Perception of Biology Instructors on Using Student Evaluations to Inform Their Teaching

Genevieve Newton[1,4], Kim Poung[1], Amar Laila[1], Zoe Bye[1], William Bettger[1,4], Karl Cottenie[3,4], John Dawson[2,4], Steffen P. Graether[2,4], Shoshanah Jacobs[3,4], Coral Murrant[1,4] & John Zettel[1,4]

[1] Human Health and Nutritional Sciences, University of Guelph, Guelph, Canada

[2] Department of Molecular and Cellular Biology, University of Guelph, Guelph, Canada

[3] Department of Integrative Biology, University of Guelph, Guelph, Canada

[4] BioEd Research Hub, College of Biological Science Office of Educational Scholarship and Practice, University of Guelph, Guelph, Canada

Correspondence: Genevieve Newton. Address: Human Health and Nutritional Sciences, University of Guelph, 50 Stone Road East, Guelph, N1G 2W1, Canada. E-mail: newton@uoguelph.ca

Received: January 24, 2019 Accepted: February 12, 2019 Online Published: February 14, 2019

doi:10.5430/ijhe.v8n1p133 URL: https://doi.org/10.5430/ijhe.v8n1p133

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

Abstract

Student evaluations of teaching (SETs) provide both summative and formative feedback. Although it is clear that SETs are used by administrators for summative purposes, such as providing data to support personnel decisions, it is uncertain how instructors use them for formative development such as to inform overall teaching practice. The objective of our study was to determine the frequency and nature of SET use for formative purposes, to explore the perception of SET utility to inform teaching practice, and to determine how perception of SET utility might be improved to enhance its use in a formative context. Participants were all biological sciences instructors at a large, research-intensive University. This research was conducted in two phases, using a combination of focus groups, interviews, and a survey to yield both qualitative and quantitative data. We found that while instructors generally perceive that SET feedback has formative utility, and that most instructors have used SET feedback for formative purposes at some point, there are many elements of SET administration that they are dissatisfied with, and they suggest several ways in which SETs could be improved (such as allowing in class time for SET administration or doing multiple administrations per semester) to yield more useable feedback that could inform their teaching. The results of this study can be used to further inform the ongoing debate about the role that SETs should play in higher education, as they demonstrate both the utility and concerns about using SETs for formative purposes.

Keywords: course evaluations, formative, student evaluations of teaching, student ratings, teacher evaluations, thematic analysis

1. Introduction

The student evaluation of teaching (SET) process dates back almost 100 years (Murray, 2005). Although the earliest reports of using SETs to provide feedback about teaching effectiveness to instructors was in 1920, they did not become widely used across most North American post-secondary institutions until the 1960s and 1970s (Murray, 2005). There were three main rationales behind the use of SETs: (1) students wanted a say in how they were taught, (2) administrators wanted a way to provide evidence for institutional accountability with respect to teaching, and (3) new faculty desired an additional metric for salary determination, promotion, and tenure decisions unrelated to research publication (Murray, 2005). Since then, the literature has examined, and continues to examine, many aspects related to SETs including their construction (e.g. Guder & Malliaris, 2013; Veeck et al.,, 2016), reliability (e.g. Ewing, 2012; MacNell et al., 2014; Jackson & Jackson, 2015; Boswell, 2016), and validity (Bacon et al., 2016; Boring, 2017; Boring, Ottoboni & Stark, 2016; Kelly, 2012; McNeil, Driscoll & Hunt, 2015; Miles & House, 2015; Wagner, Rieger & Voorvelt, 2016; Uttl et al., 2016).

Though SETs have evolved over time and differ based on the needs and purposes of the institution, they share many common characteristics. They are typically administered at the end of a semester where, in most institutions, students voluntarily complete them individually in the absence of the course instructor; are almost always anonymous, unless students choose to have their feedback known; and frequently contain both qualitative and quantitative responses. Usually, students respond to quantitative questions using a five- to seven-point Likert scale (Gravestock & Gregor-Greenleaf, 2008). Common questions relate to elements of both the instructor and the course, such as the instructor's communication skills, student-teacher interaction, course design, course content, course workload, assessment practices and more. Some institutions have standardized SET forms for all courses, while others have forms prepared at the departmental level (Gravestock & Gregor-Greenleaf, 2008). Universities can administer paper-based SETs in class at a specific time during the evaluation period, which is usually the last two weeks of the semester, and/or they can administer online-based SETs throughout the evaluation period (Groen & Herry, 2017). Interest in online SETs has increased significantly over the past 20 years, with many post-secondary institutions choosing to entirely or partially adopt online administration as the preferred method (Gravestock & Gregor-Greenleaf, 2008).

Despite the existence of alternative means of evaluating and rewarding good teaching, SETs remain the most common tool to assess teaching effectiveness across most North American post-secondary institutions. They provide data for administrators to support personnel decisions regarding salary, promotion, and tenure (Murray, 2005; Gravestock & Gregor-Greenleaf, 2008; Kelly, 2012). Notably, however, a recent arbitration at Ryerson University in Canada ruled that SETs should not be used in tenure and promotion decisions due to evidence that these data are significantly influenced by the personal biases of respondents (Farr, 2018). This has generated uncertainty about how SETs should be used in summative evaluations. There is also concern about the inadmissibility of SETs as a measure of student learning, with a recent meta-analysis showing very little correlation between SET ratings and student achievement (Uttl, White & Gonzalez, 2017). Although there is a large body of evidence indicating that SET feedback is often used by administrators for summative purposes (Murray, 1984; Beran et al., 2005; Beran et al., 2007; Prugh, Campbell & Bozeman, 2007), there is limited research available about how instructors use SET feedback to inform their teaching, which is a~~n~~ non-summative use for SETs that may have important potential.

The formative use of SETs may allow instructors to improve their teaching practice. A recent study by Darwin (2017) used qualitative methods to investigate how instructors understand SETs and how they influence decisions around teaching practice. Focusing on the instructor's perception of student evaluation of their teaching practice, they noted many concerns and much confusion about how to use SETs. More specific to this issue of how SETs are used formatively in higher education, a study conducted by Brickman et al. (2016) surveyed 399 biology instructors from post-secondary institutions across the USA and asked them 'To what extent do you use student evaluations to improve teaching?' The results of this study suggest that a large proportion of biology instructors in post-secondary institutions do use SET feedback to improve their teaching practice, but this is highly variable. For those that ~~do~~ use SET feedback, it is unclear precisely *how* they use it to inform their teaching, *what* their perceptions are of the utility of SET feedback for formative purposes, and *how* the process of obtaining SET feedback could be improved to make it more valuable.

Despite the extensive literature on the summative use of SETs, there is limited data on how this widely administered tool is used formatively. This is an important issue to address, particularly in light of the recent decision related to Ryerson University in Canada regarding the inadmissibility of SETs for purposes of tenure and promotion. The goal of this study was to build on the limited existing research regarding the formative use of SETs specifically by biology instructors, who have been previously shown to report frequent, although diverse, use of SETs to improve teaching. Specifically, this study has three main objectives: (1) To determine the frequency of SET use for formative purposes, and to characterize how they are used in this context; (2) to explore the perception of SET utility to inform teaching practice, with an emphasis on their value and limitations; and (3) to determine how SET utility might be improved to enhance its use in a formative context.

In Phase One of this study, biology instructors were recruited to participate in focus groups and one-on-one interviews. Focus groups and one-on-one interviews are the most commonly used form of data collection in qualitative assessments (Gill et al., 2008; Lambert & Loiselle, 2008), and when used in combination, they provide a deeper understanding of a social phenomenon (Lambert & Loiselle, 2008). In Phase Two of the study, a survey developed based on the responses from the focus groups and interviews was administered across three biological science departments housed in one college, and a mixed methods approach was used to further probe the perception and use of SETs by instructors.

1.1 Context of the Study

This study was conducted in the College of Biological Sciences (CBS) at the University of Guelph, a large, comprehensive, research intensive university in Canada. The CBS consists of three departments based in the biological sciences – Human Health and Nutritional Sciences (HHNS), Integrative Biology (IB), and Molecular and Cellular Biology (MCB). The college has approximately 132 full-time instructors (that is, individuals that present course content in classrooms and are responsible for the submission of final grades for courses), and a variable number of part-time, or sessional, instructors that teach a range of biology-based courses within these three departments.

SETs are currently used to evaluate the teaching of all instructors in CBS. The same SET form is used across all departments in CBS. Student completion is on a voluntary basis, and the forms are administered either in class or online during the last two weeks of each semester, or two weeks from the last lecture. The SET form contains 12 quantitative questions related to elements of both the instructor and the course (see Supplementary Materials). Students have the option to rate their agreement with each question using a five-point Likert scale ranging from 'strongly agree' to 'strongly disagree', which is then translated into numerical scores. Students also have the option to provide open-ended signed or unsigned comments regarding the instructor or course. Both signed and unsigned comments are provided to instructors, along with the rest of the SET feedback (provided as numerical averages and number of students selecting each Likert score) after final grades are released. The chairs of the Department and the Tenure and Promotion Committee are also provided with SET feedback but can only view signed comments.

1.2 Ethics

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.

2. Methods

2.1 Phase One - Focus Groups & Interviews

2.1.1 Participants

Using departmental email lists, all faculty and sessionals (n=132) in CBS were sent an email that described and invited them to participate in the research study. The email invitations gave participants an option to participate either in an interview or focus group to give feedback on their perception and use of SETs. Nine participants agreed ot join the focus groups, while nine participants agreed to partake in interviews, with different participants partaking in each format.

2.1.2 Focus Groups and Interviews

The same question route was used in the focus groups and interviews. This consisted of nine predetermined questions (see Supplementary Materials) that asked participants to reflect on their experience on the end-of-semester SET feedback, to describe how the feedback has or has not been used to inform their teaching practice, and to discuss how the utility of SET feedback might be improved. These questions were developed by a team of researchers who were familiar with the purpose of the study.

All participants provided written informed consent prior to the focus groups meeting. Each focus group consisted of four or five participants, separated according to rank (one group was tenured professors, the other was untenured sessionals and assistant professors), who taught a variety of biology-based undergraduate and/or graduate courses. Separate focus groups were used as it was thought that combining them could discourage untenured participants from contributing openly to the discussion. A Primary Moderator and an Assistant Moderator facilitated each focus group using the question route described above. The conversations varied to some extent, with different follow-up questions asked depending on the responses of the participants to each predetermined question. Each focus group discussion was audio-recorded by the Assistant Moderator, and lasted for about 90 minutes on average. Participants were provided with a free lunch.

All participants provided written informed consent prior to the interviews. Each one-on-one interview was audio-recorded by the interview facilitator, and lasted approximately 30 minutes on average. Interviews were conducted in participant offices, and followed the question route described above. As with the focus groups, the conversations in each interview varied to some extent, with different follow-up questions asked depending on the responses of the participant to each predetermined question. There was no compensation provided for participating in an interview.

2.1.3 Data Analysis

All focus group and interview audio recordings were transcribed by a professional transcription company. The transcripts were verified to the audio recordings by the research team, which assisted in familiarizing the researchers with the data. The transcription data were then uploaded to NVIVO 11 Pro (QSR International), which was used to organize the responses and identify trends. For each question in the question route, several categories of responses were identified. Two researchers first worked independently to analyse each transcript and identify response categories, which were then compared to each other for further refinement until consensus was reached.

2.2 Phase 2 - Survey

2.2.1 Survey Design

The survey (see Supplementary Materials) was based on the questions used in the Phase 1 focus groups and interviews. Demographic questions related to gender, academic rank, and years teaching were also included. Close-ended responses to questions based on analysis of Phase 1 were offered instead of open-ended format ones, although an open-ended "other" option was also provided for each question. Two questions that asked about perceptions of (1) SET questions and (2) SET procedures in general were kept open-ended, as it was not possible to identify clear categories of responses from the Phase 1 data.

2.2.2 Participants

Using the department email lists, all instructors (n=132) in CBS were sent an email that described the study and invited them to participate in the study. A link was provided to the online survey in the invitation email. Participants were informed that they could complete the online survey even if they had participated in a focus group or interview in Phase 1, so some of the same participants contributed to both phases. 36 participants completed the survey.

2.2.3 Survey Data Collection

The survey was hosted online by *Qualtrics* and was open from October 10th, 2018 to November 2nd, 2018 (24 days), and was estimated to take approximately 15 minutes to complete. The link to the survey was sent out via email to all faculty in the college, on three occasions: (1) initial invitation (October 10th), (2) first reminder (October 17th) and (3) second reminder (October 24th). Participants were offered an incentive to participate in the form of a lottery for one of four CAD$25 hospitality gift cards.

2.2.4 Data Analysis

Descriptive statistics are presented for survey responses. Open-ended responses were analysed for recurring themes as outlined in Braun and Clarke's (2006) guide for interpreting qualitative research data.

3. Results

3.1 Phase 1

The demographic characteristics of focus group participants are summarized and compared to the demographic characteristics of interview participants in Table 1. Participant demographics are comparable to that of the College of Biological Sciences as a whole.

Table 1. Demographic characteristics of the participants in Phase One (total n=18).

Variable	Focus Group Participants (n=9)	Interview Participants (n=9)
Sex		
Female	2	3
Male	7	6
Academic Rank		
Professor	1	1
Associate Professor	4	4
Assistant Professor	2	2
Sessional Instructor	2	2
CBS Department		
HHNS	3	6
IB	6	1
MCB	0	2

The analysis of focus group and interview transcripts resulted in the identification of several response categories for each of the nine open-ended questions asked during the question route. These response categories are shown in the survey developed in Phase 2 and provided in the Supplemental Materials.

3.2 Phase Two

Table 2 shows the demographic characteristics of the participants in Phase Two of the research study.

Table 2. Demographic characteristics of the participants in Phase Two. (total n=36)

	n	% Total
Sex		
Female	15	42%
Male	21	58%
Academic Rank		
Professor	10	28%
Associate Professor	19	53%
Assistant Professor	4	11%
Sessional	2	6%
Other	1	3%
Years Teaching		
Less than 5	2	5%
6-10 years	8	22%
11-20 years	19	53%
21-30 years	5	14%
31 + years	2	66%

Results from the survey were organized according to four categories:

(1) Are SETs used for formative purposes, and if so, how?

(2) What is the perception of the utility of SETs for formative purposes?

(3) What are the perceptions of the SET questions and procedures specifically?

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

3.2.1 Are SETs for Formative Purposes, and if so, How?

The vast majority of instructors in CBS reported using SETs to inform their teaching practice (Table 3). A small minority of the population reported that SETs were not used to inform their teaching practice. When instructors were asked what they do with the information provided by SETs and in what timeframe this occurs, the most prevalent response was that faculty read SETs immediately, including both the numerical scores and the comments. The next most common response was that SETs were read at a later date, with it being reported that SETs were read anywhere from one week to two months later, or simply when time permitted. A small minority of instructors reported that they avoided reading SETs or that they scanned the results for later use.

Table 3. Use of SETs for formative purposes and the timeline of SET use.

	n	% Total
Have you ever used SETs to inform your teaching practice?		
Yes	33	92%
No	3	8%
When you receive your SETs, what do you do with them?		
I read them immediately, both comments and numerical scores	26	72%
I read them immediately, but only the comments	-	-
I read them immediately, but only the numerical scores	-	-
I wait until later to read them, either in part or in full	8	22%
I do not read them at all	-	-
Other	2	6%

When asked specifically *how* SETs were used in a formative context, responses included that SETs influenced both the instructor and the course (Table 4). Within the context of the instruction, a majority of respondents reported using SET feedback to change an aspect of how they teach. Aspects of teaching that were changed focused on personal methods of teaching and method execution. At the level of the instructor, a smaller subset of instructors reported using SET feedback to validate a facet of how they teach, gaining justification for their style of teaching. These facets include qualities such as confidence, enthusiasm, and care for the course material.

Within the context of the course itself, many instructors reported using SET feedback to change a feature of their course, including improving course materials such as lecture slides, seminar content and audio-visual aids. They also found SETs useful to gain feedback on the value of selected topics and the use of critical thinking strategies. Instructors also used SET feedback to validate the level of difficulty of a course or use of peer-review evaluations.

Consistent with the very small number of respondents that previously reported not using SETs to inform their teaching practice, a very small number reported that student feedback is not useful and that instructors must trust their own judgement on informing their teaching practice while practicing self-improvement. When instructors were asked if SETs had any utility other than informing their teaching practice, the majority reported using SETs to enhance their teaching dossier. In addition to dossiers, instructors report using SETs for teaching awards, curricula vitae, and in applications for teaching positions.

Table 4. Formative use of SETs by faculty and other non-summative uses of SETs.

	n	% Total
How have you used SET's to inform your teaching practice?		
SET feedback was used to change an aspect of how I teach		
SET feedback was used to change an aspect of my course	28	78%
SET feedback was used to validate an aspect of how I teach	23	64%
SET feedback was used to validate an aspect of my course	18	50%
	15	42%
Do you use your SETs in any other way, or do they have any other utility to you?		
Teaching dossier	22	61%
Curriculum vitae	11	31%
Teaching awards	10	28%
Applications (such as faculty positions)	5	14%
Other	8	22%

3.2.2 What is the Perception of the Utility of SETs for Formative Purposes?

When asked about their perception of the utility of comments on SETs, the majority of instructors in CBS perceive that the SETs providing constructive feedback were the most useful, followed by the perception that written comments are more useful than numerical SET scores (Table 5). A minority of instructors believe that both signed and unsigned comments are equally useful. Some instructors were preoccupied with negative feedback while a small minority chose not to read student comments on SETs. When asked about how they perceived the numerical scores of SETs, the majority of instructors consider certain numerical scores more useful than others, such as the overall average score or the scores for categories that instructors perceive align most closely with their course design.

Several themes emerged from the analysis of these responses, such as perceptions of: (1) limited value of questions viewed as irrelevant to the course, (2) preoccupation with students' experience in contrast to teaching effectiveness, and (3) inclusions of questions that are not suitable to evaluate teaching effectiveness. A small minority found all numerical scores to be equally useful, while an equal number found the average numerical score to be the most useful. Only a very small minority chose to not read numerical scores on SETs.

Overall, a small majority of instructors in CBS report being dissatisfied with SETs as they are currently used. Instructor agreement with specific concerns (including clarity, polarization, constructive utility, instrument construction, and bias) raised in the focus groups and interviews about the utility of SETs is shown in Figure 1A. Instructor satisfaction with the current administration of SETs is shown in Figure 1B.

Table 5. Instructor perception of the formative utility of written comments and numerical scores.

	n	% Total
How do you perceive the utility of comments on SETs?		
I read all student comments and consider both signed and unsigned equally useful	10	28%
I read all student comments, but consider signed comments more useful		
I read all student comments, but consider unsigned comments more useful	11	31%
I read all student comments, but consider those that provide constructive feedback more useful	-	-
I read all student comments, but am preoccupied with those that have positive feedback	19	53%
I read all student comments, but am preoccupied with those that have negative feedback	1	3%
I do not read the student comments	10	28%
Other	3	8%
	2	6%
How do you perceive the utility of numerical scores on SETs?		
All of the numerical scores are equally useful	5	14%
Certain numerical scores are more useful than others	23	64%
The average numerical score is most useful	5	14%
I don't read the numerical scores	2	6%
Other	3	8%

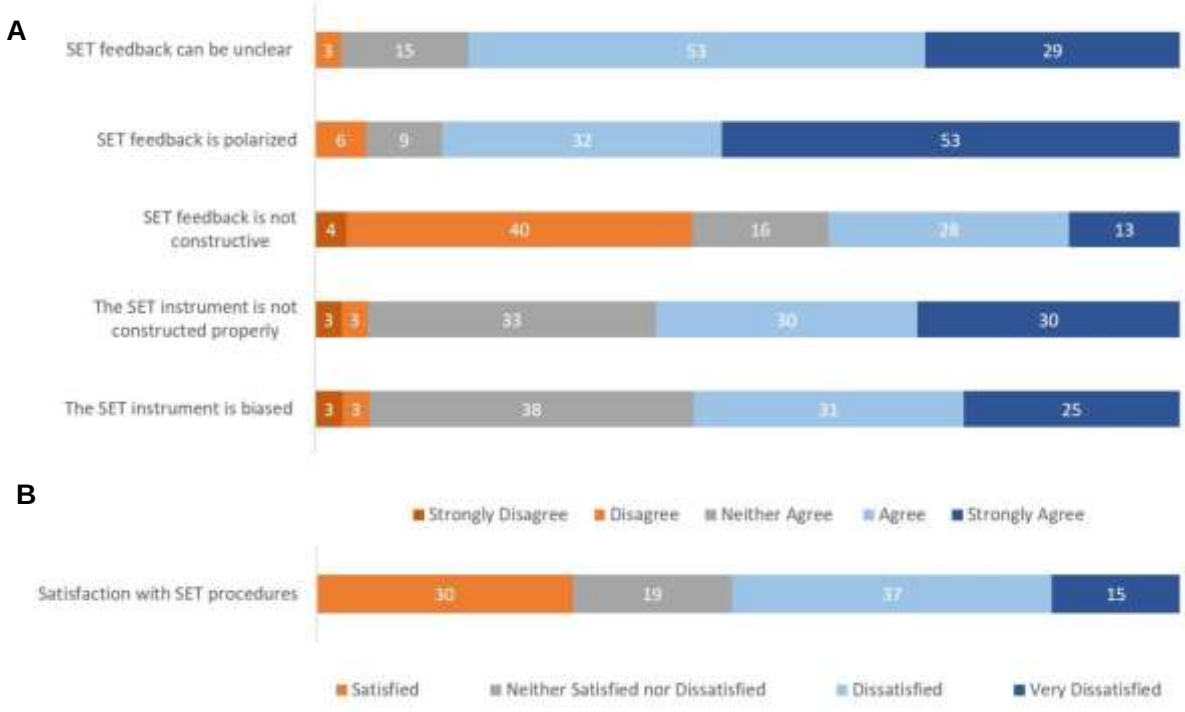

Figure 1. Instructor agreement with concerns raised in the focus groups and interviews about the utility of SETs. **A**. Concerns with the utility of SETs. **B.** Satisfaction with the administration of SETs. Values are expressed as a percentage of the participants.

3.2.3 Open-Ended Response to SET Perceptions

The perceptions of SET questions and procedures are described here as general themes without response frequency since questions were open-ended. Themes that appeared regarding negative attributes of SET questions centred primarily around the ability of students to provide valid feedback, the appropriateness of certain questions, and question interpretation. Instructors commonly reported their belief that students were not capable of evaluating teaching effectiveness, and were not "qualified" to provide this type of feedback. Moreover, instructors perceived that some questions required skills to answer that students may not have developed, such as assessing whether learning outcomes were achieved. Instructors also perceived that SET feedback was influenced by student effort or engagement in a course, which could lead to biased evaluations of teaching effectiveness, and that the validity of SETs to provide feedback that can be used for formative purposes was questionable since they were likely influenced by the students' experience in the course rather than the qualities of the course itself. When asked about their perception of the utility of the current complement of questions, instructors described that there were attributes of questions that make them difficult or not useful to support formative development. Instructors felt that SET questions could be subjective, and that many of them were difficult to interpret.

Themes that appeared regarding the attributes of SET procedures were largely negative and were based around the perception that the procedures encourage bias. When asked about their perception of the utility of the current procedures, instructors mainly reported that they do not perceive that the current procedures support formative development. One perception is that SETs are biased due to their polarized and low response rates. Some instructors perceived that students who completed SETs represent a skewed population, related to the voluntary nature of SET responses. Additionally, SET completion at the end of the semester when students are busiest was a barrier to participation, such that only students who were strongly opinionated about the course complete SETs. The timing of SET administration limit their formative utility because instructors are not provided feedback before the final exam and are not able to make changes to their course for the present cohort. Furthermore, the online administration of SETs allows students who do not regularly attend class to provide feedback even though they may not be able to provide valid opinions when they have been mostly absent. Allowing unsigned comments provides opportunities for students to provide highly negative and even cruel feedback, which can be harmful to instructor morale and mental health.

Despite these negative perceptions regarding SET questions and procedures, an important positive theme emerged regarding SET questions and procedures; SETs provide information regarding the student experience, which was valued by instructors. Instructors reported that SETs allow students to provide constructive feedback on important and relevant aspects of teaching, thus allowing for formative development. Many perceived that SETs reflected the student experience and perception of the course, providing informative student insights. As a result, it was frequently reported that SETs have potential to be used to inform course teaching development to increase learning.

3.2.4 How might the SET Process be Changed to Improve the Formative Utility of SET Feedback?

When asked what changes, if any, faculty would make to the current complement of questions to improve the formative utility of SETs, instructors reported several possible modifications. Many instructors (n=15, 42% total) believed that changing the wording of the questions would improve formative utility. It was assumed by instructors that this would improve student understanding of questions and allow students to provide more appropriate feedback. Many instructors (n=14, 39% total) also saw a benefit in removing specific questions to use SETs for formative development (Table 6). Instructors presumed removing questions that are less geared towards their course will generate more accurate SET scores which will be better suited to formative utility. Some instructors perceived a benefit in the addition of more course-specific questions (n=10, 28% total) or personalised questions written by the instructor (n=9, 25% total). Some instructors (n=9, 13%) thought that reorganizing questions would improve formative utility because reorganization would improve students' perception and understanding of the questions. Only a small number (n=3, 8%) of instructors thought that the current complement of questions should be maintained.

Table 6. Suggested modifications to SET questions to improve formative utility

Suggestions	Example Responses
Change the wording of specific questions	- Focus on the role of the instructor not course specific issues - Assessment of the word "fair" varies between individuals
Remove questions that do not suit numerical extrapolation, and that are not appropriate to the course	- Assessment of the word "effective teaching" - Questions that do not relate well to specific course get rated poorly and bring down the average score
Allow personalization of questions by the instructor	- Allow instructor to add in their own questions based on the course - Add questions about student self-reflection to understand a student's effort in the course - Add questions pertaining to how well the course aligned with learning outcomes
Reorganize questions to promote more useful responses and allow students to understand differences in questions asked	- Reorganize questions into two categories, questions about instructor and questions about the course - Weight more useful questions higher than others to produce a more accurate average of overall teaching effectiveness
Allow instructors to design their own questions	- Ability to seek feedback on particular elements of the course would prove beneficial - Perceived as an interesting and valuable idea

When asked what changes, if any, they would make to the current procedures to improve the formative utility of SETs, instructors reported several possible modifications (Table 7). Many instructors believed that providing in-class time for students to complete online SETs (n=15, 42% total) or to complete paper SETs (n=15, 42% total) would act to improve formative utility. Instructors thought that changing the timing of administering the SETs would help improve formative utility. Some (n=10, 28% total) thought that the timing should be changed to after final exams and only a few (n=3, 8% total) believed that the timing should be changed to after students receive their final grades. Many instructors (n=17, 47% total) suggested that providing SET feedback in the middle of a semester in addition to providing SETs at the end would benefit formative utility. A few instructors (n=4, 11% total) believed there would be no benefit to modifying the timing of administration.

Many instructors (n=13, 36% total) also believed that providing incentives for student completion of SETs would provide a benefit to formative utility of SETs. When instructors were asked what changes, if any, should be made to incentivize participation, many instructors (n=16, 44% total) believed making SET completion a mandatory component of the course would act to improve their formative development. Instructors suggested incentives such as entry into draws for hospitality gift cards, free beverages, or credit to the campus bookstore. A few instructors (n=3, 8%) believed that the current procedure should not be modified. If incentives were used, some instructors mentioned concerns regarding student responses becoming non-thoughtful, and thus not valuable in formative utility.

When instructors were asked what changes, if any, should be made to the pre-determined questions across each course and department to improve formative utility some instructors (n=10, 28% total) believed that standardized questions should not be used across each course and department. However, some instructors (n=10, 28% total) also believed this administration procedure should not be modified. Some instructors (n=5, 14% total) provided responses addressing the use of adding course or department specific questions to aid in the formative utility of SETs.

When instructors were asked what changes, if any, should be made to written comments, many instructors (n=17, 47% total) thought that this procedure should be kept as it is currently administered. Some instructors (n=11, 31% total)

thought that SETs should only allow for non-anonymous comments. A small number of instructors (n=2, 6% total) thought that SETs should only allow for anonymous comments.

Table 7. Suggested modifications to SET procedures to improve formative utility.

Suggestions	Example Responses
Provide in-class time for SET completion or switch administration period to after final exams	- Allow time in-class for student completion of SETs - After the semester is over and final grades have been received allow students to complete SETs
Allow optional SET administration in the middle of the semester and continue SET administration at the end	- Administer a general version of SETs in the middle of the semester - Switch administration of SETs to once, in the middle of the semester
Make SETs a mandatory course component or provide an incentive to complete SETs	- Incentives for SET completion should be tokens of appreciation, not tied to grades - If using incentives must be organized currently to avoid non- thoughtful responses
Continue with standard SET questions but allow addition of course or department specific questions	- Allow addition of department specific questions to better address learning objectives in SETs - Keep standardized questions in order to maintain equal evaluation of instructors
Continue to keep a section for both signed and unsigned comments	- For written comments add emphasis on meeting learning outcomes to ensure valuable responses from students

4. Discussion

The purpose of this study was to determine the frequency and nature of SET use for formative purposes, to explore the perception of SET utility to inform teaching practice, and then to explore how SET utility might be improved to enhance its use in a formative context. We found that while instructors generally perceived that SET feedback has formative utility, and that most instructors have used SET feedback for formative purposes at some point, there are many elements of SET administration that they are dissatisfied with, and they suggested several ways in which SETs could be improved to yield more relevant feedback to inform their teaching.

The primary objective of this study was to determine whether SETs are used by instructors for formative purposes. Similar to Brickman et al. (2016), we observed that the majority of instructors have used SET feedback to inform their teaching practice, with only a few instructors reporting that they had never used their SETs in this way. In fact, SETs were most often read immediately upon receipt, with both numerical scores and comments being considered; although consistent with the widespread use of SETs for formative purposes, many instructors focused primarily on the comments that provides "constructive" or actionable feedback. In response to SET feedback, instructors report making changes both to the way they teach and to their course itself, in addition to using SET feedback to validate both their teaching styles and courses. As well, instructors reported using SET feedback in teaching dossiers, CVs, and to apply for faculty positions and teaching awards. Particularly in light of the recent arbitration in Canada regarding the inadmissibility of SETs in decisions related to tenure and promotion (Farr, 2018), our results identify an alternative utility for SETs; that is, that they may be used by instructors to inform and improve their teaching practice. In this context, however, our study also highlighted many concerns about using SETs that are important to consider.

Despite the widespread use of SETs for formative purposes, faculty expressed many concerns with the SET process in general. One frequently expressed concern related to the ability of students to provide valid feedback. This concern was often couched in the context of summative assessment, even though our questions focused on formative feedback; that is, instructors reported that students were not qualified to evaluate teaching effectiveness. This finding is consistent with those of Ackerman, Gross & Vigneron (2009), who found that seven out of eight instructors surveyed indicated that students were not experts at evaluating teaching effectiveness, but rather that faculty peers were. However, even

when considered in the context of formative assessment, instructors reported concerns about whether students are qualified to provide valid feedback. A common theme in both phases of our research was that SETs primarily reflect student satisfaction, either with the course or instructor, which is not a valid reflection of either. SETs are perceived as being more inwardly focused on the students' own personal experience rather than being focused outwardly on the qualities or characteristics of the teaching. Although somewhat in conflict, instructors acknowledged that there is formative utility in getting feedback about the student experience. It seems that while instructors value this feedback, they interpret it with caution, and generally do not think that it should be used to evaluate teaching effectiveness.

This concern about the validity of SET feedback has long been a primary focus in the SET literature. Unlike for reliability, where there is a general consensus among academics that properly designed SETs are reliable because they provide consistent measurements for specific items associated with teaching effectiveness (Murray, 2005; Gravestock & Gregor-Greenleaf, 2008; Benton, 2009; Kelly, 2012), there is less consensus regarding the overall validity of SETs. While an early review by Geenwald (1997) expressed that SETs are valid, other early reports (Abrami, d'Appollonia & Cohen, 1990; Adams, 1997; Yunker & Yunker, 2003) raised concerns, and emerging research shows that multiple variables, such as instructor gender and appearance, can threaten validity (Boring, 2017; Boring, Ottoboni & Stark, 2016; Braga, Paccagnella, & Pellizzari, 2014; Hoefer, Yurkiewicz, & Byrne, 2012; McNeil, Driscoll & Hunt, 2015; Miles & House, 2015; Spooren, Brockx, & Mortelmans, 2013; Stark & Freishtat, 2014; Wagner, Rieger & Voorvelt, 2016). A commonly cited reference regarding the overall validity of SETs is the report by Gravestock and Gregor-Greenleaf (2008), who argue that even when statistically significant, variables that affect validity usually only change overall SET ratings by less than 0.1%, so validity is maintained if ratings are reported in broad categories to no more than one decimal place. However, it should be noted that since the publication of this assertion in 2008, many research studies regarding bias in SETs have been published, including Boring (2017); Boring, Ottoboni & Stark (2016); Braga, Paccagnella, & Pellizzari (2014); Hoefer, Yurkiewicz, & Byrne (2012); McNeil, Driscoll & Hunt (2015); Miles & House (2015); Spooren, Brockx, & Mortelmans (2013); and Wagner, Rieger & Voorvelt (2016). Since the validity of SETs is contestable, as recently demonstrated through legal arbitration (Farr, 2018), we argue that concern about validity is the main reason why many instructors are cautious about using SET feedback for formative purposes. It is difficult to extricate the concerns about the formative use of SETs from their summative use, because the issue of validity extends readily across both uses.

While many studies have focused on qualitatively measuring the validity of SETs, the present study unpacks why so many instructors perceive SETs to be invalid, which seemed to be largely independent of their familiarity with the SET validity literature. Many of the concerns stem from criticisms of SET questions and SET procedures, which although variable across institutions, often share many similar characteristics, such as a voluntary one-time online administration at the end of the semester using questions about both the course and the instructor. For example, instructors were concerned when the same questions are used across courses, despite courses being highly variable in terms of their pedagogical characteristics (a course with a laboratory component *versus* lecture only), as they may lack relevance. As well, there is concern about student interpretation of questions, which may ask about relatively vague or subjective concepts such "fairness" and "assessment" that may mean different things to different people. Moreover, voluntary participation in SETs can contribute to polarized or biased feedback from students at the top and bottom performance levels, who either seek to praise or criticize, while ignoring the middle majority who may not be sufficiently incentivized to offer their opinions. Negative perceptions regarding SET questions and procedures were highly pervasive in the present study, and undoubtedly contributed to the high level of overall dissatisfaction with the administration of SETs among the surveyed instructors. A small majority of instructors reported either being "very dissatisfied" or "dissatisfied" with the current processes, although a large minority reported some degree of satisfaction. Brickman, Gormally & Martella (2016) similarly found that 41% of biology instructors were dissatisfied with SETs and 46% were satisfied "in some way", while Beran & Rokosh (2009) observed that 70% of instructors surveyed expressed concerns with their institution's SET instrument.

This finding of both satisfaction and dissatisfaction with SETs, as reflected by the seemingly incongruent observations of widespread use of SET feedback to inform teaching practice alongside multiple criticisms of the SET process, speaks both to the potential utility of the SET tool and the need for improvement in the way the tool is administered. In the present study, we asked instructors how they would modify SET questions and procedures to improve their formative utility. Response themes included changing SET questions, changing the mode of administration of SETs, educating students and instructors about SETs, and integrating peer evaluations alongside SETs. With respect to SET questions, it was suggested that these should be better aligned with the learning outcomes of a course to make the questions more relevant. In terms of administration, instructors suggested that there should be in class time for students to complete SETs online on their laptops or mobile devices to increase participation rate, while keeping the

convenience and advantage of electronic SETs. It was also suggested that SETs should continue to be voluntary, but that there could be an incentive for participation. Regarding preparing students to complete SETs, it was recommended that students might take a course or a workshop on this topic. The goal here would be to help students understand their role in the SET process, the impact of SETs on instructors or future students and how learning works. Lastly, although this is not a recommendation that involves direct modification of the SET process, some instructors suggested that in addition to SETs, that they should be evaluated by their peers. Peer evaluation of teaching has long been recommended as a tool for faculty evaluation (Chism, 1999), and is widely used at many institutions, although still it is more commonly used for summative rather than formative purposes. Ackerman, Gross & Vigneron (2009) similarly found that many instructors reported that other instructors, not students, should be providing feedback about teaching.

The primary limitation of the present study is the small sample size, particularly for the focus groups conducted in Phase 1. The number of focus groups recommended by Krueger & Casey (2015), deemed to be three or four for each category of participant, was not reached. However, we supplemented the focus groups with one-on-one interviews, which increased the sample size from which we received qualitative feedback. We allowed participants to self-select to either interviews or focus groups due to the sensitive nature of the subject, as we anticipated that many participants would not feel comfortable talking about their perceptions and use of SETs in an open forum. It was deemed that it was more important to permit participants to disclose their perceptions confidentially than to aim for a greater number of focus groups. As a result, however, we may not have reached saturation (the sample size at which a social phenomenon is fully captured; Krueger & Casey, 2015). To mitigate this issue, we included an "other" option for each question in the survey where close-ended answer options, derived from the Phase 1 feedback, were presented.

In Phase 2 of the research, we attempted to recruit more participants by using an incentive for participation. This did not seem to motivate participation; only about a quarter of eligible instructors in the CBS participated, and of those that did, only about a third disclosed their identity to be entered in the compensation lottery. This lack of participation seems to suggest that research regarding instructor perception of SETs is indeed highly sensitive, which is an important issue of consideration for future research. Interestingly, we might expect that the same issue of polarization related to student participation in SETs might affect research on SETs with instructors. That is, that instructors who feel strongly about the issue would participate, while those with more moderate views would not. This situation does not seem to be what we observed, however, with only a small majority of instructors reporting dissatisfaction with SET administration, although it certainly needs to be considered as a potential bias due to the low rate of participation, and acknowledge that these results may not be generalizable across populations.

The objective of this study was to determine the frequency and nature of SET use for formative purposes, to explore the perception of SET utility to inform teaching practice, and to determine how perception of SET utility might be improved to enhance its use in a formative context. To our knowledge, this study is the first to demonstrate that biology instructors frequently use SET feedback to inform their teaching practice in a variety of ways. However, they also expressed widespread concern and dissatisfaction with several aspects of the SET process, including questions and procedural administration. As such, modifications are recommended to improve the utility of SETs in a formative context. Similar to the argument recently raised by Darwin (2017), we suggest that the cumulative body of SET research suggests that it is time to consider "alternative conceptions" of SETs, whereby the focus shifts to a productive investigation of the student voice, with a reorientation towards students learning. As such, we recommend that the role of SETs shift from being a tool used to evaluate teaching performance for summative purposes to one that is used to provide constructive feedback to instructors so that they can use to become better teachers. Future research should seek to extend this finding to other disciplines, as these findings may not be generalizable to other contexts.

Funding

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

Disclosure Statement

No potential financial interest to the authors were reported.

References

Abrami, P.C., d'Apollonia, S., & Cohen, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*, 219-231. https://doi.org/10.1037/0022-0663.82.2.219

Ackerman, D., Gross, B.L., & Vigneron, F. (2009). Peer Observation Reports and Student Evaluations of Teaching: Who Are the Experts? *Alberta Journal of Educational Research, 55*(1), 18-39. https://ajer.journalhosting.ucalgary.ca/index.php/ajer/article/view/679/660.

Adams, J.V. (1997). Student evaluations: The ratings game. *Inquiry, 1*, 10-16. https://ajer.journalhosting.ucalgary.ca/index.php/ajer/article/view/679/660.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The Utility of Student Ratings of Instruction for Students, Faculty, and Administrators: A "Consequential Validity" study. *The Canadian Journal of Higher Education, 35*(2), 49-70. http://journals.sfu.ca/cjhe/index.php/cjhe/article/view/183500/183449.

Beran, T., Violato, C., & Kline, D. (2007). What's The "Use" of Student Ratings of Instruction for Administrators? One University's Experience. *The Canadian Journal of Higher Education, 37*(1), 27-43. http://journals.sfu.ca/cjhe/index.php/cjhe/article/view/183545/183490.

Beran, T., & Rokosh, J.L. (2009). The Consequential Validity of Student Ratings: What Do Instructors Really Think?. *The Alberta Journal of Educational Research, 55*(4), 497-511. https://ajer.journalhosting.ucalgary.ca/index.php/ajer/article/view/751/722.

Benton, S.L. (2009). Student Ratings of Teaching: A Summary of Research and Literature. Kansas: The IDEA Center. https://www.ntid.rit.edu/sites/default/files/academic_affairs/Sumry%20of%20Res%20%2350%20Benton%202012.pdf.

Braga, M., Paccagnella, MC., & Pellizari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71-88. https://doi.org/10.1016/j.econedurev.2014.04.002

Brickman, P., Gormally, C., & Martella, A.M. (2016). Making the Grade: Using Instructional Feedback and Evaluation to Inspire Evidence-Based Teaching. *Cell Biology Education-Life Sciences Education, 15*(4), 1-14. https://doi.org/10.1187/cbe.15-12-0249

Boring, A. (2017). Gender biases in student evaluations of teaching. *Qualitative Journal of Public Economics, 145*, 27-41. https://doi.org/10.1016/j.pubeco.2016.11.006

Boring, A., Ottoboni, K., & Stark, P.B. (2016). "Student evaluations of teaching (mostly) do not measure teaching effectiveness." *ScienceOpen Research* January 07. 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Chism, N. (1999). Peer review of teaching: A sourcebook. Anker Publishing Co, Bolton MA.

Darwin, S. (2017). What contemporary work are student ratings actually doing in higher education? *Studies in Educational Evaluation, 54*: 13-21. https://doi.org/10.1016/j.stueduc.2016.08.002

Farr, M. (2018). Arbitration evaluation on student evaluations of teaching applauded by faculty. *University Affairs*, August 28th. https://www.universityaffairs.ca/news/news-article/arbitration-decision-on-student-evaluations-of-teaching-applauded-by-faculty/.

Gill, P., Stewart, K.F., Treasure, E.T., & Chadwick, B.L. (2008). Methods of Data Collection in Qualitative Research: Interviews and Focus Groups. *British Dental Journal, 204*(6), 291-295. https://doi.org/10.1038/bdj.2008.192

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.

Greenwald, A.G. (1997). Validity Concerns and Usefulness of Student Ratings of Instruction. *American Psychologist, 5* (11), 1182-1186. https://doi.org/10.1037/0003-066X.52.11.1182

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://eric.ed.gov/?id=EJ1154163

Hoefer, P., Yurkiewicz, J., & Byrne, J.C. (2012). The association between students' evaluation of teaching and grades. *Decision Sciences Journal of Innovative Education, 10*, 447-459. https://doi.org/10.1111/j.1540-4609.2012.00345.x

Hornstein, H.A. (2017). Student Evaluations of Teaching Are an Inadequate Assessment Tool for Evaluating Faculty Performance. *Cogent Education, 4*(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Kelly, M. (2012). Student Evaluations of Teaching Effectiveness: Considerations for Ontario Universities. COU Academic Colleagues Discussions Paper.

http://cou.on.ca/wp-content/uploads/2015/07/Academic-Colleagues-Paper-Student-Evaluations-of-Teaching-Effectiveness.pdf.

Krueger, R.A., & Casey, M.A. (2015). Focus Groups: A Practical Guide for Applied Research. Thousand Oaks: SAGE Publishing.

Lambert, S.D., & Loiselle, C.G. (2008). Combining Individual Interviews and Focus Groups to Enhance Data Richness. *Research Methodology, 62*(2), 228-237. https://doi.org/10.1111/j.1365-2648.2007.04559.x

MacNeil, L., Driscoll, A., & Hunt, A.H. (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

Miles, P., & House, D. (2015). The tail wagging the dog: An overdue examination of teaching evaluations. *International Journal of Higher Education, 4*(2), 116-126. https://doi.org/10.5430/ijhe.v4n2p116

Murray, H.G. (1984). Impact of Formative and Summative Evaluation of Teaching in North American Universities. *Assessment and Evaluation in Higher Education, 9*(2), 117-132. https://doi.org/10.1080/0260293840090204

Murray, H.G. (2005). Student Evaluations of Teaching: Has It Made a Difference? Paper presented at the annual meeting for the Society for Teaching and Learning in Higher Education, Charlottetown, June, 1-15.

Ory, J.C., & Ryan, K. (2001). How Do Student Ratings Measure Up to a New Validity Framework? *New Directions for Institutional Research, 109*, 27-44. https://doi.org/10.1002/ir.2

Peeters, M.J., Beltyukova, S.V., & Martin, B.A. (2013). Educational Testing and Validity of Conclusions in the Scholarship of Teaching and Learning. *American Journal of Pharmaceutical Education, 77*(9), 1-9. https://doi.org/10.5688/ajpe779186

Prugh Campbell, J., & Bozeman, W.C. (2007). The Value of Student Ratings: Perceptions of Students, Teachers, and Administrators. *Community College Journal of Research and Practice, 32*(1), 13-24. https://doi.org/10.1080/10668920600864137

Santhanam, E., & Hicks, O. (2002). Disciplinary, Gender and Course Year Influences on Student Perceptions of Teaching: Explorations and Implications. *Teaching in Higher Education, 7*(1), 17-31. https://doi.org/10.1080/13562510120100364

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). https://doi.org/10.3102/0034654313496870

Stark, P.B., & Freishtat, R. (2014). An evaluation of course evaluations. doi.10.14293/S2199-1006.1.SQR-EDU.AOFRQA.v1Stark. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

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

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review, 54*, 79-94. https://doi.org/10.1016/j.econedurev.2016.06.004

Yunker, P.J., & Yunker, J.A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business, 78*, 313-317. https://doi.org/10.1080/08832320309598619