Institutional Approaches to Evaluate Teaching Effectiveness: The Role of Summative Peer Review of Teaching for Promotion and Tenure

Keif Godbout-Kinney & Gavan P. L. Watson
Memorial University of Newfoundland and Labrador

Abstract

A growing body of literature has identified student evaluations of teaching (SETs) as introducing bias against minority faculty members and not serving as a reliable or valid measure of teaching effectiveness. This lack of reliability and validity presents issues for university tenure and promotion committees, as these institutional processes necessarily require accurate, objective, and holistically informed modes of evaluation to recognize teaching achievements. Summative peer review of teaching (SPRT) is an alternative mode of assessment that aims to provide evidence of teaching effectiveness to inform promotion and tenure. SPRT, as an institutional practice, has been adopted at a small cohort of institutions of higher education, marking a potential shift in practice. This article examines SETs to articulate the problematic elements introduced by SETs, specifically to examine if SPRT can serve as a viable alternative. By describing the SPRT processes that four institutions have taken, the authors aim to articulate these emerging approaches to collecting evidence of teaching effectiveness. In this descriptive work, it is our secondary contention that SPRT, through intentional design and facilitation, can offer a process that does not introduce bias in the same way as SETs and thus, can also be used to satisfy the growing need for practices that help achieve, in part, institutional goals related to equity, diversity, and inclusion (EDI).

Keywords: Student evaluations of teaching, promotion and tenure, summative peer review, Canadian universities, EDI

Introduction

In higher education, teaching is a foundational value (Bharuthram, 2012). Yet, the evaluation of teaching effectiveness continues to remain elusive. Elusive, perhaps, when compared to evaluating research productivity: rather than counting publications, reporting bibliometric data, or a count of funding secured, effective teaching is not so easily measured. Chism (2007) found that "when teaching is viewed as professional knowledge, there must be an accepted way to define characteristics of teaching excellence and to make judgments based on a stated set of criteria and standards that reflect the complexity of teaching" (p. 13). Teaching effectiveness remains challenging to measure and account for (Jereb et al., 2018), and the issue of how to measure teaching practice has been a rich topic of debate (Bernstein et al., 2006). The most widespread institutional process associated with measuring and reporting teaching effectiveness are survey instruments called—broadly—student evaluations of teaching (SET). SETs' proliferation, though often varied in processes, metrics, and designs, has been critiqued as problematic as a tool for evaluating teaching effectiveness (Hammersley-Fletcher & Orsmond, 2004; Hornstein, 2017; Peterson et al., 2019; Uttl, 2017). Yet, SETs persist.

In this work, the authors describe the shortcomings of using SETs as a regular part of a faculty

evaluative process and offer the summative peer review of teaching (SPRT) model as an alternative for institutions to consider. It is not our contention that SPRT is itself a novel practice, rather, that as an administrative and institutionally supported process to inform promotion and tenure, SPRT has not yet been widely adopted. To help illustrate a way forward with SPRT, we turn to the field to describe current SPRT practices used to inform promotion and tenure at four universities: two Canadian, one American, and one Australian. We close by detailing concerns for institutions to consider with a transition to SPRTs. Our overall purpose is to describe how SPRT could be implemented as an institutional process to offer evidence of teaching effectiveness to inform the promotion and tenure process while addressing, in part, the limitations articulated concerning SETs.

Critiques of SETs

SETs have been, to date, the widespread institutional mechanism through which individual instructors' teaching quality is evaluated. SETs are most commonly understood as standardized evaluations where students rate their instructor and learning experience towards the end of a course. Questions often ask students to rate answers on a Likert scale and can include categories such as instructor clarity, instructor enthusiasm or passion, if instructors were available to meet and how often, and if the evaluations were fair (Murray, 2005). These groups of institutional surveys are ubiquitous in the academe because they have been seen, relatively speaking to other forms of assessment of teaching, as being easily implemented and not significantly resource-intensive for the scale of assessment. However, a body of literature on teaching evaluations indicates there are problems with SETs: they lack validity, lack reliability, and do not uphold the principles of EDI (Griffen, 2021). We summarize these concerns below.

A seemingly universal critique of current SETs is that they appear to measure how much an instructor was *liked* by their students rather than how *effective* or knowledgeable they were as an instructor (Hornstein, 2017). A recent meta-analysis of the literature by Uttle et al. (2017) argued that an underpinning logic of all SETs is an assumption that students learn more from instructors who have higher ratings. The authors demonstrate how studies in cognitive psychology indicate that any correlations between SETs and student learning are more likely to be a "fluke" (p. 23) rather than any kind of actually valid measure of teaching effectiveness.

The next emergent theme in the literature surrounding SETs details how the instruments are not practical measures of teaching efficacy but rather are indications of reliability (Kreitzer & Sweet-Cushman, 2021). That is, SETs seems to show if a diverse range of students rate an instructor similarly or dissimilarly. On average, men-identified faculty tend to score better on SETs than women-identified faculty and are often perceived as more competent, organized, and professional (p. 2). The implications of low reliability in SETs for black, Indigenous, and people of colour (BIPOC) and women faculty are highly damaging, especially within the context of formal evaluation processes. These lower evaluations result in fewer promotions, tenures, or awards, and it negatively affects groups who already suffer from low representation and poor treatment in the academe.

Concerning gender, the literature suggests there is evidence that gender biases exist within SETs, to the point where even seemingly objective factors result in different scores based on the instructor's gender (Hornstein, 2017; Kreitzer & Sweet-Cushman, 2021). According to Hornstein (2017), this bias is "statistically significant" (p. 5), and he firmly asserted that it is irresponsible, unconscionable, and might even be illegal for universities in North America to continue to employ SETs for evaluative purposes.

In undertaking this work, the authors conducted an extensive review with the intention to present the broad scope of literature that both supported and critiqued the use of SETs. Through our review, which included several meta-analyses focused on the history, application, and current utilization of SETs in higher education, it became evident to the authors that very little of the current published work supports the continued use of SETs as a tool for promotion and tenure. As such, and with the risk of appearing biased, we believe that we are justified in articulating the limitation of SETs without incorporating research that supports their validity. This is not to say that SETs are without value altogether; indeed, they can serve as a helpful metric to inform instructors about students' feelings or thoughts about a course in terms of what does or does not work from a student's perspective. The evidence, however, is that as a tool to inform promotion and tenure, SETs are imperfect, primarily due to their introduction of bias and, in some cases, a lack of validity.

Thus, given this evidence, developing an alternative evaluation methodology of teaching effectiveness would not solely rely on SETs but instead, could draw on the process of peer review. Our assumption here is that peer review, used for summative evaluation purposes, would allow academic administrators and academic unit promotion and tenure committees to judge instructors' teaching more effectively, primarily because they would have more comprehensive data to draw on when making these evaluative decisions. We also theorize that moving away from SETs would have the added benefit of implementing more equitable review practices. Exploring summative peer review is where we turn next.

Introducing Summative Peer Review of Teaching

To address the shortcomings of SETs, a small cohort of Canadian Universities have implemented a new institutional approach to collecting evidence of teaching effectiveness: summative peer review. While informally inviting peers' feedback is not a novel approach within the academe, work to develop formal institutional processes is an emerging practice. The term *summative* before peer review marks that the process is being used for evaluative purposes rather than solely for continuous improvement. As an approach, SPRT is a form of peer review that examines teaching for various outcomes such as effectiveness, rigor, knowledge of the material, and preparation. As the name suggests, SPRT takes place between faculty colleagues and should be an informed process to highlight areas for teaching improvement, evaluation for awards and promotion, and to make decisions regarding personnel. SPRT is also useful to university administrators because of the comparability it allows. That is, reviewers can review specific characteristics against pre-established standards and guidelines, theoretically allowing for better evaluation regarding tenure and promotion (Chism, 2007). But SPRT is not a homogenous approach. These processes draw upon various evaluative methods such as syllabus evaluation, classroom observation, or interviews with students to establish measures of teaching effectiveness. In this way, SPRT is understood as a multidimensional approach to evaluating teaching effectiveness. See Table 1 below for a comparison of typical characteristics between SETs and SPRT.

Table 1

Comparing and Contrasting General Characteristics of Student Evaluations of Teaching and Summative Peer Review of Teaching

	Student Evaluations of Teaching (SET)	Summative Peer Review of Teaching (SPRT)
Source(s) of data on teaching effectiveness?	Students enrolled in a course.	Faculty peer reviewers, interpreting classroom teaching and/or related teaching materials.
When is data collected	Within the last weeks of semester.	Dependent on what is being peer-reviewed. If classroom practice, typically not within the first weeks of a class. If prepared materials like syllabi or assignments are being reviewed, material can be collected at any time.
How long does data collection take?	15-20 minutes.	From hours up to days; dependent on scope or purpose of review.
How often is data collected?	With each offering of a course.	Preceding significant milestones aligned with promotion and tenure review processes but collection dependent on institution.

	Student Evaluations of Teaching (SET)	Summative Peer Review of Teaching (SPRT)
What kind of data is collected?	Students' responses to a standardized survey instrument, collecting student opinion typically through Likert-scale responses, often closing with a section for written comments. Primarily quantitative in nature.	Student feedback, faculty peer observation(s), interview(s), teaching materials. A mix of qualitative and quantitative but primarily qualitative information.
Choice in scope of review?	Limited: some SET processes allow an instructor to ask customized questions.	Scope and focus of review led by instructor being reviewed.
Is it a collaborative process between assessors / reviewers and instructor?	No.	Yes.
What is the output?	Quantitative data with descriptive statistics (e.g., measures of frequency, measures of central tendency, measures of variation), individual responses for any text-based questions. Limited to no interpretation of data.	Summative observations and comments submitted by peer reviewers, often focused based on scope and focus of elements selected for review. Significant interpretation of data.
How much data is collected?	Typically, one submission per registered student.	Dependent on scope; volume of data collected not a measure of validity of observations.
How is the data used?	To evaluate instructor performance and inform promotion and tenure process; can inform future classroom practice.	To inform instructor's teaching practice; primarily to inform promotion and tenure process.
Resources required	Can be as simple as paper forms and pencils; many SETs transitioning to electronic collection delivered by third-party companies. Automation throughout the typical processes.	Resources could include preparation time, review resource development and maintenance, peer reviewer time, and peer reviewer training. Comparatively more human resources are required.
What training do assessors/ reviewers receive?	Minimal training is required; training most often focuses on how to use form or tool to submit responses.	Peer reviewers typically receive training before they engage in peer review.

As an approach, SPRT promises to provide a more thorough understanding of the intellectual qualities and preparation necessary for effective teaching to administrators and committees for promotion

and tenure assessments. The overarching purpose is to centralize the importance of teaching within the academic career path (Noben et al., 2020). This is often an aspect of one's academic career that is given less priority than research and publishing, but good teaching is vital for the continuation of the academe (Dell'Alba, 1994; Wood & Su, 2017). In this way, SPRT offers a multifaceted and nuanced approach to evidencing teaching effectiveness. In that spirit of nuance, rather than only examining SPRT from a theoretical lens, we explore how current practices of SPRT function.

Data Collection and Method of Analysis

Our examination began by reviewing all Canadian public universities with a publicly detailed SPRT process and where the results of that process could be used to inform faculty tenure and promotion. From this scoping review, 14 Canadian universities were identified as having such processes. Each of the universities' published SPRT processes was reviewed to determine if SPRT was used for the explicit purposes of tenure and promotion. Consequently, Universities that listed details about SPRT on a website but did not incorporate it into formal tenure and promotion processes were excluded from this study. Based on this exclusion criteria, the University of British Columbia (UBC) and the University of Toronto (U of T) were selected for detailed analysis. When the Canadian universities' SPRT was reviewed, two other universities outside of Canada were mentioned as having served as exemplars from which institutions had drawn inspiration and resources for developing their institutional approach: Texas A&M University (Texas A&M) in the United States of America and the University of New South Wales (UNSW) in Australia.

The four Universities' SPRT processes were then analyzed with the goal of describing similarities and differences amongst the institutions. In addition to the general comparison, the authors also explored how the different SPRT processes address, if at all, SET shortcomings. Analytic induction, which is a qualitative process used primarily in the social sciences, was the first approach taken by the authors. This form of content analysis is useful for analyzing a small number of case studies, and identifying common factors beginning with smaller samples to study the hypothetical explanation for a specific phenomenon (Hammersley, 1989). The phenomenon under study in this paper includes the SPRT processes themselves and how, if at all, they address the commonly cited shortcomings of SETs.

The second complementary approach employed for data analysis was critical textual analysis. Critical textual analysis is a form of "rhetorical critique" and is a method that seeks to "describe the content, structure, and functions of the messages contained in texts" (Frey et al., 1999, p. 231). This method was selected because it allows researchers to speak about and interpret the characteristics of texts. For this study, the published SPRT processes were selected as the primary textual sources to understand better the potential implications of implementing SPRT.

Overview of the SPRT Process

We summarize SPRT processes (see Table 2 below for a comparison of the Universities' SPRT processes), and then assess whether these processes could introduce new sources of concern for institutional reviews of teaching effectiveness.

Table 2

Summary Table of Institutional Summative Peer Review Process' Characteristics

	UBC	U of T	Texas A&M	UNSW
Institutional name of process	Summative peer review of teaching	Peer observation of teaching	Teaching observation	Summative peer review of teaching
Number of peer reviewers	Minimum of two	One	Two	Two

	UBC	U of T	Texas A&M	UNSW
Institutional term used to describe peer reviewers	Peer reviewers, or reviewers	Observers	External reviewers, or peer reviewers	Peer reviewers
Discipline-specific peer reviewers?	Yes, with at least one outside discipline to act as external reviewer	No, but recommended that the observer have a passing familiarity with the specific context.	No	No
Do reviewers have tenure?	At least two	Not explicitly stated, but implied	Not explicitly stated, but implied	Not explicitly stated, but implied
Institutional level that coordinates the process.	Department or school-level	Department and division-level	Department-level	Department-level
Are peer reviewers trained in peer review of teaching before conducting evaluation?	Yes	No	Yes	Yes
Are real or perceived conflicts of interest actively managed?	Need to disclose noted in unit documentation	Not otherwise noted	Not otherwise noted	Yes
If provided, who provides the peer-review training?	Centre for Teaching, Learning, and Technology	n/a	Centre for Teaching Excellence	Portfolio of the Pro-Vice Chancellor, Education & Student Experience
How is a roster of current peer reviewers managed?	Managed by each department or school. Not shared publicly	Managed by each department or school. Not shared publicly	Managed by each department or school. Not shared publicly	Portfolio of the Pro Vice-Chancellor, Education & Student Experience

	UBC	U of T	Texas A&M	UNSW
How often do reviews occur?	At least once every five years; must be conducted during the year prior to the year candidate is up for review	At least two observations per academic cycle for one report	Varies	Varies, can be once every 12 months. Mandated for all promotion and teaching awards
Each peer review process overseen by?	Head of unit	Dean	Dean or chair	Portfolio of the Pro-Vice Chancellor, Education & Student Experience
Do reviewers make recommendation on tenure or promotion?	No	Yes	Yes	Yes
Final report submitted to?	Head, copy shared with candidate	Entire committee	Varies, left up to each individual department	Candidate, the report can then be used by the candidate for tenure and promotion applications
Is process aligned with the relevant collective agreement?	Yes	Yes	Yes	Yes
Are SETs still used?	Yes	Yes	Yes	Yes
Noted limitations	Process excludes review of educational leadership, educational service, and student supervision	Process excludes review of educational leadership, educational service, and student supervision	Process lacks a formal structure	Applicants are required to attend a workshop only offered at certain times of the year before they can undergo review

The University of British Columbia

The University of British Columbia (UBC) first implemented a peer review of teaching in 2009 that was revised in 2013 for distance and blended online courses. UBC's general approach to summative peer review sees it employed concurrently with formative reviews of teaching. For our purposes here,

we focus only on their summative peer review of teaching protocols. They recommend a review team of two reviewers, with one of the reviewers having a familiarity with the subject matter and one of the reviewers having relevant training to conduct SPRTs. The institution's Centre for Teaching and Learning facilitates any training and education associated with peer review. Reviews should occur once every five years for tenured faculty members, with at least two reviews within that same five-year period for non-tenured members and per term appointments. Thus, it is necessary to cultivate several trained reviewers in each Faculty (i.e., different reviewers for the Faculty of Science and Faculty of Arts). It is left up to individual departments if it is feasible to have discipline-specific reviewers. If not, there is a pool of reviewers in each Faculty from which the departments can draw when needed. UBC utilizes trained reviewers (a characteristic shared with UNSW) and reviewers familiar with the subject matter (a characteristic shared with U of T). The belief that underpins this approach is reviewers with a familiarity with the subject matter are better able to make an informed decision regarding knowledge on the part of the reviewee. This approach to the practice of SPRT is contrasted with SETs, where the evaluations draw on students to make judgments about instructor knowledge (Hornstein, 2017; Uttl et al., 2017), allowing for a more-informed evaluation when compared to SETs alone.

The University of Toronto

Structurally, the approach to SPRT taken by the University of Toronto (U of T) is quite similar to that of UBC. An important aspect of U of T's approach to SPRT has to do with power dynamics and the nature of feedback. Often, in academic settings, receiving feedback can be a difficult and, at times, harrowing experience (Sword, 2017). Those undergoing observation for summative reviews are in a vulnerable position – this is true at any stage of their academic career, but particularly those in adjunct, contract, or assistant professor positions where positive reviews are necessary for continued employment or promotion. As a result, the critical feedback must be constructive and presented in a way that will contribute to new pedagogical practices and understandings. U of T goes so far as to strongly suggest that negative judgment has no place in the evaluation process, and care must be taken on the part of the observers/evaluators to abstain from such (Hammersley-Fletcher & Orsmond, 2005). Such feedback should endeavor to be forward-looking; that is, it should improve teaching quality. There should be a challenge or obstacle to teaching identified, from which a collaborative plan can be formulated about how best to approach it. Lastly, there should be a focus on both alternatives and options instead of prescriptions and judgments by reviewers or academic administrators.

The SPRT process at U of T can be summarized in the following steps: it begins with a pre-observation conference, typically comprising the instructor to be observed, the reviewers conducting the observations, and the department or administrative unit representative (i.e., the dean). After this, the classroom observation occurs, which can take place in an in-person or online class, depending on the mode of delivery. During this phase, the instructor's pedagogical approach, as well as their interaction with students, are observed. There is a specific template or rubric that is utilized for this process to (attempt to) standardize results, though narrative notes are also employed. When this has concluded and some time to collate notes and data has elapsed, there is a post-observation conference. This conference allows the reviewers to debrief the reviewee and deliver constructive feedback.

For the most effective results, U of T suggests that multiple peer observations for summative purposes should occur for one report (Chism, 2007). This allows for richer data collection by utilizing the same observational techniques over an extended period, perhaps once per semester for two semesters total. As well, this report should only be one evidentiary source among many for teaching assessment. For best results, it is recommended that timelines for the observation report include both formative and summative assessments. This will allow for a more effective evaluation of the observed teaching and a stronger opportunity for the person under review to benefit from the feedback on their pedagogy. Conducting multiple observations over a period of time sets SPRT apart from SETs where data collection infrequently occurs and only draws on one data point, that of student opinions. In this way, the information gathered through SPRT would be comprised of more data points.

Texas A&M University

Texas A & M University identifies its approach to peer review of teaching to be of a more "holistic" nature. They also make specific mention that in 2010, summative and formative assessments were separated into two processes comprising their own specific protocols and methods for data collection. As mentioned with U of T, summative assessments are the focus here. For this, Texas A&M highlights three basic stages of their peer-review process.

The first stage is the pre-review meeting or dialogue with the entire committee. The person under review lists any relevant details of the course and what they think is essential for the committee to know before conducting an observation. They can also request feedback on specific areas of their teaching, and this is also the stage where the overall expectations from the rest of the committee will be discussed. The second stage is the observation itself. This stage is relatively straightforward: there is an agreed-upon time of review, i.e., a specific class the reviewer(s) will attend, along with any necessary instructional and pedagogic material. The third and final stage consists of the post-review meeting and reflection. This is another meeting of the full committee, where they provide constructive feedback that is also action-oriented to the person under review. This includes specific points and examples to facilitate a helpful learning experience. The person under review will then have the opportunity to reflect on this information and formulate a plan moving forward. Regular meetings of the full committee will, ostensibly, facilitate open and honest dialogue between the members, leading to an effective evaluation of teaching. This is contrasted with SETs, where reviews lack detail and nuance, and do not allow for a kind of collaborative approach, and are only unidirectional. By using these methods, SPRT could contribute to increased teaching effectiveness and for more departmental collaboration vis-à-vis teaching goals as well as open communication.

The University of New South Wales

Of the universities reviewed, the University of New South Wales (UNSW) has one of the most well-developed SPRT processes. UNSW also has the most integrated SPRT protocols by making peer reviews necessary for applications for promotion or awards. UNSW provides information for applicants, information for reviewers, and a reviewer profile page that allows the applicant to learn about some of the potential reviewers they could select. This includes information such as the academic background, inspiration for becoming a reviewer, and FAQs they have for new applicants. The faculty member is responsible for applying for SPRT and demonstrating how they have utilized the six Principles of Quality Teaching in their work. These Principles (detailed on the UNSW teaching and learning website) were implemented in January of 2021 and replaced the previously utilized Nine Dimensions of Teaching. They draw on work by Marsh and Roche (1994), and some examples of teaching Principles include how one engages students in active learning, ways to challenge and support student learning, and demonstrating awareness of student diversity.

The process to apply for summative peer review of teaching can be broken down into the following steps: the applicant begins by registering for an SPRT information session. Next, the applicant plans their timelines. After this, applicants must formally register for a peer review of teaching through the university portal. Two reviewers are then allocated from the available pool of committee members. Importantly, while the applicant cannot select their reviewer, they can identify reviewers they do not wish to be reviewed by, for example, because of a conflict of interest. The application submission is followed by a pre-observation meeting, followed by the actual observation of teaching, and then a post-observation meeting. Here, the applicant responds to the material contained in the report. The data collected informs committees on tenure and promotion about the specific capabilities of the applicants.

Analysis: On the Implementation of An SPRT Process

SPRT is a process that shows promise as an approach to better determine eligibility and excellence for promotion and tenure by drawing on a wide range of metrics and reviewer expertise (Chism, 2007). This contrasts with the current widespread alternative, SETs, which only utilize a metric of student satisfaction under the auspices of teaching effectiveness (Hornstein, 2017). To properly assess the viability of a University implementing a new review process, this section contrasts SETs and SPRTs by focusing on some of the more prevalent criticisms that have been raised against each process.

One of the major concerns identified in the literature relates to the subjective nature of SETs (Hornstein, 2017; Kreitzer & Sweet-Cushman, 2021; Peterson et al., 2019; Uttl et al., 2017). Subjective student evaluations have introduced significant and problematic instances of gender and racial bias against women-identifying/presenting and BIPOC instructors because of either unconscious or overt biases on the part of the students. SPRT strives to be more of an objective process through its implementation of trained reviewers (e.g., UNSW), established protocols (e.g., UBC, U of T, Texas A&M, UNSW), and faculty consultation (e.g.,UBC). One concern that researchers have identified with SETs is that students could be rating a variety of factors unrelated to teaching effectiveness (Hornstein, 2017). During the SPRT processes, the reviewee is evaluated by a group of peers who have similar academic training and classroom experience and are typically informed by guidelines and rubrics (e.g., UBC and U of T). This shift from subjective, student-informed reviews to more objective measures based on multiple points of data collection can help provide a more comprehensive framework to inform promotion and tenure committee members.

In attempting to measure how effective an instructor may have been in the classroom, Kreitzer & Sweet-Cushman (2021) suggest that through SETs, students' comprehension of course material is not evaluated; rather, the student's *feelings* about how much they enjoyed the class are measured. In examining how SPRT functions, the authors found that the issue of failing to evaluate student learning properly is not present to the same degree. This is mainly attributable to trained SPRT reviewers likely surveying multiple data sources and triangulating amongst the evidence to establish a decision of teaching effectiveness. These potential data sources include course syllabi, assessment reviews, classroom observation, and outcome observations (e.g., UBC and Texas A&M). Additional sources of information on teaching practice, such as philosophy of teaching statements, sample assignments, or teaching dossiers, can be examined by review committees (e.g., U of T and UNSW). When compared to SETs, a SPRT process collects more multivarious data for interpretation, with fewer issues of interpretation (Hornstein, 2017).

Multiple studies have concluded that SETs directly contribute to racialized and gendered forms of bias, such as women-identified and BIPOC faculty members at most higher education institutions (Hornstein, 2017; Kreitzer & Sweet-Cushman, 2021; Peterson et al., 2019). This typically translates into fewer opportunities for promotion and tenure, awards, and, in the case of adjuncts, continued teaching appointments. Concerns have been raised in the literature pertaining to unconscious bias (Kreitzer & Sweet-Cushman, 2021), and this no doubt exists among peer reviewers conducting SPRTs; but this is endemic to a larger systemic issue of racism and gender discrimination present in the academe and not directly attributable to SPRT as a process. The issue of bias is seemingly endemic to the SET process (Hornstein, 2017; Kreitzer & Sweet-Cushman, 2021; Peterson et al., 2019; Uttl et al., 2017) and to-date, there has been no evidence published that bias (implicit or otherwise) affects SPRT to the same degree. It is the authors' opinion that, based on the available evidence, the hallmarks of a well-designed SPRT process (e.g., reviews based on multiple sources of data, reviewer training, including implicit bias training) would not allow for the introduction of the same degree of bias. Testing this assumption, however, is an obvious area for future research. For institutions working to implement more inclusive institutional practices, SPRT offers the promise of a more equitable institutional process to evaluate teaching practice than SETs.

Several obstacles to the implementation of SPRT have been identified, including faculty trepidation regarding summative peer review as a tool of managerial oversight or quality assurance, where it is perceived that SPRT could function as a tool of surveillance by university administration to assess teaching quality (Napier et al., 2014; Parsell, 2010). A way of mitigating this concern would be to ensure a transparent process to co-develop SPRT processes and protocols that invite collegial collaboration from the departments (e.g., UBC, U of T, Texas A&M, UNSW). As well, the current standard of SETs is already a measure of managerial oversight. It is uncertain at this point if SPRT would invite less oversight, but at the very least, the data collected is done so in a transparent way between colleagues and from multiple observations without relying on students who are more likely to make uninformed assessments about pedagogy.

Attracting and training assessors for conducting peer reviews has been identified as a threat to the successful implementation of SPRT processes. As detailed in this paper, SPRT comprises a specific and deliberate process. At its best, SPRT requires training for those conducting the reviews, which contrasts

with SETs, where no specific training is necessary for its implementation (Uttl et al., 2017). As a result, SPRT requires a different investment of resources on the part of the university to facilitate the proper orientation and training of assessors. However, this is not an insurmountable issue; for example, UNSW has provided a clear and vibrant framework of what the peer review assessor model could look like. For example, because UNSW has linked SPRT to tenure and promotion—a trend that seems to be increasing based on the literature—the institution has allocated resources to host regular information sessions for faculty interested in requesting a review of their teaching, evidence (in part) of time and resources necessary to train reviewers in the protocols.

A further barrier to the successful implementation of SPRT would be the potential for a conflict of interest between colleagues conducting peer reviews. Many committees and review boards within the academe are made up primarily of faculty members. Members of these committees are expected to interact with their colleagues and peers as a regular part of their official duties. While we are not naïve to the fact that, at times, there are issues on review boards and committees, these tend to be related to an interpersonal issue rather than stemming from the organization itself, i.e., there would be nothing about the SPRT *process* that would cause undue conflicts of interest if there was adherence to established protocols.

An analogous process exists in institutional Research Ethics Boards (REBs), or Institutional Review Boards (IRBs), which are comprised of trained assessors drawn from various faculties and departments to read and render judgments on applications for research involving human participants. In almost every instance, there is a mechanism to declare a potential conflict of interest, either on the part of the applicant or the part of the reviewer, to forestall any issues. There is a similar mechanism in the UNSW SPRT process, where applicants can state if there is any individual reviewer, they would not wish to have on their review committee—a direct example of offering a remedy to the issue.

As it could be deduced from the review of case studies presented, launching, and supporting a SPRT process is a complex undertaking. As detailed, the SPRT processes examined above have a significant number of roles, procedures to follow, and materials to support their use. The development of processes and material to support SPRT will, in many cases, represent new novel ways of working. Overall, SPRT could lead to issues of complexity, including the potential for confusion between the roles and responsibilities in a peer review of the teaching process when used for formative purposes and summative purposes. And aside from the groundswell to address the inequity of SETs use for promotion and tenure, an institution's teaching evaluation inertia could hold back that shift from SETs to a new process. Inertia here means the time, effort, and resources invested in a specific institutional process, all of which could offer barriers to slow change. For example, if a specific evaluation of teaching is named in a faculty union's collective agreement, the terms and conditions related to the evaluation will need revision before a new SPRT process can be adopted. At institutions where SETs are fully automated through IT resources, agreements with third-party companies contracted to deliver this service could lengthen the adoption timeline and adds complexity to the magnitude of change.

Conclusion

Universities have committed to improving the lived experience of all faculty. As institutions reflect and act on their obligations to increase the diversity of the professoriate, they also have an obligation to shift their approach to evaluating teaching effectiveness. Previous research suggests that the "evidence" generated through SETs is problematic and can asymmetrically affect equity-deserving instructors. When compared to these current institutionally supported processes, summative peer review offers a significant improvement in terms of the depth, scope, and quality of its review process to provide faculty members evidence of their teaching effectiveness. Not only does the SPRT process offer a more balanced assessment of teaching effectiveness, drawn from multiple lines of evidence and multiple observers, but when deployed in replacement of SETs to inform promotion and tenure decisions, it would also serve to lessen the bias experienced by racialized and gendered faculty members (Burrell et al., 2020). While there are challenges in making the shift away from student evaluations of teaching, the four institutional practices explored in this paper show how these challenges can be addressed, as well as exemplars of multiple ways forward for other universities interested in navigating this change.

In addition to exploring the presence of bias in SPRTs, an area for future research would be to establish how concretely valid, as a process, SPRTs are for evidencing effective teaching. Such future

research would benefit from incorporating a comparative multiple case study method as a preliminary data-gathering tool, or an institutional analysis to better understand how and why SPRT was selected as a review process for tenure and promotion.

From the instructor's perspective, classroom teaching can be seen as an intensely private act: while students are engaged in learning and are active participants in a classroom, often the classroom is not seen as a space open for public observers. While classroom observation forms a significant portion of all the SPRT processes reviewed, teaching is *more* than the performative act: in taking a more holistic approach to evaluation that SPRT represents teaching materials such as lessons plans, course materials, course outlines, examples of student feedback, and interviews with a reviewee are just *some* of the points of information that would be drawn on when making a judgment. This opening of practice and widening of review material incumbent in formal peer review offers instructors the potential for more reliable and valid judgments of teaching effectiveness. Teaching is a pillar of the university. That institutions can support a process that helps recognize faculty members' teaching effectiveness, be it to support promotion and tenure or external recognition of teaching excellence, also promises to enhance the status of teaching within the academe.

References

Bernstein, D., Burnett, A., Goodburn, A., & Savory, P. (2006). *Making teaching and learning visible: Course portfolios and the peer review of teaching.* Anker Pub. Co. Inc.

Bharuthram, S. (2012). Making a case for the teaching of reading across the curriculum in higher education. *South African Journal of Education*, *32*(2), 205–214. https://doi.org/10.15700/saje.v32n2a557

Blackmore, J. A. (2005). A critical evaluation of peer review via teaching observation within higher education. *The International Journal of Educational Management, 19*(2), 218-232.

Burrell, S. L., Donovan, S. K., & Williams, T. P. (Eds.). (2020). *Breaking down silos for equity, diversity, and Inclusion (Edi): Teaching and collaboration across disciplines*. Rowman & Littlefield.

Canadian Association of University Teachers. (2018, April). *Underrepresented & underpaid: Diversity & equity among Canada's post-secondary education teachers*. https://www.caut.ca/sites/default/files/caut_equity_report_2018-04final.pdf

Chism, N. V. (2007). *Peer review of teaching: A sourcebook* (2nd ed.). Anker Pub. Co. Inc.

Dall'Alba, G. (1994). The role of teaching in higher education: Enabling students to enter a field of study and practice. *Learning and Instruction*, *3*(4), 299–313. https://doi.org/10.1016/0959-4752(93)90021-q

Donnelly, R. (2007). *Perceived impact of peer observation of teaching in higher education*. http://arrow.dit.ie/cgi/viewcontent.cgi?article=1023&context=ltcart

Frey, L., Botan, C., & Kreps, G. (1999). *Investigating communication: An introduction toresearch methods* (2nd ed.). Allyn & Bacon.

Griffen, A. J. (Ed.). (2021). *Challenges to integrating diversity, equity, and inclusion programs in organizations*. IGI Global.

Hammersley-Fletcher, L., & Orsmond, P. (2004). Evaluating our peers: Is peer observation a meaningful process? *Studies in Higher Education*, *29*(4), 489-503.

Hammersley, M. (1989). *The dilemma of qualitative method: Herbert Blumer and the Chicago tradition*. Routledge.

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

Jereb, E., Jerebic, J., & Urh, M. (2018). Revising the importance of factors pertaining to student satisfaction in higher education. *Organizacija*, *51*(4), 271–285. https://doi.org/10.2478/orga-2018-0020

Kreitzer, R. J., & Sweet-Cushman, J. (2021). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics, 20,* 73-84. https://doi.org/10.1007/s10805-021-09400-w

Marsh, H. W., & Roche, L. A. (1994). *The use of students' evaluations of university teaching to improve teaching effectiveness*. Australian Government Publishing Service.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, *75*(1), 138–149. https://doi.org/10.1037/0022-0663.75.1.138

Murray, H. G. (2005). Annual Meeting of the Society for Teaching and Learning in Higher Education. In *Student evaluation of teaching: Has it made a difference?* (pp. 1–15). St Peter's Bay: Society for Teaching and Learning in Higher Education.

Napier, J., Riazi, M., & Jacenyik-Trawoger, C. (2014). Leadership: A cultural perspective on review as quality assurance versus quality enhancement. In J. Sachs & M. Parsell (Eds.), *Peer review of learning and teaching in higher education: International perspectives* (pp. 53-66). Springer.

Noben, I., Deinum, J. F., & Hofman, W. H. (2020). Quality of teaching in higher education: Reviewing teaching behaviour through classroom observations. *International Journal for Academic Development*, 1–14. https://doi.org/10.1080/1360144x.2020.1830776

Parsell, M. (2011, September 22). *The PEER model* [Plenary paper]. International Symposium on Leadership and Communication in Peer Review, Macquarie University.

Peterson, D. A., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLOS ONE*, *14*(5), e0216241.. https://doi.org/10.1371/journal.pone.0216241

Sword. H. (2017). *Air & light & time & space: How successful academics write*. Harvard University Press.

Texas A&M University. (n.d.). *Center for teaching excellence: Home*. https://cte.tamu.edu/

University of Alberta. (n.d.). *Centre for teaching and learning*. https://www.ualberta.ca/centre-for-teaching-and-learning/index.html

University of British Columbia. (n.d.). *Centre for teaching, learning and technology*. http://ctlt.ubc.ca/

University of Toronto. (n.d.). *Centre for teaching support & innovation*. https://teaching.utoronto.ca/

UNSW. (n.d.). *Summative peer review of teaching*. https://www.teaching.unsw.edu.au/summative-peer-review

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

Wood, M., & Su, F. (2017). What makes an excellent lecturer? Academics' perspectives on the discourse of 'teaching excellence' in higher education. *Teaching in Higher Education*, *22*(4), 451–466. https://doi.org/10.1080/13562517.2017.1301911