salt.”

Increased instructor control of SETs	“I'd like the opportunity to develop custom SET questions that are keyed to the material in the course - this would make the numerical responses more useful in revising teaching practice.” “There should be an option for instructor specific questions.” “It would be particularly useful to new instructors if we could design some of our own questions. We receive little hands-on training in teaching and course development in [my department], so I think getting feedback on personalized questions would be helpful.” “There are some strange questions in our SET questionnaire, in my opinion.” “I would remove most of the questions, and refocus the actual questions by linking them directly to the learning outcomes outlined in the course outline. That way it becomes a final instructional activity, where they have to reflect on what the instructor intended the course to be (and not what the student thought the course should be/was).” “The current set of questions are very general and are not specific to individual courses.”	“I would very much appreciate being able to [add] a few course-specific questions, particularly to solicit feed-back when I try something new. I see this as a small set of questions, in addition to the standard set...” “I think it should be routine to allow faculty to add 2-3 questions to the SET so they can ask about specific aspects of their course. I have administered my own evaluation in class to provide me with input but it would be helpful to have it all in one place.” “I would greatly prefer my own method - at the beginning of the class getting students to write where they are/what they know/what they don't know/what they want to know. At the end of the class getting them to redo the exercise and to discuss their responses and insights about learning in the group.”

Discussion

The purpose of this study was to assess the variability in SET procedures within our university and the degree to which SETs are currently being used by faculty for formative purposes, and to identify the factors that influence this potential use, including SET questions and procedures. Through quantitative and qualitative analyses, we found that certain parameters of SET perceptions are significantly influenced by both factors. Recurring concepts in participant responses across all categories led to the identification of four important themes (Figure 4) to enhance formative SET use.

In this study, perception and formative use of SETs varied across disciplines. Specifically, disciplines were categorized into (1) STEM, and (2) Non-STEM. This categorization was based on hypothesized differences in educator perception of SETs among university faculties (Shannon et al.; Kwan, 1999). STEM courses tend to score higher on SET questions than their non-STEM counterparts, although the reason is unclear (Neumann & Neumann, 1985; Kwan, 1999). Out of the eleven parameters of SET use investigated in this study (statement variables in Tables 2-4), six were statistically significant when discipline was assigned as an independent variable. No significant disciplinary differences were observed with respect to value placed on continuous learning.

While it is unclear why different perceptions exist, an analysis of instruments used by STEM and non-STEM departments revealed notable differences. Non-STEM departments tended to have more freeform comments, which was expected to improve formative value of SETs based on written responses of participants in the survey and previous studies. The divergence may be due to non-STEM SET instruments focusing more on the course, rather than the instructor. It is possible that SETs that focus on the course (e.g., “Was the course well-organized?”) rather than the instructor (e.g., “Did the instructor explain ideas clearly?”) may be limited in formative potential, at least as perceived by the individual instructor. The observation that STEM instructors perceive greater value in SETs may be because their SET instruments provide more opportunities for students to give feedback on instructor-specific teaching style (e.g., “How could the instructor’s teaching be improved?”) instead of course-specific material. Therefore, although several significant differences in how different disciplines perceive SETs were detected, differences between SET instruments used across disciplines may be an important consideration.

We also observed that perception and formative use of SETs varied across academic rank. In particular, participants at the rank of Professor reported the highest level of satisfaction with SETs as being useful for informing teaching practice. Although not explored in this study, we hypothesize that this observation may be due to Professors having more experience interpreting SET feedback. A case study by Yao & Grady (2005) suggests that lower-ranked faculty lack experience in navigating feedback they feel is unfair, leading to less formatively useful SET feedback compared to Professors with more extensive teaching experience. Like Professors, significantly more Associate Professors reported having previously used SETs to inform their teaching practices than Untenured instructors. This observation may be related to Associate Professors consistently having the most positive perceptions of SET feedback among ranks. No significant differences were reported between disciplines with respect to value in teaching improvement, nor perception of ability to use SETs to inform teaching practice. Overall, academic rank demonstrated less variation in SET use and perception compared to discipline, suggesting that rank may have a smaller influence.

Some parameters of SET perception were statistically different between discipline and rank analyses. In particular, satisfaction with SETs was higher in STEM instructors, and academic ranks had differential perspectives on SETs. This observation was also observed in the analysis of composite categories of Tenured and Untenured STEM and non-STEM, along with mutual (overlapping) findings with the two previous analyses. Overlapping parameters may have been significant in both analyses because they are particularly sensitive to instructor characteristics (i.e., discipline/academic rank). Analysis of discipline and academic rank as independent variables using a two-level analysis mostly displayed the significant relationships of both factors. However, this analysis did not demonstrate statistically significant differences in one overlapping parameter (opinion on SET clarity), while amplifying the significance of other overlapping parameters (overall SET satisfaction). Therefore, the effects of factors such as discipline and rank may work additively or competitively to influence SET perception.

Two-level analysis of both rank and discipline revealed that Untenured STEM instructors had the highest levels of satisfaction, which appears contrary to the outcome of the rank-divided analysis which revealed no clear rank-dependent patterns in differences of SET perception. Because the Untenured STEM group was small (n=26) compared to the others, it is possible that this effect is influenced by other factors. Nearly all 26 participants are represented by one College, including 15 from the same department. In addition to the statistical limitations, smaller groups may be more susceptible to selection bias, which may have also contributed to this outcome. Future research should repeat this analysis with large, equal representation between disciplines and ranks to clarify how big an effect these instructors have on SET perception.

The present study demonstrates that student feedback has been used by instructors to validate and change both aspects of instructor teaching practice and individual courses, in similar proportions. This practice suggests that SET feedback is used and useful not only to improve teaching content and methods of assessment. Most participants identified that both positive and negative comments informed their teaching practice, indicating that recurring critical comments can be used to identify areas of improvement while positive comments can be used to validate aspects of teaching. Zimmaro et al. (2006) observed that critical comments on SETs tend to be more specific compared to affirmative comments, providing more opportunity for formative use. Regarding the utility of numerical values as compared to written comments, there was a preference for written comments, and most participants reported that they only consider certain numerical scores which they perceive as being relevant. This practice suggests that written comments have more formative value than numerical questions. The usefulness of open-ended feedback for formative use over statistical summaries is well-established (e.g., Smith & Welicker-Pollak, 2008; Hammond et al., 2003). Some participants suggested that the value of numerical-based questions might be improved if instructors were able to add numeric questions that are specific and relevant to them. The validation of formatively important SET components in this survey, such as open-ended comments, are important in designing SETs that are effective for formative purposes. Smith & Welicker-Pollak (2008) similarly identify the importance of designing SETs that assure quality of instruction, rather than narrow elements which are typically used for administrative and summative purposes.

Participants who had not used SETs formatively (18% of all participants) were asked about SET characteristics that discourage formative use. The most frequent responses were that SETs are not useful and that SETs were not representative of the general student population. While the former is a general statement about SET use that may be due to multiple reasons (e.g., lack of instructor training on how to interpret student feedback), the latter may be explained by the

polarity, bias, or low participation rate that are known to affect student feedback (Adams & Umbach 2012; Capa-Aydin 2016; Hoel & Dahl 2019). Knowledge of the factors that lessen SET value to instructors for formative improvement is important for SET development and planning. When asked about how SETs can be improved, changes to current SET questions (i.e., add, remove or modify) were suggested with the greatest frequency along with the closely-related suggestion of allowing instructors to modify SETs, indicating that generic department SET instruments may hinder the formative use of SETs. The importance of individualized SETs is supported by Wolfer & Johnson's (2003) suggestion that when designing SETs to improve instruction, instructor weaknesses should be addressed. Procedurally, instructors most frequently reported that making SET completion mandatory for students would enhance the formative value of SETs because it may decrease student feedback polarity and solve issues of low completion rate (see Kelly, 2012). In our study, changing the time, location (e.g., online or in-class) and frequency of SET administration were also identified as possible improvements to SETs; rather than singular, online, end-of-semester SETs. Our participants suggested administering SETs mid-semester and adding an in-class option to improve representativeness and completion rates of SETs. There is evidence that individual variables, such as delivery format (online or in-class) or length of SETs affects response rates, suggesting that such changes may positively improve the formative potential of SETs (Hardy, 2003; Johnson, 2002; Nowell et al., 2010; Spiller and Ferguson, 2011).

The survey used in the present study also probed for detailed feedback regarding how SETs could be modified to improve their formative utility. Qualitative analysis of open-ended survey responses revealed four interrelated themes central to improving formative potential of SETs. These themes were identified across all strata, including academic rank and discipline, indicating that these SET suggestions are broadly applicable guidelines for improving instructor ability and SET perception. The presence of these themes among participants is not novel to our study (see also Ulker 2021); though with our broad disciplinary and year level representation, and an understanding of the diversity in administration and data management of SETs, we make practical recommendations for the future and formative use of SETs in the context of these themes.

1. **More open-ended questions and freeform text and fewer scale questions**

   Survey participants reported that freeform comments are more useful than numerical scores. Studies focusing on the content of freeform responses in SETs have found that the dimensions of teaching (e.g., engagement with students) is represented in written responses along with other important parameters relevant to teaching, such as course-specific material (Alhija & Fresko, 2009). Moreover, past SET studies have showed that instructors place a strong formative value on written comments compared to numerical scores (Nasser & Fresko, 2002; Smith & Welicker-Pollack, 2008). Unlike numerical questions, open-ended comments allow students to provide instructor-specific feedback (Hammond et al., 2003). This evidence, in tandem with participant reports in our study, indicates that thematic analysis of multiple open-ended comments as outlined in Lewis (2001) may be useful to improve effectiveness of SET as a formative tool.

2. **Training students for SET completion**

   A common critique of the current use of SETs by this study's participant pool is the lack of student education in SET completion, thereby amplifying the polarity of SET feedback.

Response errors including the ceiling/floor effects and the halo effect (i.e., non-teaching qualities of instructor affecting student feedback) have been frequently observed in SETs (Bernardin, 1978; Keeley et al., 2013; Hugh Feeley, 2002). Furthermore, the extent of polarity in responses is decreased in students educated about or given the opportunity to reflect upon common biases (Bernardin et al 1978; Peterson et al., 2019; Hoorens et al., 2020) prior to completing SETs. Thus, student rater training regarding (1) response errors and biases, including poor response rate and non-constructive feedback, and (2) the formative importance of SET feedback for future cohorts may improve instructor perception of formative SETs. Relevant sources and guides, such as Svinicki (2001), may be valuable for instructors to incorporate this training into their teaching.

3. Increase instructor control of SETs

One key observation of the present study is the distinction between how different academic ranks and disciplines approach SETs as a formative tool. Nasser & Fresko (2002) observed a similar lack of consensus between Tenured and Untenured instructors when investigating belief in the validity of SETs. The present study's results include several significant differences in different parameters of SET perception between strata (i.e., rank, discipline, or both), showing that there are multiple factors that shape how an instructor perceives SETs for formative use. It is likely that other factors, such as teaching practices, professional background, and the perception of the existence of learning styles also affect SET perceptions. Moreover, within a department, there may be added or missing components (e.g., online portion, labs) that reduce the formative relevance of generic departmental SETs for some instructors. Thus, to increase SET formative value, departments should provide a template to instructors that can be modified to their own teaching style.

4. Acknowledging inherent bias of SETs

The responses of the present survey indicate that SET bias is a serious concern to instructors that inhibits SET formative potential. Feedback provided by students on SETs have previously been observed to be affected by instructor characteristics unrelated to teaching quality, such as language fluency (Carpenter et al., 2013), difficulty of material (Clayson, Frost & Sheffet, 2006; Kogan, Schoenfeld-Tacher & Hellyer, 2010), physical appearance (Riniolo et al., 2006), academic rank and discipline (Chen & Watkins, 2010). Quantitative analysis of survey responses in this study divided by discipline revealed near-significant results ($p=0.07$) between non-STEM and STEM in their perception of whether SETs are biased, supporting the idea that certain disciplines observe more bias than others. Although educating students of such bias may contribute to its reduction, acknowledging these biases by instructors and institutions may improve the perception of SETs as a formative tool.

This study has both strengths and limitations that should be considered. One strength is the inclusion of a wide range of departments in the analysis (as recommended by Kelly 2012, see also, e.g., the sciences: Chan, Luk, and Zeng 2014; Arts, Science, Business and Education: Sojka, Gupta, and Deeter-Schmelz 2010; undergraduate students: Spooren and Christiaens 2017). In the

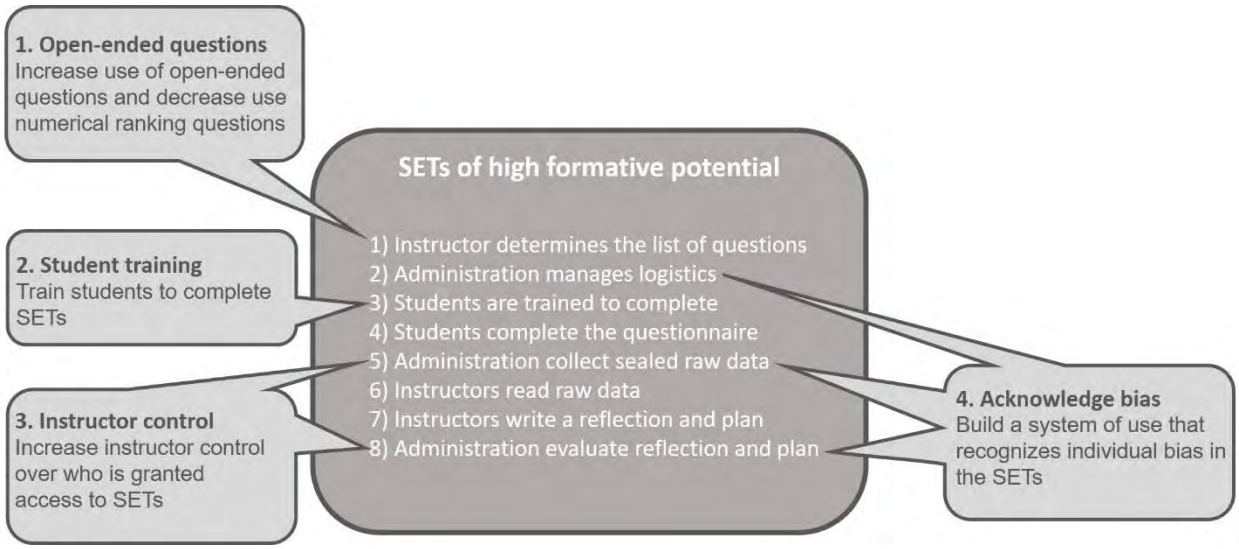
present study, all instructors, regardless of teaching history and experience, were invited to participate. The inclusion of diverse demographics increases the external validity of the results. Another strength is the collection of SET instruments used in each department. These instruments were used to support the interpretation of observed disciplinary differences in formative value placed on SETs. The main limitation of this study is an unequal number of groups across strata, including discipline and rank. Furthermore, division of participants into only two broad disciplines, STEM and Non-STEM, may have masked differences between individual departments, but this was necessary to achieve statistical relevance. Though we had 30% of our total actual instructor population represented in our participant pool, without access to confidential human resource files, it is not possible to know whether the demographic distribution of our participants is representative of the actual population.

Overall, this study has demonstrated that there are significant and notable differences between how faculty of different disciplines and academic rank perceive SETs for formative use. Compared to STEM, non-STEM instructors tended to report more dissatisfaction with SETs as a tool to inform teaching. Unlike the disciplinary analysis, no trend was observed among academic ranks, although statistically significant differences were observed in SET perception. Being a member of STEM faculty appeared to have a stronger influence on SET perception than rank, however, as observed in a composite analysis of rank and discipline where STEM categories had more satisfaction with SETs than non-STEM faculty groups. This study also identified ways to improve the formative value of SETs, such as training students to accurately interpret SETs and providing instructors with the ability to modify SETs to their fit their unique teaching styles and courses. Future studies should (1) validate these suggestions by conducting a similar study in institutions that adhere to them, and (2) further elucidate other factors that affect perception of SETs, such as instructor teaching style. Knowledge of instructor characteristics that influence SET perception is important for the development of SETs conducive to teaching scholarship.

Recommendations to support formative use of SETs:

Figure 5.

Schematic of a Modified Procedure to Support More Formative Use of SETs.



The belief that SETs are only useful for summative purposes may affect their application by instructors for formative purposes (Golding and Adam 2016). Because most instructors determine course delivery, the four themes identified in this study, derived from instructors as participants, should be incorporated into the administration of SETs to promote formative use for course improvement.

We therefore recommend several modifications to the generalized model of SET administration (Figure 5):

Recommendation 1. Instructors should be included in the process of determining the questions asked of students. Here lies the opportunity to include more open-ended questions that will allow instructors to understand more fully the student perspective.

Recommendation 2. Students should have some training in how to complete the SETs in ways that are more helpful to the instructor. This practice will improve the quality of the feedback and also help to reduce bias.

Recommendation 3. Only the instructor is able to see the raw data of the SETs. This restriction will increase instructor control and reduce bias in their interpretation by administrative bodies.

Recommendation 4. In supporting the formative nature of the SETs, instructors should submit a reflection and course design plan that is based upon the SETs in lieu of raw data more typically submitted for evaluation by administration. This SET reflection could be used as part of a multiple source of evidence approach to evaluating teaching effectiveness.

References

Adams, M. J., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*(5), 576-591. https://doi.org/10.1007/s11162-011-9240-5

Alhija, F. N. A., & Fresko, B. (2010). Student evaluation of instruction: what can be learned from students' written comments? *Studies in Educational Evaluation, 35*(1), 37-44. https://doi.org/10.1016/j.stueduc.2009.01.002

Avi-Itzhak, T. E., & Kremer, L. (1986). An investigation into the relationship between university faculty attitudes toward student rating and organizational and background factors. *Educational Research Quarterly, 10*(2), 31-38.

Ballantyne, R., Borthwick, J., & Packer, J. (2000). Beyond student evaluation of teaching: Identifying and addressing academic staff development needs. *Assessment & Evaluation in Higher Education, 25(*3), 221-236. https://doi.org/10.1080/713611430

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education, 3*(4), 245. https://doi.org/10.1037/a0020763

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research. 0(*0), 1-11. https://.14293/S21991006.1.SOR-EDU.AETBZC.v1

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63*(3), 301. https://doi.org/10.1037/0021-9010.63.3.301

Cain, J., Stowe, C. D., Ali, D., & Romanelli, F. (2017). How faculty recognized for teaching excellence interpret and respond to student ratings of teaching. *American Journal of Pharmaceutical Education*, *83*(4), 6680. https://doi.org/10.5688/ajpe6680

Capa-Aydin, Y. 2016. Student Evaluation of Instruction: Comparison between in-Class and Online Methods. *Assessment and Evaluation in Higher Education, 41*(1), 112–126. https://doi.org/10.1080/02602938.2014.987106

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*(6), 1350-1356. https://doi.org/10.3758/s13423-013-0442-z

Chan, C. K. Y., L. Y. Y. Luk, & M. Zeng. (2014). Teachers' perceptions of student evaluations of teaching. *Educational Research and Evaluation*, *20*(4), 275–289. https://doi.org/10.1080/13803611.2014.932698

Chen, G. H., & Watkins, D. (2010). Stability and correlates of student evaluations of teaching at a Chinese university. *Assessment & Evaluation in Higher Education, 35*(6), 675-685. https://doi.org/10.1080/02602930902977715

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education*, *5*(1), 52-65. https://doi.org/10.5465/amle.2006.20388384

Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: A multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education*, *52*(7), 717-737. https://doi.org/10.1007/s11162-011-9214-7

Edström, K. (2008). Doing course evaluation as if learning matters most. *Higher education research & development*, *27*(2), 95-106. https://doi.org/10.1080/07294360701805234

Farr, M. (2018). Arbitration decision on student evaluations of teaching applauded by faculty. *University Affairs*. Aug. 28, 2018

Golding, C., & Adam, L. (2016). Evaluate to improve: Useful approaches to student evaluation. *Assessment & Evaluation in Higher Education*, *41*(1), 1-14. https://doi.org/10.1080/02602938.2014.976810

Gravestock, P., & Gregor-Greenleaf, E. (2008). Student course evaluations: Research, models and trends. *Toronto: Higher Education Quality Council of Ontario*. http://www.heqco.ca/SiteCollectionDocuments/Student%20Course%20Evaluations_Research,%20Models%20and%20Trends.pdf.

Groen, J.F., & Herry, Y. (2017). The online evaluation of courses: Impact on participation rates and evaluation scores. *Canadian Journal of Higher Education, 47*(2), 106-120. https://doi.org/10.47678/cjhe.v47i2.186704

Gupta, P., & Bajaj, N. (2018). Perceptions of the students and faculty of a dental college towards student evaluation of teaching (SET): A cross-sectional study. *Cureus*, *10*(3). e2390. https://doi:10.7759/cureus.2390

Hardy, N. (2003). Online ratings: Fact and fiction. *New directions for teaching and learning*, *96*, 31-38. https://doi.org/10.1002/tl.120

Hammond, I., Taylor, J., & McMenamin, P. (2003). Value of a structured participant evaluation questionnaire in the development of a surgical education program. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, *43*(2), 115-118. https://doi.org/10.1046/j.0004-8666.2003.00037.x

Hendry, G. D., Lyon, P. M., & Henderson-Smart, C. (2007). Teachers' approaches to teaching and responses to student evaluation in a problem-based medical program. *Assessment & Evaluation in Higher Education*, *32*(2), 143-157. https://doi.org/10.1080/02602930600801894

Hoel, A., & Dahl, T. I. (2019). Why bother? Student motivation to participate in student evaluations of teaching. *Assessment & Evaluation in Higher Education*, *44*(3), 361-378. https://doi.org/10.1080/02602938.2018.1511969

Hoorens, V., Dekkers, G., & Deschrijver, E. (2021). Gender bias in student evaluations of teaching: Students' self-affirmation reduces the bias by lowering evaluations of male professors. *Sex Roles*, *84*, 34-48. https://doi.org/10.1007/s11199-020-01148-8

Hugh Feeley, T. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education*, *51*(3), 225-236. https://doi.org/10.1080/03634520216519

Johnson, T. (2002). Online student ratings: Will students respond? *Online student ratings of instruction: New directions for teaching and learning*, *96*, 49-59. https://doi.org/10.1002/tl.122

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement*, *73*(3), 440-457. https://doi.org/10.1177/0013164412475300

Kelly, M. (2012). Student evaluations of teaching effectiveness: Considerations for Ontario universities. *Toronto: Council of Ontario Universities*, (COU #866)

Kember, D., Leung, D. Y., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching?. *Assessment & Evaluation in Higher Education*, *27*(5), 411-425. https://doi.org/10.1080/0260293022000009294

Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2010). Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, *15*(6), 623-636. https://doi.org/10.1080/13562517.2010.491911

Kwan, K. P. (1999). How fair are student ratings in assessing the teaching performance of university teachers? *Assessment & Evaluation in Higher Education*, *24*(2), 181-195. https://doi.org/10.1080/0260293990240207

Lederman, D. (2020, April 8). Evaluating teaching during the pandemic. *Inside Higher Ed.* https://www.insidehighered.com/digital-learning/article/2020/04/08/many-colleges-are-abandoning-or-downgrading-student-evaluations

Lewis, K. G. (2001). Making sense of student written comments. *New Directions for Teaching and Learning*, *87*, 25-32. https://doi.org/10.1002/tl.25

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in educational Evaluation, 54*, 94-106. https://doi.org/10.1016/j.stueduc.2016.12.004

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291-303. https://doi.org/10.1007/s10755 014-9313-4

Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education, 8*(3), 3351-33514.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253-388. https://doi.org/10.1016/0883-0355(87)90001-2

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30*(1), 217-251. https://doi.org/10.1016/0883-0355(87)90001-2

Moore, S., & Kuol, N. (2005). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education, 10*(1), 57-73.

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *The International Journal for Academic Development, 2*(1), 8-23. https://doi.org/10.1080/1360144970020102

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*(2), 187-198. https://doi.org/10.1080/02602930220128751

Neumann, L., & Neumann, Y. (1985). Determinants of students' instructional evaluation: A comparison of four levels of academic areas-a. *The Journal of Educational Research, 78*(3), 152-158. https://doi.org/10.1080/00220671.1985.10885591

Newton, G., Poung, K., Laila, A., Bye, Z., Bettger, W., Cottenie, K., Dawson, J., Graether, S.P., Jacobs, S., Murrant, C., & Zettyl, J. (2019). Perception of biology instructors on using student evaluations to inform their teaching. *International Journal of Higher Education, 8*(1), 133. http://doi.org/10.5430/ijhe.v8n1p133

Nowell, C., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education, 35*(4), 463-475. https://doi.org/10.1080/02602930902862875

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of educational research, 74*(2), 215-253. https://doi.org/10.3102/00346543074002215

Peterson, D. A. M., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLoS One, 14*,1–10. https://doi.org/10.1371/journal.pone.0216241

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology, 133*(1), 19-35. https://doi.org/10.3200/GENP.133.1.19-35

Ryerson University v Ryerson Faculty Association, 2018 CanLII 58446 (ON LA). https://canlii.ca/t/hsqkz

Shannon, D. M., Twale, D. J., & Hancock, G. R. (1996). Use of instructional feedback and modification methods among university faculty. *Assessment and Evaluation in Higher Education*, *21*(1), 41-53. https://doi.org/10.1080/0260293960210104

Smith, C. (2008). Building effectiveness in teaching through targeted evaluation and response: Connecting evaluation to teaching improvement in higher education. *Assessment & Evaluation in Higher Education*, *33*(5), 517-533. https://doi.org/10.1080/02602930701698942

Smith, K., & Welicker-Pollak, M. (2008). What can they say about my teaching? Teacher educators' attitudes to standardised student evaluation of teaching. *European Journal of Teacher Education*, *31*(2), 203-214.

Sojka, J., Gupta, A. K. & Deeter-Schmelz, D. R. (2010). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching*, *50*(2), 44-49. https://doi.org/10.1080/87567550209595873

Spiller, D., & Ferguson, P. B. (2011). Student evaluations: do lecturers value them and use them to engage with student learning needs? *University of Salford, Manchester*. http://usir.salford.ac.uk/id/eprint/16999/

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*(4), 598-642. https://doi.org/10.3102/0034654313496870

Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in Educational Evaluation, 54*, 43-49. https://doi.org/10.1016/j.stueduc.2016.12.003

Stark, P., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*, *0*(0), 1-7. http://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Stein, S. J., Spiller, D., Terry, S., Harris, T., Deaker, L., & Kennedy, J. (2013). Tertiary teachers and student evaluations: never the twain shall meet? *Assessment & Evaluation in Higher Education*, *38*(7), 892-904.

Stein, S. J., Spiller, D., Terry, S., Harris, T., Deaker, L., & Kennedy, J. (2012). Unlocking the impact of tertiary teachers' perceptions of student evaluations of teaching. *Ako Aotearoa National Centre for Tertiary Teaching Excellence.* https://doi.org/10.1080/02602938.2013.767876

Surgenor, P. W. (2013). Obstacles and opportunities: addressing the growing pains of summative student evaluation of teaching. *Assessment & Evaluation in Higher Education, 38*(3), 363-376. https://doi.org/10.1080/02602938.2011.635247

Svinicki, M. D. (2001). Encouraging your students to give feedback. *New Directions for Teaching and Learning*, *87*, 17-24. https://doi.org/10.1002/tl.24

Tucker, B. (2014). Student evaluation surveys: anonymous comments that offend or are unprofessional. *Higher education*, *68*(3), 347-358. https://doi.org/10.1007/s10734-014-9716-2

Ulker, N. (2021). How can student evaluations lead to improvement of teaching quality? A cross-national analysis. *Research in Post-Compulsory Education, 26*(1), 19-37. https://doi.org/10.1080/13596748.2021.1873406

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

van Lierop, M., de Jonge, L., Metsemakers, J., & Dolmans, D. (2018). Peer group reflection on student ratings stimulates clinical teachers to generate plans to improve their teaching. *Medical teacher, 40*(3), 302-309. https://doi.org/10.1080/0142159X.2017.1406903

Wieman, C. (2015). A better way to evaluate undergraduate teaching. *Change: The magazine of higher learning, 47*(1), 6-15. https://doi.org/10.1080/00091383.2015.996077

Wolbring, T., & Treischl, E. (2016). Selection bias in students' evaluation of teaching. *Research in Higher Education, 57*(1), 51-71. https://doi.org/10.1007/s11162-015-9378-7

Wolfer, T. A., & Johnson, M. M. (2003). Re-evaluating student evaluation of teaching: The teaching evaluation form. *Journal of Social Work Education, 39*(1), 111-121. https://doi.org/10.1080/10437797.2003.10779122

Yao, Y., & Grady, M. L. (2005). How do faculty make formative use of student evaluation feedback?: A multiple case study. *Journal of Personnel Evaluation in Education, 18*(2), 107. https://doi.org/10.1007/s11092-006-9000-9

Zimmaro, D. M., Gaede, C. S., Heikes, E. J., Shim, M. P. & Lewis, K.G. (2006). A study of students' written course evaluation comments at a public university. *Measurement and Evaluation Center, University of Texas at Austin*.

Appendix A

Properties of disciplinary SET instruments used across all departments during a 12-week semester.

College	Department	Questions about instructor or course?	Instructor: Course question ratio	In-class or online?	Timing of survey during semester	Incentive for participation	Different SET for different types of courses?	Signed/ unsigned comments?	Question type		
									Scale	Open	Demographic
College of Arts	History	Both	5:8	Instructor choice	Week 10-12	No	No	Yes	13	0	4
	Philosophy	Both	6:11	Online and in-class	Week 10-12	No	No	Yes	17	1	4
	English & Theatre Studies	Both	5:12	Instructor choice	Week 10-12	No	No	Yes	17	1	4
	Fine Art & Music	Instructor	n/a	Online	Week 10-12	No	Yes program types	Yes	14	1	0
	Language & Literature	Both	5:12	In-class and online	Week 10-12	No	Yes languages/ civilization	Yes	17	1	4
College of Biological Science	Cellular & Molecular Biology	Both course + lab	12:5	Instructor choice	Week 10-12	No	No	Yes	17	0	0

College of Biological Sciences	Integrative Biology	Both course + lab	12:5	Instructor choice	Week 10-12 or end of instructor rotation	No	No	Yes	17	0	0
	Health & Nutritional Sciences										
College of Business & Economics	Department Management										
	Economics & Finances and Finance	Both	6:10	Online	Week 10-12	No	No	Yes	13	3	6
	Marketing & Consumer Studies										
	Hospitality, Food & Tourism Management	Both	6:11	Instructor choice	Week 10-12	No	No	Yes	14	3	5
College of Physical & Engineering Science	Chemistry										
	Computer Science	Both	11:12	In-class and online	Week 10-12	No	No	Yes	23	0	2
	Math & Statistics										
	Physics	Both	7:5	Instructor choice	Week 10-12	No	No	Yes	12	1	0
	Engineering	Both	6:3	Online	Week 10-12	No	No	Yes	9	1	0

College of Social & Applied Human Sciences	Family Relations & Applied Human Nutrition	Both	3:4	Instructor choice	Week 10-12	No	Yes undergrad/ grad	Yes	7	2	0
	Geography	Both +TA	7:5	Online	Week 10-12	No	No	Yes	12	3	5
	Psychology	Both +TA	15:12	Instructor choice	Week 10-12	No	Yes DE/in-class /grad	Yes	24	3*	1
	Political Science	Both	7:8	Online	Week 10-12	No	No	Yes	15	0	4
	Sociology & Anthropology	Both	8:1	Instructor choice	Week 10-12	No	No	Yes	9	0	7
	International Development Studies										
Ontario Agriculture College	Food, Agricultural & Resource Economics	Both	5:9	Instructor choice	Week 10-12	Yes exam hints	No	Yes	14	1	0
	Animal & Poultry Science	Both	7:8	Instructor choice	Week 10-12	No	No	Yes	15	1	0
	School of Environmental Science	Both	3:3	Online and in-class	Week 10-12	No	Yes DE/in-class	Yes	6	0	1
	Plant Agriculture	Both +TA	14:6	Instructor choice	Week 10-12	No	No	Yes	20	1	0
	Environmental Design & Rural Development										

Ontario Veterinary College	Biomedical Sciences	Both	4:5	Online	Week 10-12	No	Yes undergrad/ grad/ veterinary	Yes	6	3	0
	Clinical Studies	Instructor	n/a	Instructor choice	End of instructor rotation	No	No	Yes	6	2	0
	Pathobiology										
	Population Medicine	Instructor	n/a	In-class	Week 10-12 or end of instructor rotation	No	No	Yes	1	2	0

Appendix B

Study survey distributed to departments throughout the University of Guelph.

Demographic Information

1) What is your sex? – Male, Female, Other, I prefer not to say

2) What is your age? - 24-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66+, I prefer not to say

3) What department do you work in? – All departments in institution listed

4) What is your academic rank? - Sessional instructor - if so, are you a current graduate student? – Yes/No, Assistant professor, Associate professor, Professor, Other – please specify

5) How many years have you been teaching? - Fewer than 5, 6-10, 11-20, 21-30, 31+

Your Teaching at the University of Guelph

5) How many courses do you typically teach each year? – 1, 2, 3, 4, 5, 6 or more

6) Have you received SETs? – Yes, No

7) How well do you agree with the statement “I feel like I have the ability to make changes to my teaching practice in response to feedback?”

Strongly agree, Agree, Neutral, Disagree, Strongly disagree

8) How well do you agree with the statement “I have the ability to make changes to the course(s) that I teach in response to feedback (for example, to content or structure)?”

Strongly agree, Agree. Neutral, Disagree, Strongly disagree

9) How well do you agree with the statement “I value continuous improvement of my teaching?”

Strongly agree, Agree. Neutral, Disagree, Strongly disagree

Are SETs for formative purposes, and if so, how?

10) Have you ever used SET’s to inform your teaching practice? - Yes, No

11) If you have never or rarely used your SETs to inform your teaching practice, can you explain why? Please select all of the options that are applicable to you. –

- □ I don't think that the SET feedback is useful (if possible, please explain why),
- □ I don't think that the SET feedback is sufficiently representative of the student voice (if possible, please explain why),
- □ I think the SET instrument used in my department is flawed (if possible, please explain why),
- □ I prefer to use my own teaching evaluation method (if possible, please explain this method),
- □ Other (please explain)

12) How have you used SETs to inform your teaching practice? Please select all of the options that are applicable to you. –

- □ SET feedback was used to change an aspect of how I teach (if possible, please provide an example),
- □ SET feedback was used to change an aspect of my course (if possible, please provide an example),
- □ SET feedback was used to validate an aspect of how I teach (if possible, please provide an example),
- □ SET feedback was used to validate an aspect of my course (if possible, please provide an example),
- □ Other (please explain)

What is the perception of the utility of SET's for formative purposes?

13) How do you perceive the utility of comments on SET's in terms of informing your teaching practice? Please select the option that is most applicable to you. –

a) I use primarily positive comments that provide constructive feedback to inform my teaching practice (if possible, please provide an example),
b) I use primarily negative comments that provide critical feedback to inform my teaching practice (if possible, please provide an example),
c) I use both positive and negative student comments to inform my teaching practice (if possible, please provide an example),
d) I am preoccupied with critical or negative comments and these detract from my perception of the utility of SET feedback (if possible, please provide an example),
e) I don't consider the student comments,
f) Students do not have an option to provide written comments as part of the SET process in my department,
g) Other, please explain

14) How do you perceive the utility of numerical scores on SET's in terms of informing your teaching practice? Please select the option that is most applicable to you. –

a) I focus equally on all of the numerical scores to inform my teaching practice,
b) I focus primarily on specific questions that I feel apply readily to my course/teaching to inform my teaching practice (if possible, please provide an example),
c) I focus primarily on the average numerical score to inform my teaching practice
d) I don't consider the numerical scores, I do not receive numerical scores as part of the SET process in my department,
e) Other, please explain

15) Do you agree with the following statements? (Likert scale agreement response options for each) –

a) SET feedback can be unclear,
b) SET feedback can be polarized,
c) SET feedback is not constructive,
d) The SET instrument is not constructed properly,
e) The SET instrument is biased

16) How satisfied are you with the questions asked in your SET instrument? That is, do they provide feedback that can be used for formative purposes?

Very satisfied, Satisfied, Neither satisfied nor dissatisfied, Dissatisfied, Very dissatisfied

17) How satisfied are you with the way SET's are procedurally administered (such as in class or online, or at the end of the semester) in your department? That is, do the procedures allow for provision of feedback that can be used for formative purposes? –

Very satisfied, Satisfied, Neither satisfied nor dissatisfied, Dissatisfied, Very dissatisfied

18) Overall, how satisfied are you with SET's in your department from the perspective of providing feedback that can be used for formative development?

Very satisfied, Satisfied, Neither satisfied nor dissatisfied, Dissatisfied, Very dissatisfied

How might the SET process be changed to improve the formative utility of SET feedback?

19) What changes, if any, would you make to the current complement of SET questions used in your department in order to improve their formative utility? Please select all of the options that are applicable to you. –

☐ Change the wording of certain questions (if possible, provide an example),
☐ Remove certain questions (if possible, provide an example),
☐ Add questions (if possible, provide an example),
☐ Reorganize questions (if possible, provide an example),
☐ Allow instructors to design their own SET questions (if possible, provide an example),
☐ I'm happy with the questions that are currently used,
☐ Other (please explain)

20) What changes, if any, would you make to the SET procedures as they are currently used in your department in order to improve their formative utility? Please select all of the options that are applicable to you. –

- □ Change the timing of SET administration (if possible, provide an example),
- □ Change the location (in class, online, or both) of SET administration (if possible, provide an example),
- □ Change the frequency of SET administration (if possible, provide an example),
- □ Make completion of SET's by students mandatory (if possible, provide an example),
- □ I'm happy with the procedures as they are currently (if possible, provide an example),
- □ Other (please explain)

21) Is there any other information you would like to provide regarding your perception of how the SET process might be changed to improve its formative utility? – Open responses

For entry into draw for compensation:

If you wish to enter the draw for compensation, please click here to access a survey that will ask you for your name and email address.