CSSHE
SCÉES

Canadian Journal of Higher Education
Revue canadienne d'enseignement supérieur
Volume 49, No. 1, 2019, pages 85 - 103

"Somebody has to teach the 'broccoli' course": Administrators Navigating Student Evaluations of Teaching (SET)

Luis Francisco Vargas-Madriz
Norma Nocente
Rebecca Best-Bertwistle
Sarah Forgie
University of Alberta

Abstract

Student Evaluations of Teaching (SET) have been the most consistently administered tool, and they are still extensively used in higher education institutions to assess teaching effectiveness. The purpose of this study was to explore how SET are used by administrators in the teaching evaluation process at a large, research-intensive Canadian university. A basic qualitative research design was used in this project, and semi-structured interviews were used to obtain administrators' experiences. The research question that guided this study was: How are SET (and other tools) used in the evaluation of teaching at this university? Findings showed that although participants mostly utilized a couple of SET statements as indicators of effective teaching, they were certainly aware of the intrinsic issues concerning these tools, and that they are continually seeking to obtain more evidence if SET results are below their benchmarks.

Résumé

L'évaluation de l'enseignement par les étudiants (EEE) a été dans le passé, et est encore aujourd'hui, l'outil le plus largement utilisé dans les établissements d'enseignement supérieur pour évaluer l'efficacité de l'enseignement. Le but de cette étude était d'explorer comment directeurs utilisent l'EEE pour le processus d'évaluation de l'enseignement dans une grande université de recherche Canadienne. Un modèle d'étude qualitative de base, ainsi que des entrevues semi-structurées, ont été utilisés pour recueillir les expériences des directeurs. La question de recherche qui a guidé cette étude était la suivante: comment les EEE sont-ils utilisés dans l›évaluation de l›enseignement dans cette université? Les résultats ont montré que, bien que les participants aient principalement utilisé quelques déclarations dans les EEE comme indicateurs d›un enseignement efficace, ils étaient certainement conscients des problèmes liés à ces outils, et qu›ils cherchaient continuellement à obtenir davantage de preuves si les résultats des EEE étaient inférieurs à leurs critères.

Introduction

Although there are several potential ways of evaluating teaching, Student Evaluations of Teaching (SET) have been the most consistently administered tool that uses ratings for appraising the quality of faculty teaching, and they are still extensively used in higher education institutions to assess teaching effectiveness (Bassett, Cleveland, Acorn, Nix, & Snyder, 2015; M. J. Brown, 2008; Iqbal, Lee, Pearson, & Albon, 2016). Furthermore, because of the relative ease and consistency of the administration of these evaluations, it is understandable why institutions use SET as the main source of input in the evaluation of teaching.

SET shape the quality of instruction being offered to students by providing insight into specific areas of strength or improvement related to areas of teaching such as planning and organization, communication, or assessment (formative evaluation) (Curwood, Tomitsch, Thomson, & Hendry, 2015; Gaillard, Mitchell, & Vahwere, 2006; Iqbal et al., 2016); they also affect the careers of individuals when used to make decisions about faculty retention, promotion, tenure, and pay increments (summative evaluation) (Ahmadi, Helms, & Raiszadeh, 2001; M. J. Brown, 2008). Nevertheless, using SET in summative evaluations is controversial since many claim these tools are biased, invalid, and unreliable, and because they are used in isolation, as opposed to using them with other forms of teaching evaluation (Ahmadi et al., 2001; Gaillard et al., 2006).

Consequently, there is a wealth of research around their potential biases, validity, and reliability. Some investigations have even focused on the small impact of SET for formative purposes, in addition to the concerns regarding the use of SET for summative purposes. However, previous studies have not examined how administrators use the information from SET for formative and summative teaching evaluations.

Therefore, the purpose of this study was to explore how SET are used by administrators in the teaching evaluation process at a large research-intensive Canadian university. A basic qualitative research design was used in this project, and semi-structured interviews were used to obtain administrators’ experiences with evaluating teaching using SET. The main research question that guided this study was: How are SET (and other tools) used in the evaluation of teaching at this university?

This study is of significance because it sheds light on how administrators evaluate teaching using SET in a fair, equitable, and meaningful way despite limitations around implementing multifaceted approaches. Thus, this study is of potential interest for (1) higher education institutions seeking to understand the challenges and opportunities of using SET for teaching evaluation purposes, (2) administrators looking for guidance on how to use SET for the teaching evaluation process, and (3) faculty trying to discern how administrators use SET for formative and summative evaluation purposes.

Conceptual Framework

This section describes strengths and weaknesses of evaluations, potential biases affecting responses, student willingness to provide feedback, and general views about the evaluation process.

Strengths and Weaknesses of Evaluations

Although SET validity and reliability have been frequently disputed, some authors state that these are valid tools to evaluate teaching (Grammatikopoulos, Linardakis, Gregoriadis, & Oikonomidis, 2014; Khong, 2014), and in some cases remain valid tools years after their initial implementation (Nargundkar & Shrikhande, 2012). Though assessing an instrument's validity is a continuous process, some researchers have indicated that SET have good overall reliability and validity with relatively few biases (Socha, 2013; Wright & Jenkins-Guarnieri, 2012). For instance, SET have been successfully implemented in quality assurance and improvement processes in higher education institutions (Ginns, Prosser, & Barrie, 2007), and students in these institutions have been able to distinguish between excellent and poor teaching quality as a result (Dolmans, Janssen-Noordman, & Wolfhagen, 2006). But other investigators have been more cautious by indicating that even when these tools are seemingly reliable, their validity needs to be carefully assessed, particularly when there is a low student response rate (Al-Eidan, Baig, Magzoub, & Omair, 2016), or when administrators follow an inconsistent or ineffective approach to interpretation (Fraile & Bosch-Morell, 2014).

For some authors, student satisfaction does not necessarily imply teaching quality (Bedggood & Donovan, 2012), because it is context-dependent (G. D. Brown, Wood, Ogden, & Maltby, 2015). Furthermore, students may focus more on characteristics that make a course appealing rather than on learning (Chonko, Tanner, & Davis, 2002). For example, their mood (Zumbach & Funke, 2014) and motivation (Chen & Hoshower, 2003) have also been found to affect SET results, and some students have rated instructors inattentively (Uijtdehaage & O'Neal, 2015). These findings lead to questions concerning students' ability to accurately distinguish teaching quality (Grayson, 2015; Lama, Arias, Mendoza, & Manahan, 2015), and they also raise validity concerns (Dodeen, 2013; Martin, Dennehy, & Morgan, 2013; Morley, 2012; Rantanen, 2013; Spooren, Brockx, & Mortelmans, 2013; Uttl, White, & Gonzalez, 2017).

The actual impact of SET on teaching quality and performance is also frequently questioned. While some studies have suggested that SET are beneficial as instructors perceive these results as a valuable piece of information for improving teaching quality (Makondo & Ndebele, 2014) and for refining their teaching skills (Curwood et al., 2015), others have argued that most of this data do not have an actual impact on teaching quality (Asassfeh, Al-Ebous, Khwaileh, & Al-Zoubi, 2013; Campbell & Bozeman, 2007), mostly due to gaps in the way instructors engage with SET results (Stein et al., 2013) or due to the potential negative emotional impact SET results have in instructors' personal lives (Lindahl & Unger, 2010; Zhu, White, Rankin, & Davison, 2018).

However, despite these limitations, some research supports the use of SET for teaching evaluation purposes when administrators follow a careful and consistent approach to interpretation (Fraile & Bosch-Morell, 2014); acknowledge the concerns around their use for faculty retention, promotion, tenure, and pay increments purposes (Jackson & Jackson, 2015; Jones, Gaffney-Rhys, & Jones, 2012; Palmer, 2012); and recognize the potential misinterpretations of SET results (Boysen, 2015; Boysen, Kelly, Raesly, & Casner, 2013; Mitry & Smith, 2014).

Potential Biases Affecting Responses

There are diverse biases that could potentially affect SET responses. Several authors have highlighted that a gender bias influences SET results, yielding lower ratings for female instructors (Boring, Ottoboni, & Stark, 2016; MacNell, Driscoll, & Hunt, 2014) compared to male instructors (Gehrt, Louie, & Osland, 2014; Huebner & Magel, 2015; Mengel, Sauermann, & Zölitz, 2017; Miles & House, 2015; Wilson, Beyer, & Monteiro, 2014). Others have shown evidence of female instructors obtaining higher SET results than males (Centra & Gaubatz, 2000; Smith, Yoo, Farr, Salmon, & Miller, 2007), while others have shown no gender difference (Wright & Jenkins-Guarnieri, 2012).

Instructor characteristic bias may also influence SET responses. Instructor personalities that are outgoing and engaging have positively correlated with SET results (Clayson, 2013; Kim & MacCann, 2016), and instructor physical attractiveness has also positively correlated with evaluations on ratemyprofessor.com (Felton, Mitchell, & Stinson, 2004). Instructor title seems to have affected ratings that resulted in sessional lecturers obtaining higher ratings than full-time faculty (J. Cho, Otani, & Kim, 2014). Instructor age has negatively impacted student perceptions of teachers (Wilson et al., 2014), and has negatively correlated with evaluations on ratemyprofessor.com (Stonebraker & Stone, 2015). Additionally, instructor race and ethnicity appeared to have affected SET results when teachers were visible minorities (Merritt, 2012).

Similarly, non-response bias occurs when SET results are skewed due to a large number of students choosing not to respond. SET ratings have been shown to not only be affected by a low response rate (Al Kuwaiti, AlQuraan, Subbarayalu, & Piro, 2016) but also by the characteristics of the students who complete them (Macfadyen, Dawson, Prest, & Gašević, 2015) and by the characteristics of those who do not (Reisenwitz, 2015).

Lastly, non-instructional bias has occurred when circumstances beyond the control of an instructor have influenced SET results, such as class size (Al Kuwaiti et al., 2016) or subjects that rely on quantitative methods (e.g., mathematical and statistical sciences; Royal & Stockdale, 2015).

Willingness to Provide Feedback

A possible correlation between grades awarded and SET results has also been identified. Students with high grades provided more favourable ratings (Blackhart, Peruche, DeWall, & Joiner, 2006; Miles & House, 2015), as did students with better-than-expected grades (D. Cho, Baek, & Cho, 2015). Conversely, students provided lower ratings for their instructors if they received failing grades (Backer, 2012; Maurer, 2006). Other studies found no correlation between student grades and SET results (Centra, 2003; Gump, 2007), while some suggested that SET results are more sensitive to grade expectation than the effectiveness and quality of the teaching (Boring et al., 2016).

Views about the Evaluation Process

Incorporating multiple sources of information as often as possible for teaching evaluation purposes is optimal (Berk, 2013; Hughes & Pate, 2013; Lyde, Grieshaber, & Byrns, 2016; Ridley & Collins, 2015), but this is difficult in practice and many universities continue to solely rely on SET for teaching evaluation purposes. Alternative teaching evaluation methods have been proposed, such as peer-assessment tools (Cox, Peeters, Stanford, & Seifert, 2013) or student focus groups (Martin et al., 2013), but some authors suggest that these methods should be used in concert with SET of high psychometric quality regarding experiences with the course instruction and learning environment (Fraile & Bosch-Morell, 2014).

Given the complexities around the tools used for the evaluation of teaching, this study sought to examine how administrators evaluate teaching, and specifically how they use and interpret SET and other sources of information despite the aforementioned limitations of these tools.

Method

This study used a basic qualitative research design (Merriam & Tisdell, 2016), with the purpose of exploring how SET are used by administrators in teaching evaluation processes. Ethics approval was sought and obtained from the Human Research Ethics Board.

This study was conducted at a large research-intensive Canadian university with over 37,000 students (30,000 undergraduate) enrolled in the last academic year. Although the university's institutional policy explicitly states the compulsory use of 10 SET items (see Table 1) and of multifaceted teaching evaluation, neither is systematically enforced. SET at this university are available online once the withdrawal deadline for classes has passed and can be completed until the last day of classes. Students receive an email with instructions and an explanation of the purpose of the tool, as well as a link to sign into the rating system once SET become available. During the rating period, both instructors and students receive email reminders to encourage participation. Once the rating period is complete, students, instructors, and administrators can view results online. Overall, the institution has reached a 60% SET completion rate.

This university has a total of 73 department chairs (or their equivalents in non-departmental faculties) across 18 faculties, each appointed for a five-year term. These administrators attend a leadership program at the start of their appointment, but currently, this program does not convey specifics around teaching evaluation. All potential participants for this study (N = 73) were emailed directly with detailed information about the study and asked to participate in a 45-minute semi-structured interview.

The interview protocol consisted of questions regarding the participants' experiences evaluating teaching using SET, and it also included two SET case studies based on real instructor ratings (see Table 2) that participants were asked to interpret and evaluate. Interviews were audio recorded, transcribed, and coded by research team members into NVivo 11. A hybrid thematic analysis approach was conducted on each of the transcribed interviews. Fereday and Muir-Cochrane (2006) describe thematic analysis as "a search for themes that emerge as being important to the description of the phenomenon" (p. 82). This hybrid approach included both identifying themes through careful reading of the literature and reading of the data. Therefore, interview excerpts were coded in NVivo 11 using nodes previously generated from the conceptual framework, and also from recurring themes found in these interviews that went beyond the conceptual framework. Interview data were *quantitized*, which "refers to the process of assigning numerical (nominal or ordinal) values to data conceived as not numerical" (Sandelowski, 2009, pp. 209-210). This provided descriptive statistical information about these interviews. Finally, an external research assistant determined an inter-coder percentage agreement of 95% with 10% of the total number of interviews.

Findings

Strengths and Weaknesses of Evaluations

Forty-three department chairs (or their equivalents in non-departmental faculties) participated in the study, which corresponded to 59% of the total number of this type of administrator. Most of them (86%) used SET results (i.e., both ratings and comments) in their teaching evaluation process. A few of them (4.7%), though, did not use the prescribed SET even when their use was part of the institution's policy. These administrators believed that, due to the nature of their departments, the institutional SET did not capture the essence of their teaching practices, and consequently they had decided to discontinue their use for teaching evaluation purposes and used alternative student feedback mechanisms. The rest (9.3%) did not provide a clear response or did not seem to know about the SET: "I have never seen this. Your email was the first time that I heard the term ever" (Participant 43). As a result, all of the following findings will only consider participants who reported using the institution's prescribed SET, since the purpose of this study was to explore how these tools are used by administrators in the teaching evaluation process. Thus, these administrators were asked to explain how they examine the SET information. According to one respondent:

> We key in right away on "overall the instructor was excellent." You always look at that one first. "Overall the course content was excellent" is the second thing you look at. And then, only if there's problems in either of those two scores, you look in more detail at the other questions. There's around 300 faculty members in our FEC [Faculty Evaluation Committee], so we're only finding ways to efficiently go through SET. (Participant 01)

Although most participants considered all SET statements during their initial examination of the results, they seemed to primarily focus on only a few out of the 10 institution-wide items. Almost all participants (97.3%) indicated centring their attention on "overall this instructor was excellent," more than half (67.6%) on "overall the quality of the course content was excellent," and only one-third (35.1%) on "the instructor treated students with respect." The remaining seven statements were only identified by 20% or fewer of participants as statements commonly considered in their teaching evaluation process (see Table 1). Administrators believed that focusing on a few seemingly important statements allowed them to be more efficient during biannual faculty evaluation committee (FEC) meetings, especially since they would still examine the rest of the items if any major concerns were raised after an initial examination.

Table 1. SET Statements Usage in the Teaching Evaluation Process

Statement	Percentage
The goals and objectives of the course were clear.	21.6%
In-class time was used effectively.	16.2%
I am motivated to learn more about these subject areas.	24.3%
I Increased my knowledge of the subject areas in this course.	18.9%
Overall, the quality of the course content was excellent.	67.6%
The instructor spoke clearly.	2.7%
The instructor was well prepared.	16.2%
The instructor treated students with respect.	35.1%
The instructor provided constructive feedback throughout this course.	10.8%
Overall, the instructor was excellent.	97.3%

Note. All items are rated on a 5-point Likert scale from (1) strongly disagree to (5) strongly agree.

One respondent reported:

> If somebody gets a high score on [the above-mentioned] items, generally the remaining ones will also be high. If an instructor does not get a good score on those items, you'll see mixed scores in the remaining ones. So, these other statements actually give you an idea of what specifically went wrong in the course. (Participant 03)

Therefore, even though they still recognized that all statements were important to more holistically determine what went on in the course, only one or two SET statements were actually used by administrators to determine if the teacher was effective in that course. Administrators had benchmarks in mind as they initially reviewed SET results, indicating whether they needed to pay more attention to some of the other SET items. For them, this was a way to make the teaching evaluation process more efficient, especially when having to evaluate a large number of instructors.

Knowledge and Support Gaps in the Evaluation Process

While having such benchmarks in mind might aggravate some of the intrinsic SET issues, administrators stated that they were aiming for efficiency and had no alternative methods for evaluating teaching because of the lack of training and supports for department chairs. Most participants (83.7%), regardless of the number of years in their administrative roles, were determined when voicing their need for actual supports to evaluate teaching better. One participant said, "I was hoping the result of this study would give me some ideas of what this [teaching evaluation] actually was" (Participant 24). In fact, one-quarter (27.9%) even suggested the institution should be the one to provide clear guidelines to aid them to evaluate teaching in a fair, equitable and meaningful way. Another participant revealed:

> My learning curve coming in to the chair role has been huge. We used to have the chairs' school, but now there's only the leadership college, and it's a very differ-

> ent thing. So, now you transition into your chair role on your own. You've got to go figure it out on your own or ask people for coffee to ask questions and learn up because there's no real orientation to being a chair. (Participant 42)

Many of these administrators found learning the components of their new role overwhelming, considering they were not aware of any institutional supports to assist them with learning these components. Some (20.9%) highlighted that it was critical for newly appointed administrators to have access to teaching evaluation training. In addition, they identified a need for assistance in determining appropriate instruments that can be used to both support and evaluate teaching, such as discipline-specific concept inventories to better determine student knowledge increase (11.6%), peer-support initiatives to assess and improve pedagogical practices (11.6%), video-recorded lectures for later review and analysis (7%), individual pedagogical self-reflections to comprehend instructors' teaching approaches (7%), and class materials to obtain a more complete overview of the course (4.7%). Indeed, participants not only felt unsupported, but also felt that they were without the necessary knowledge when transitioning to their new administration roles, and particularly when attempting to understand how to evaluate teaching. One participant commented:

> We need support to develop our own teaching evaluation skills more comfortably, so we can help develop excellent teachers. But it is also important to make sure our instruments are valid, and that we can actually use them on a journey of self-improvement. And to do that, having some facilitation from people who can work with us and help us would be better than just having a list on a website. That's not enough. (Participant 39)

Administrators understood that receiving institutional support was critical not only to improve the currently available teaching evaluation process but also to improve teaching itself. As one respondent said, "We don't expect people to be great teachers initially, but we expect them to improve based on student feedback. So, we help them find pathways to improve the quality of their teaching reflected on their teaching evaluations" (Participant 02). Hence, more than having access to online repositories with decontextualized evaluation practices, what administrators truly required was access to different contextualized tools that are useful to assess and improve teaching in their departments. For many of them, getting this support would be the first step toward assessing teaching in a fair, equitable, and meaningful way.

Potential Biases Affecting Responses

Improving the teaching evaluation process was important for participants, and they understood that they needed to adopt supplementary tools beyond the use of SET to better evaluate teaching. They expressed their concerns about the inability of SET to effectively assess diverse approaches to teaching (46.5%), noticing lower SET ratings for female and visible minority instructors (11.6%), and their futility at persuading some tenure-faculty to improve their teaching (9.3%). One participant revealed:

> I know one female member of my department that most people would agree she is a very good instructor, and even the SET written comments support it, but her

> SET scores are too low in comparison to her peers. It's concerning, it's unfair, and I think it's a big issue. I also find that people with strong accents are at a disadvantage. I know another colleague who is not as easy to understand, and some of our students struggle for a few weeks. But then students seem to have a hard time differentiating between his accent as the one and single issue, and the fact that otherwise the instructor was prepared, knowledgeable, and organized. (Participant 14)

Although many participants (67.5%) admitted the need to move beyond the use of SET because of their concerns regarding biases and validity, others (27.9%) confessed not having enough time or resources to implement a multifaceted teaching evaluation approach. They even suggested that perhaps, considering these restricted conditions, the statements should be reconsidered to better represent the wide diversity of teaching practices and instructors. One participant observed:

> That question set doesn't serve the diversity and the kind of pedagogy we have now, and really needs fixing. I think there needs to be a conversation about what this is going to look like over time. I also think the institution has to take very seriously the concerns that equity-seeking groups have about what happens in teaching evaluations. What happens to women? What happens to visible minority? What happens to people that are perceived to have strong accents? And I think there's a huge responsibility on chairs on FEC to really be educated and understand how much you can extrapolate from SET. (Participant 42)

Therefore, even though participants were aware of the need to improve the teaching evaluation process, they also admitted that they face several challenges in order to actually implement a multifaceted approach. This was a concern for these administrators not only because they were largely dissatisfied with the effectiveness of SET, but also because they were aware of the various potential biases affecting these ratings. As a result, they openly acknowledged how these issues intensified when relying exclusively on SET to evaluate teaching.

Willingness to Provide Feedback

Most participants (67.5%) believed that the SET response rate has decreased since 2014, when the institution implemented an online evaluation system instead of a paper-based one. Another group (21.6%) estimated that there was a similar student response rate to the paper-based method previously implemented, and only a few (8.1%) believed that response rates had increased when the online system was implemented. One participant responded that

> [Students] get completely annoyed because they're being bombarded with emails and professors in their last week of classes reminding them to do SET. I think they just go: 'I'm really annoyed! I'm not going to do them at all.' I don't know what system they use, but it's almost like every student receives one for every class. So, they're just harassing them to death, and they get mad about it. (Participant 27)

Many participants had analyzed data from their own SET results or from their department's general student response rate from previous years to back-up these claims; others were only echoing changes in the student response rate with little or no data to support their views. Nevertheless, when prompted to reflect about the possible explanations for these changes as well as the overall student reluctance to provide feedback, some participants (8.1%) suggested that a major issue with SET response rates was that students are continually asked to volunteer for studies, complete assessments, and respond to the SET. One respondent said, "I give time at the beginning of class for students to complete the online evaluation, and I tell them how critical it is to get feedback about the course and why we do with the information" (Participant 18). The perception of students' general unwillingness to provide feedback was an additional concern for these administrators when relying solely on SET to evaluate teaching.

Views about the Evaluation Process

Regardless of these institutional, resource, and time limitations, many participants had already implemented supplementary tools and additional information sources to move toward a multifaceted approach and better evaluate teaching in their departments. Peer in-class teaching observations (70.3%), individual pedagogical self-reflections (37.8%), examination of class materials (i.e., syllabi, assignments, exams; 29.7%), and department-specific surveys (21.6%), were the most commonly employed extra tools. These participants were truly aware of the importance of including one or more of these additional sources of information and not relying exclusively on SET to evaluate teaching. One participant commented:

> I don't think that SET are very useful by themselves. They're incomplete, and I'd feel uncomfortable judging somebody's fate just based on that. I'm not saying SET are wrong, but they're only one piece of understanding. We take teaching seriously, and it's not just a bunch of simple numbers pouring at us. We don't just look at 'you're above this number, or below this number, and we're done.' We're looking at you much more carefully than that, but SET are a good start. (Participant 29)

Administrators understood the potential impact of teaching evaluations and what was at stake as a result of this process. However, the implementation of these tools varied in different departments. Some administrators only requested additional information on a voluntary basis (37.8%), while others implemented supplementary tools as a departmental standard (27%). A few others only used other tools when evaluating teaching for tenure purposes (18.9%), or when assessing sessional lecturers and new instructors (10.8%). Only a small group indicated not using any kind of additional tools or information for evaluating teaching beyond SET (5.4%). One participant commented:

> SET are useful, as long as they're not used by themselves. They're supplemented by all kinds of other measures, but at the end of the day, there's not one correct measurement. But by having several different kinds of measurements, hopefully we at least get a better idea about what instructors are doing. They know we're paying attention because we talk to them as well, so we can get a fairly deep understanding of what's going on, and that we're not just mechanically following those numbers. (Participant 29)

A small group of participants (8.1%), motivated to attain a fair, equitable, and meaningful teaching evaluation in their departments, had even implemented yearly audits that included a wider selection of supplementary tools and sources of information. These audits, nonetheless, were only possible by reviewing a portion of their professorate each year.

Many administrators also indicated that an array of contextual factors was considered in the teaching evaluation process, and that they tried to obtain as much information as possible to provide an informed interpretation of SET results. This approach was made more explicit during the two SET case studies that participants were asked to evaluate (see Table 2), when a considerable group of administrators (45.9%) refrained from assessing these results due to the importance of the missing contextual factors in both sample cases.

Table 2. Sample SET Case Studies

Instructor A		Tukey	Reference Data		
Statement	Median	Fence	25%	50%	75%
The goals and objectives of the course were clear.	3.4	2.7	3.9	4.3	4.7
In-class time was used effectively.	3.6	2.5	3.8	4.3	4.7
I am motivated to learn more about these subject areas.	3.5	2.9	4.1	4.5	4.8
I increased my knowledge of the subject areas in this course.	4.4	3.0	4.1	4.6	4.8
Overall, the quality of the course content was excellent.	3.8	2.4	3.8	4.3	4.8
The instructor spoke clearly.	4.5	3.8	4.5	4.8	4.9
The instructor was well prepared.	4.6	3.4	4.3	4.8	4.9
The instructor treated the students with respect.	4.0	4.2	4.7	4.9	5.0
The instructor provided constructive feedback throughout this course.	4.5	2.8	4.0	4.5	4.8
Overall, this instructor was excellent.	4.0	3.2	4.2	4.7	4.9
Instructor B		Tukey	Reference Data		
Statement	Median	Fence	25%	50%	75%
The goals and objectives of the course were clear.	4.0	2.7	3.9	4.3	4.7
In-class time was used effectively.	4.2	2.5	3.8	4.3	4.7
I am motivated to learn more about these subject areas.	3.7	2.9	4.1	4.5	4.8
I increased my knowledge of the subject areas in this course.	4.1	3.0	4.1	4.6	4.8
Overall, the quality of the course content was excellent.	4.2	2.4	3.8	4.3	4.8
The instructor spoke clearly.	4.7	3.8	4.5	4.8	4.9
The instructor was well prepared.	4.4	3.4	4.3	4.8	4.9
The instructor treated the students with respect.	4.8	4.2	4.7	4.9	5.0
The instructor provided constructive feedback throughout this course.	4.0	2.8	4.0	4.5	4.8
Overall, this instructor was excellent.	4.5	3.2	4.2	4.7	4.9

Note. All items are rated on a 5-point Likert scale from (1) strongly disagree to (5) strongly agree.

One respondent stated:

> In the abstract, I don't know what I would say, without knowing the circumstances. If one of those instructors is in their first year of teaching, and the other was an experienced professor, I think that interpretation is dramatically different than if they're both experienced professors or if they're both new professors. If we look at the overall averages I can say they're both scoring in the lower percentile, and that sort of thing, but to be perfectly honest that means very little to me, because I think that understanding a person's position is crucial to being able to read any of these numbers. (Participant 04)

What mechanisms are administrators employing to evaluate teaching in a meaningful way despite all their time and resource limitations? How are they approaching an equitable and fair evaluation of teaching when admittedly they have concerns about SET? Many participants showed that, when possible, they were implementing supplementary tools and additional information to evaluate teaching. They comprehended the potential impact of the teaching evaluation process and recognized that SET were only one piece of understanding. Unfortunately, institutional, resource, and time limitations and the limitation that additional evidence is mostly obtained on a voluntary basis have hindered administrators' explicit efforts to improve this process. Nevertheless, many administrators still defended their use of a contextual approach to inform their interpretation of SET results whenever it was feasible.

Discussion and Conclusions

Although many institutional policies in higher education encourage administrators to employ a multifaceted approach to evaluate teaching, the reality is that most departments heavily rely on SET results alone. Thus, it was relevant to explore how these tools are used by administrators to evaluate teaching despite their limitations.

Taken together, these findings showed that although participants mostly utilized only a couple of SET statements as indicators of effective teaching, they were certainly aware of the intrinsic issues concerning these tools. Administrators indicated devoting a lot of energy into implementing multifaceted evaluations of teaching, but reported meeting obstacles along the way. These hurdles were not only due to limited time, resources, and institutional support, but also because some of the administrators only obtained supplementary information on a voluntary basis.

However, despite these challenges, administrators attempted to navigate SET by trying to get a contextualized understanding of the courses, the students, and most importantly the instructors. Participants believed that, without any contextual information, these results remained abstract ratings that did not provide the necessary information to evaluate teaching in a fair, equitable, and meaningful way. One participant mentioned, "I'd talk to the instructor and ask what's going on in the class. I would support them and help them try to build a case. And we would get ready together, and stick together, before FEC" (Participant 31). Although administrators seemingly had benchmarks in mind as they initially examined ratings and even when they could interpret the numbers on their own, they highlighted that without having any context, they could not appropriately assess SET results.

These findings are noteworthy considering that SET validity and reliability are still under continuous scrutiny, especially for making decisions about faculty retention, promotion, tenure and pay increments. One participant commented, “Somebody has to teach the ‘broccoli’ course, not everybody gets to teach the ‘dessert.’ Especially when you get into courses which are heavily directed towards application, you are forced to give more critical feedback, which tends to be unpopular” (Participant 26). This study helped unveil that administrators are certainly aware of these issues; that they understand “broccoli” courses are unpopular, are often required, and are classes that get lower than average ratings; and that they are continually seeking to obtain more evidence if SET results are below their benchmarks. In other words, these administrators would not take poor SET results at face value, but instead would further investigate to see what contextual factors could have contributed to the ratings. Although participants are still trying to implement a multifaceted teaching evaluation process, the overall strategy they keep implementing is consistent with the constraints and limitations they repeatedly face.

Previous studies have focused on the potential biases with these tools—the correlation between expected grades and ratings, the deficiencies regarding the validity and reliability of these tools, the small impact on teaching quality, and the concerns related to the use of these tools for summative evaluation purposes. To our knowledge, however, none of these studies have addressed how administrators navigate SET despite their perceived limitations. However, these findings only represent the experience of administrators in a large research-intensive Canadian university, and they may not be representative of different institutions. Thus, future research should address how administrators in other institutions navigate SET and evaluate teaching. Future studies should also address how to help administrators develop a multifaceted teaching evaluation process that fosters a sustainable approach that could be implemented despite the limitations they face.

Lastly, although these findings are drawn from data collected in only one institution, other higher education institutions may begin to understand the various challenges and opportunities that administrators face when navigating SET for teaching evaluation purposes. The lessons learned in this study could encourage institutions to implement different events (e.g., workshops, sessions, etc.) to help administrators become aware of such challenges and opportunities, and fairly assess teaching using SET as part of the standard evaluation process. Similarly, these findings could help administrators recognize the importance of better communicating to faculty how they use SET results, so that faculty realize many administrators navigate SET results with a contextual approach.

References

Ahmadi, M., Helms, M. M., & Raiszadeh, F. (2001). Business students’ perceptions of faculty evaluations. *International Journal of Educational Management, 15*(1), 12–22. doi:10.1108/09513540110366097

Al Kuwaiti, A., AlQuraan, M., Subbarayalu, A. V., & Piro, J. S. (2016). Understanding the effect of response rate and class size interaction on students evaluation of teaching in a higher education. *Cogent Education, 3*(1). doi:10.1080/2331186x.2016.1204082

Al-Eidan, F., Baig, L. A., Magzoub, M., & Omair, A. (2016). Reliability and validity of the faculty evaluation instrument used at King Saud bin Abdulaziz University for Health Sciences: Results from the haematology course. *Journal of Pakistan Medical Association, 66*(4), 453–457.

Asassfeh, S., Al-Ebous, H., Khwaileh, F., & Al-Zoubi, Z. (2013). Student faculty evaluation (SFE) at Jordanian universities: A student perspective. *Educational Studies, 40*(2), 121–143. doi:10.1080/03055698.2013.833084

Backer, E. (2012). Burnt at the student evaluation stake: The penalty for failing students. *e-Journal of Business Education & Scholarship of Teaching, 6*(1), 1–13.

Bassett, J., Cleveland, A., Acorn, D., Nix, M., & Snyder, T. (2015). Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations. *Assessment & Evaluation in Higher Education, 42*(3), 431–442. doi: 10.1080/02602938.2015.1119801

Bedggood, R. E., & Donovan, J. D. (2012). University performance evaluations: What are we really measuring? *Studies in Higher Education, 37*(7), 825–842. doi:10.1080/03075079.2010.549221

Berk, R. A. (2013). Top five flashpoints in the assessment of teaching effectiveness. *Med Teach, 35*(1), 15–26. doi:10.3109/0142159X.2012.732247

Blackhart, G. C., Peruche, M., DeWall, C. N., & Joiner, T. E. (2006). Faculty forum: Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33*(1), 37–39. doi:10.1207/s15328023top3301_9

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*(1), 1-11. doi:10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching. *Teaching of Psychology, 42*(2), 109–118. doi:10.1177/0098628315569922

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2013). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641–656. doi:10.1080/02602938.2013.860950

Brown, G. D., Wood, A. M., Ogden, R. S., & Maltby, J. (2015). Do student evaluations of university reflect inaccurate beliefs or actual experience? A relative rank model. *Journal of Behavioral Decision Making, 28*(1), 14–26. doi:10.1002/bdm.1827

Brown, M. J. (2008). Student perceptions of teaching evaluations. *Journal of Instructional Psychology, 35*(2), 177–181.

Campbell, J. P., & 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. doi:10.1080/10668920600864137

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495–518. doi:10.1023/A:1025492407752

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*(1), 17–33. doi:10.2307/2649280

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education, 28*(1), 71–88. doi:10.1080/02602930301683

Cho, D., Baek, W., & Cho, J. (2015). Why do good performing students highly rate their instructors? Evidence from a natural experiment. *Economics of Education Review, 49*, 172–179. doi:10.1016/j.econedurev.2015.10.001

Cho, J., Otani, K., & Kim, B. J. (2014). Differences in student evaluations of limited-term lecturers and full-time faculty. *Journal on Excellence in College Teaching, 2*, 5–24.

Chonko, L. B., Tanner, J. F., & Davis, R. (2002). What are they thinking? Students' expectations and self-assessments. *Journal of Education for Business, 77*(5), 271–281. doi:10.1080/08832320209599676

Clayson, D. E. (2013). Initial impressions and the student evaluation of teaching. *Journal of Education for Business, 88*(1), 26–35. doi:10.1080/08832323.2011.633580

Cox, C. D., Peeters, M. J., Stanford, B. L., & Seifert, C. F. (2013). Pilot of peer assessment within experiential teaching and learning. *Currents in Pharmacy Teaching and Learning, 5*(4), 311–320. doi:10.1016/j.cptl.2013.02.003

Curwood, J. S., Tomitsch, M., Thomson, K., & Hendry, G. D. (2015). Professional learning in higher education: Understanding how academics interpret student feedback and access resources to improve their teaching. *Australasian Journal of Educational Technology, 31*(5), 556–571. doi:10.14742/ajet.2516

Dodeen, H. (2013). Validity, reliability, and potential bias of short forms of students' evaluation of teaching: The case of UAE University. *Educational Assessment, 18*(4), 235–250. doi:10.1080/10627197.2013.846670

Dolmans, D. H., Janssen-Noordman, A., & Wolfhagen, H. A. (2006). Can students differentiate between PBL tutors with different tutoring deficiencies? *Med Teach, 28*(6), 156–161. doi:10.1080/01421590600776545

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education, 29*(1), 91–108. doi:10.1080/0260293032000158180

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80–92.

Fraile, R., & Bosch-Morell, F. (2014). Considering teaching history and calculating confidence intervals in student evaluations of teaching quality: An approach based on Bayesian inference. *Higher Education, 70*(1), 55–72. doi:10.1007/s10734-014-9823-0

Gaillard, F. D., Mitchell, S. P., & Vahwere, K. (2006). Students, faculty, and administrators perception of students evaluations of faculty in higher education business schools. *Journal of College Teaching & Learning, 8*(3), 77–90. doi:10.19030/tlc.v3i8.1695

Gehrt, K., Louie, T. A., & Osland, A. (2014). Student and professor similarity: Exploring the effects of gender and relative age. *Journal of Education for Business, 90*(1), 1–9. doi: 10.1080/08832323.2014.968514

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603–615. doi:10.1080/03075070701573773

Grammatikopoulos, V., Linardakis, M., Gregoriadis, A., & Oikonomidis, V. (2014). Assessing the students' evaluations of educational quality (SEEQ) questionnaire in Greek higher education. *Higher Education, 70*(3), 395–408. doi:10.1007/s10734-014-9837-7

Grayson, J. P. (2015). Repeated negative teaching evaluations: A form of habitual behaviour? *Canadian Journal of Higher Education, 45*(4), 298–321.

Gump, S. E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly, 30*(3), 56–69.

Huebner, L., & Magel, R. C. (2015). A gendered study of student ratings of instruction. *Open Journal of Statistics, 05*(06), 552–567. doi:10.4236/ojs.2015.56058

Hughes, K. E., & Pate, G. R. (2013). Moving beyond student ratings: A balanced scorecard approach for evaluating teaching performance. *Issues in Accounting Education, 28*(1), 49–75. doi:10.2308/iace-50302

Iqbal, I., Lee, J. D., Pearson, M. L., & Albon, S. P. (2016). Student and faculty perceptions of student evaluations of teaching in a Canadian pharmacy school. *Currents in Pharmacy Teaching and Learning, 8*(2), 191–199. doi:10.1016/j.cptl.2015.12.002

Jackson, M. J., & Jackson, W. T. (2015). The misuse of student evaluations of teaching: Implications, suggestions and alternatives. *Academy of Educational Leadership Journal, 19*(3), 165–173.

Jones, J., Gaffney-Rhys, R., & Jones, E. (2012). Handle with care! An exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results. *Journal of Further and Higher Education, 38*(1), 37-56. doi:10.1080/0309877x.2012.699514

Khong, T. L. (2014). The validity and reliability of the student evaluation of teaching. *International Journal for Innovation Education and Research, 2*(9), 57–63.

Kim, L. E., & MacCann, C. (2016). What is students' ideal university instructor personality? An investigation of absolute and relative personality preferences. *Personality and Individual Differences, 102*, 190-203. doi:10.1016/j.paid.2016.06.068

Lama, T., Arias, P., Mendoza, K., & Manahan, J. (2015). Student evaluation of teaching surveys: Do students provide accurate and reliable information? *e-Journal of Social & Behavioural Research in Business, 6*(1), 30–39.

Lindahl, M. W., & Unger, M. L. (2010). Cruelty in student teaching evaluations. *College Teaching, 58*(3), 71-76. doi:10.1080/87567550903253643

Lyde, A. R., Grieshaber, D. C., & Byrns, G. (2016). Faculty teaching performance: Perceptions of a multi-source method for evaluation. *Journal of the Scholarship of Teaching and Learning, 16*(3), 82–94. doi:10.14434/josotl.v16i3.18145

Macfadyen, L. P., Dawson, S., Prest, S., & Gašević, D. (2015). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education, 41*(6), 821–839. doi:10.1080/02602938.2015.1044421

MacNell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303. doi:10.1007/s10755-014-9313-4

Makondo, L., & Ndebele, C. (2014). University lecturers' views on student-lecturer evaluations. *The Anthropologist: International Journal of Contemporary and Applied Studies of Man, 17*(2), 377–386.

Martin, L. R., Dennehy, R., & Morgan, S. (2013). Unreliability in student evaluation of teaching questionnaires: Focus groups as an alternative approach. *Organization Management Journal, 10*(1), 66–74. doi:10.1080/15416518.2013.781401

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33*(3), 176–179. doi:10.1207/s15328023top3303_4

Mengel, F., Sauermann, J., & Zölitz, U. (2017). Gender bias in teaching evaluations. *IZA Discussion Paper No. 11000.*

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (4th ed.). San Francisco, CA: John Wiley & Sons.

Merritt, D. J. (2012). Bias, the brain, and student evaluations of teaching. *St. John's Law Review, 82*(1), 235–288.

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

Mitry, D. J., & Smith, D. E. (2014). Student evaluations of faculty members: A call for analytical prudence. *Journal on Excellence in College Teaching, 25*(2), 56–67.

Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation, 38*(1), 15–20. doi:10.1016/j.stueduc.2012.01.001

Nargundkar, S., & Shrikhande, M. (2012). An empirical investigation of student evaluations of instruction: The relative importance of factors. *Journal of Innovative Education, 10*(1), 117–135. doi:10.1111/j.1540-4609.2011.00328.x

Palmer, S. (2012). Student evaluation of teaching: keeping in touch with reality. *Quality in Higher Education, 18*(3), 297–311. doi:10.1080/13538322.2012.730336

Rantanen, P. (2013). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education, 38*(2), 224–239. doi:10.1080/02602938.2011.625471

Reisenwitz, T. H. (2015). Student evaluation of teaching: An investigation of nonresponse bias in an online context. *Journal of Marketing Education, 38*(1), 7–17. doi:10.1177/0273475315596778

Ridley, D., & Collins, J. (2015). A suggested evaluation metric instrument for faculty members at colleges and universities. *International Journal of Education Research, 10*(1), 97–114.

Royal, K. D., & Stockdale, M. R. (2015). Are teacher course evaluations biased against faculty that teach quantitative methods courses? *International Journal of Higher Education, 4*(1). doi:10.5430/ijhe.v4n1p217

Sandelowski, M. (2009). On quantitizing. *Journal of Mixed Methods Research, 3*(3), 208–222. doi:10.1177/1558689809334210

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication, 30*(1), 64–77. doi:10.1080/07491409.2007.10162505

Socha, A. (2013). A hierarchical approach to students' assessments of instruction. *Assessment & Evaluation in Higher Education, 38*(1), 94–113. doi:10.1080/02602938.2011.604713

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. doi:10.3102/0034654313496870

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. doi:10.1080/02602938.2013.767876

Stonebraker, R. J., & Stone, G. S. (2015). Too old to teach? The effect of age on college and university professors. *Research in Higher Education, 56*(8), 793–812. doi:10.1007/s11162-015-9374-y

Uijtdehaage, S., & O'Neal, C. (2015). A curious case of the phantom professor: Mindless teaching evaluations by medical students. *Medical Education, 49*(9), 928–932. doi:10.1111/medu.12647

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. doi:10.1016/j.stueduc.2016.08.007

Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger teachers. *College Teaching, 62*(1), 20–24. doi:10.1080/87567555.2013.825574

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683–699. doi:10.1080/02602938.2011.563279

Zhu, C. J. Y., White, D., Rankin, J., & Davison, C. J. (2018). Making meaning from student evaluations of teaching: Seeing beyond our own horizons. *Teaching & Learning Inquiry, 6*(2), 127–142. doi:10.20343/teachlearninqu.6.2.10

Zumbach, J., & Funke, J. (2014). Influences of mood on academic course evaluations. *Practical Assessment, Research & Evaluation, 19*(4), 1–12.

Contact information

Luis Francisco Vargas-Madriz
University of Alberta
fran.vargas@ualberta.ca

Luis Francisco Vargas-Madriz is the Senior Research Coordinator at the University of Alberta's Centre for Teaching and Learning. His research interests include inclusive education, blended and online learning, 21st-century skills, social media, and assistive technology.

Norma Nocente is the Associate Director (Educational Technology) at the Centre for Teaching and Learning and an Associate Professor in the Department of Secondary Education at the University of Alberta. Her research interests include technology integration and student learning.

Rebecca Best-Bertwistle has a BA in Political Science from the University of Alberta and her research interests include public engagement and conservation. She is currently employed in the non-profit sector in community engagement and works on issues of conservation, wildlife, and public land management in southern Alberta.

Sarah Forgie is the Vice Provost (Learning Initiatives) at the University of Alberta. She is also a professor of pediatrics in the Faculty of Medicine and Dentistry. Her current research interests include faculty development in teaching and learning and the implementation and assessment of innovative teaching and learning approaches.