NPO: 2017/035104/08

Journal of Geography Education in Africa (JoGEA)

Journal of the Southern African Geography Teachers' Association - sagta.org.za

Identifying and Correcting for Bias in Students' Evaluations of Teaching: The Use of Measurable Bench-marking Questions

Jennifer M. Fitchett[1]* and Craig M. Sheridan[2]

[1] Professor of Physical Geography, School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, South Africa.
Jennifer.Fitchett@wits.ac.za
https://orcid.org/0000-0002-0854-1720
*Corresponding author

[2] Claude Leon Foundation Chair in Water Research, Centre in Water Research and Development, School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, South Africa.
Craig.Sheridan@wits.ac.za
https://orcid.org/0000-0002-5913-3428

How to cite this article: Fitchett, J. M. and Sheridan, C. M. (2023). Identifying and Correcting for Bias in Students' Evaluations of Teaching: The Use of Measurable Bench-marking Questions in teacher education, *Journal of Geography Education in Africa* (JoGEA), 6: 32 – 52.
https://doi.org/10.46622/jogea.v6i1.4280

Abstract

Student Evaluations of Teaching (SETs) are widely used to quantitatively assess the competence of university lecturers. SETs can be used formatively to direct lecturers in improving teaching and, more summatively, in determining the suitability of a candidate for employment, confirmation in post or tenure, promotion, and

performance-based salary adjustment or reward. The validity and suitability of SETs remain heavily contested, yet few papers provide workable recommendations to identify evaluations completed in a biased manner, and to improve the validity of the tool or the interpretation. This perpetuates distrust for SETs, and, in cases, lecturers behaving to artificially enhance their scores. We present a method to assess the accuracy of SETs, either informally by the lecturer in understanding the outputs or formally by the administrative body that distributes SETs and their scores. We provide recommendations for identifying biased responses and quantifying average levels of bias using benchmarking questions that can highlight biased responses and adjust output scores and assessment selections accordingly.

Keywords: Lecturer evaluations; Higher education; bias; benchmarking

Introduction

Students Evaluations of Teaching (SETs) refer to survey questionnaires that are administered to students registered for a particular course, usually at a higher education institution, at the undergraduate and postgraduate level, to quantitatively assess the quality of lecturing that the students have received (Simpson & Sigauw, 2000). The responses captured from these questionnaires include, among others, students' experiences of the audibility of the lecturer, the clarity of explanations, the legibility of handwritten notes and ease in accessing typed notes, the structure of the lectures and course, and the availability of the lecturer for consultation or feedback (Hornstein, 2017). They often include more subjective perceptions or opinions of students regarding the subject knowledge of the lecturer and their competence in teaching (Davies et al., 2007; Hornstein, 2017). The latter has been criticised on the grounds that there remains no consensus in the literature as to what constitutes 'effective' teaching, as students do not have qualifications in teaching theory or practice with which to assess this (Davies et al., 2007; Hornstein, 2017; Subbaye & Vithal, 2017). The validity of the instrument would therefore depend on whether institutional and instructional goals are consistent with student perceptions of best practices (Clayson, 2009; Montshiwa & Moroke, 2014). There is extensive critique regarding the interpretation of the student evaluation scores, the

ISSN: 2788-9114

appropriateness of statistical tests, and the understanding of the outputs thereof (Boysen, 2015; Davies et al., 2005).

Despite the ongoing debates regarding the suitability and reliability of SETs and their interpretation (Kogan et al., 2022), they remain extensively used in universities (Dommeyer et al., 2004; Mengel et al., 2019). Formatively, the feedback from SETs can be used in improving course content, lecturing style, and teaching resources (Makondo & Ndebele, 2014; Zabaleta, 2007). Summatively, the quantitative score outputs from SETs are used to assess the quality of teaching for decisions relating to hiring, confirmation (or tenure), promotion, and performance-based pay increments (Hornstein, 2017; Zabaleta, 2007). Given the significance of the consequences deriving from their application, it is important to critically assess and improve the validity both of the instrument and of the interpretation of its output (Hornstein, 2017; Machingambi & Wadesango, 2011). However, output scores are seldom tested or corrected for bias (Mengel et al., 2019; Merritt, 2008), and efforts to improve SETs are often resisted by both university administrations and the academic staff under evaluation (Simpson & Sigauw, 2000).

While there is value in studies that explore the limitations of SETs and that critically evaluate their validity and suitability as tools to assess teaching performance quantitatively, it is important to develop mechanisms to improve the effectiveness of these tools while they remain in use, often in place of more constructive methods such as peer-review (Makondo & Ndebele, 2014), and without triangulation of data from other modes of evaluation of performance and due diligence (Machingambi & Wadesango, 2011). In this study, we present a set of SET questions that can be used as benchmarking tools in identifying bias in student evaluations and disjuncts between university policies and student expectations. We argue that these can be used by an individual lecturer to understand the spread and variance in their SET responses and scores when evaluating their performance, and by administrators in more effectively identifying and correcting for bias.

Ongoing Critiques of SETs

Critiques of SETs as a tool in evaluating lecturer competence and performance can be thematically grouped into issues of the reliability

and validity of the questions posed to a non-specialist target audience and concerns regarding the interpretation of the output data (Dommeyer et al., 2004; Hornstein, 2017; Neath, 1996). In testing the validity of SETs as an instrument, big-data studies with tens of thousands of SET responses for a given course or university have been conducted. These studies evaluated biases in SET scores, the alignment between SET scores and student performance, and explored the impact of the timing and format of SETs on the output scores (see, for example, Boring, 2017; Davies et al., 2005; Kogan et al., 2022; Mengel et al., 2019). The most prominent biases that have been identified relate to the gender of the instructor, and where data are available to test it, the gender of the student (e.g. Boring, 2017; Kogan et al., 2022; Mengel et al., 2019; Wagner et al., 2016). In France, at a university focusing on the social sciences, male students were found to allocate significantly higher overall satisfaction scores to male lecturers than female lecturers, yet male and female students taught by both male and female lecturers perform relatively equally in exams and assessments (Boring, 2017). For the International Institute of Social Studies in the Netherlands, female lecturers were found to be 11% less likely to attain the teaching evaluation cut-off for promotion than men (Wagner et al., 2016). MacNell et al. (2015) provide an arguably more objective test by obscuring the gender of lecturers in an online course, and manipulating the information that students received on the gender of their instructor, and comparing these to SET scores. Gender bias is found to be heightened within the more mathematical or quantitative courses, for minority groups, and for younger and less experienced staff (Boring, 2017; Davies et al., 2005; Kogan et al., 2022; Mengel et al., 2019).

Theories to explain this gender bias often centre around the perceived anticipation among students that female lecturers will occupy a maternal role and provide the type of pastoral support that students received while at primary and secondary school, and their disappointment when this is not forthcoming at university (Bennett, 1982; Neath, 1996; Unger, 1979; Wagner et al., 2016). In an analysis of 20 197 SET responses from Science Po in France, for example, the gender bias was found to be most pronounced for questions that spoke to traditionally male stereotypes, such as the control of the classroom, and less pronounced for questions on the clarity of assessment criteria or level of preparedness for class (Boring,

ISSN: 2788-9114

2017). By contrast, in an analysis of 19 952 SET scores from Maastricht University in the Netherlands for a course group taught by both male and female lecturers, gender biases were detected in both questions specific to an individual lecturer and in those that applied to the course as a whole, such as the evaluation on the assigned textbooks and the efficacy of the online learning platform (Mengel et al., 2019). This was interpreted to be a spillover effect, whereby students anchored their average responses to those for questions relating to the instructor and, in turn, their gender (Mengel et al., 2019). The heightened gender bias for younger and less experienced staff is interpreted to relate to the perceived authority of more senior and older staff (Mengel et al., 2019). The role of racial or class background, or minority status, remains less clear, and is more apparent in some samples and countries than in others (Wagner et al., 2016). In the case of an economics department in South Africa, biases were found to be most extreme where race and gender intersect, although notably, the bias favoured female lecturers (Chisadza et al., 2019). Biases that have been identified in SET scores that do not relate to gender include the year of study of students, with higher scores when teaching at higher levels of study, whether a course is compulsory or an elective, the size of the class, the order of teaching in a team-taught course, and student satisfaction with their grades (Boysen, 2015; Kogan et al., 2022; Neath, 1996).

Further concerns regarding the instrument include the response rates for online SETs relative to those disseminated in person and the distributions of scores obtained from the two methods (Dommeyer et al., 2004). The use of grade incentives to encourage participation has been met with mixed responses, but has not been found to influence average scores. By contrast, the concerns of students regarding their anonymity in online evaluations have proven difficult to resolve (Dommeyer et al., 2004). The scope for biases, particularly through retaliatory or vindictive scoring, has been raised particularly in relation to the timing of SET dissemination before or after a major assessment (Boysen, 2008; Clayson, 2009; Mengel et al., 2019). This can be most effectively addressed by ensuring that SETs are administered and retrieved before students sit the summative examination for the subject, and that SET reports are returned to lecturers once the students' grades have been finalised (Mengel et al., 2019). Student understanding of the purpose and value of the SET has been

highlighted as important in obtaining a more accurate and well-considered collection of SET responses (Kogan et al., 2022; Sojka et al., 2002). Finally, given the importance of SET scores for hiring, confirmation or tenure, promotion, and/or performance-based pay, there are concerns that there are perverse incentives for academic staff to influence SET scores through the timing of evaluations, grading leniently, setting fewer assessments, and administering in-person evaluations when weaker or more dissatisfied students are not in class (Boysen, 2015; Clayson, 2009; Neath, 1996; Simpson & Sigauw, 2000). Indeed Neath (1996) provides a rather satirical list of 20 ways in which lecturers can improve their teaching evaluations, many of which do not involve the improvement of teaching. These practices would reduce the value of SET scores in both formative and summative applications, and are likely to skew the results of both quantitative assessments, such as those using Likert-style questions, and more open-ended qualitative assessments (Steyn et al., 2019).

Critiques regarding the analysis of SET scores relate most prominently to the more reductive use of these scores quantitatively in the summative assessment of lecturer performance (Hornstein, 2017; Steyn et al., 2019). The first relates to the quantitative comparison of averages rather than a more nuanced evaluation of the spread of responses and interpretation of what the data represent (Hornstein, 2017), or indeed an approach using qualitative rather than quantitative data (Steyn et al., 2019). Concerns have been raised that small variations in average teaching evaluation scores are often over-interpreted, particularly by university administrators, and that statistically insignificant differences are treated as being meaningful (Boysen, 2015). The averages themselves as a benchmark are problematic, as staff aiming to exceed the average score through improved teaching would if successful, increase the calculated average, and by its very nature, only half of a staff cohort could have scores above the average (Hornstein, 2017). Critique has also been made regarding the common use of Likert-style questionnaires where categories do not differ in quantity or magnitude but rather in quality, meaning that the interval distance between the categories is undefined and not fixed (Hornstein, 2017). Where numbers are then attached to these, for the purpose of calculating the aforementioned statistics, the quantitative outputs are misleading (Hornstein, 2017).

Despite these numerous concerns,

ISSN: 2788-9114

SETs remain in use, and there have been calls for the triangulation of SET data to validate results and minimise biases (Machingambi & Wadesango, 2011). Our proposed method for identifying and correcting for bias will address some of these issues. Proposals for more nuanced and qualitative assessment of SET scores, and of concurrent peer review are not mutually exclusive (Makondo & Ndebele, 2014; Simpson & Sigauw, 2000), and are welcomed.

Identifying Benchmarking Questions

SET questionnaires typically include a range of questions spanning students' experiences. These experiences concern the clarity of explanations and their perceived improvement in understanding of the content taught to them; the legibility of written notes; the audibility of the lecturer; the quality of online and hard copy resources; the quality of the mode of teaching; the efficacy of the modes of assessment; and, more contentiously, the lecturer's depth of knowledge regarding the content delivered (Davies et al., 2007; Hornstein, 2017).

Usually, a group of compulsory questions are included in assessing any teaching staff member at a particular university. These are used for the comparison between lecturers, courses and disciplines. In many instances, the individual lecturer can also choose from a set of optional questions. We would propose that the questions we identify as valuable benchmarking questions be included in the compulsory sets for any university.

We define benchmarking questions as those that could be quantitatively and objectively or independently measured and therefore provide a measure of the accuracy of students' responses and identify those which are biased. Secondary objective data could be used to quantify a measured value for those questions, or a degree of truth or falsehood, against which both the average (or modal) scores and individual responses can be evaluated for accuracy.

We present these benchmarking questions and methods for handling the inaccurate or unrepresentative responses that they highlight from a theoretical perspective, including the range of aggregated, anonymous responses we have received for each. University permission has been granted for the inclusion of this data. While reading this paper, lecturers can reflect on how their scores for these questions align with their measurable performance.

Questions Assessing the Accuracy of Student Responses

In assessing the accuracy of responses to an objectively measurable variable, the question 'The lecturer is always on time for class' is valuable. For lecturers who are always on time for their classes, or indeed always arrive early to set up and prepare ahead of the class, all students should indicate in their SET the highest ranked classification of 'always' or 'strongly agree', or similar. In this instance, their average score across the responses from the class should therefore be the highest possible category i.e. if measured out of 5, a score of 5. Those students who respond with categories below 'always' or 'strongly agree', indicating that the lecturer is not *always* on time, would represent inaccurate submissions, with the response provided either being completed out of spite — whether through the spillover effect or through deliberately inaccurately scoring that question, or out of ignorance if they are students who seldom attend class. In either case, the responses of those individuals on all other questions in the SET should be drawn into question, either representing a biased response or one of a student who has not attended sufficient lectures to make an informed assessment.

These are, however, very common. In the case of the two authors of this study, despite arriving at all lectures, for all courses, at least 10–15 minutes ahead of the scheduled start of the class, with enough time to set up their audio-visuals, neither have ever received a score of 10/10 for this question. Instead, average scores for this question were consistently below 9.40/10 and, in some cases, lower than 8.

While the assessment of mode rather than mean would generally provide a more representative assessment of a Likert-style question, it does not resolve this issue of bias, as in one instance, despite arriving in advance of every lecture, the modal value on the SET scores for one of the authors of this paper was 'agree' rather than the maximum category of 'strongly agree'. This may, in part, relate to the ambiguity in the framing of the classifications on the Likert scale, using 'agree' and 'strongly agree'. It would not be incorrect to 'agree' that a lecturer is always on time if that were the case, and the differentiation between 'strongly agree' and 'agree' is unclear from an empirical perspective, yet significant in a quantified assessment of those scores. Likewise, a response of 'neutral' may not reflect a student indicating that a lecturer

ISSN: 2788-9114

is only punctual half the time but that they feel neutral about this. A quantitative evaluation of scores would indicate a mediocre level of punctuality. Instead of 'always' and 'never' as the extremes possible for this question, the broader terms 'strongly agree' through to 'strongly disagree' are used – and a student could indeed feel 'neutral' about the member of staff being on time for every lecture, and not perceive this to indicate that they do not do this all the time, but rather that is an expectation rather than something to be commended. Indeed, we agree that all staff should be on time for all classes.

We would argue that this is also a valuable benchmarking question for the less diligent lecturers who are not consistently punctual for lectures, and that this may indeed provide greater detail regarding students' perceptions of the distance between the categories on the Likert scale. A lecturer would be able to evaluate how frequently they are punctual for lectures. If they are never on time, then an accurate representation of student responses would be the lowest category of 'strongly disagree' (that the lecturer is always on time), 'never', or similar. For a lecturer who arrives on time approximately half of the time, the middle-most category (often 'neutral') would be most representative, yet 'neutral' probably does not mean this to a student responding to the SET question and who feels agnostic to their lecturer's punctuality. Similar to the case of a lecturer who is always on time, this question allows for the average and modal scores to be compared to a real-world measure of performance. The difference between the two would represent the degree to which responses are not representative.

The first benchmarking question is useful in determining how an objective, factual measure of a lecturer's diligence can be compared to SET scores. This determines accuracy and the most egregious cases of bias in cases where the SET response directly contradicts reality.

Questions Assessing the Alignment between Student Expectation and University Policy

It is also important to assess the extent to which students' expectations of their lecturers and perceptions of good teaching align with the policies of the university. In many instances, this relates to students' perceptions of the pedagogical approach, the assessments, and the lecturer's knowledge of the subject. These are

difficult to benchmark, both in terms of the measured performance of the lecturer, and the requirements of the university. However, the questions 'work is returned timeously' and 'lecturer is available for consultation' are valuable as benchmarking questions to ascertain whether students have reasonable expectations of their lecturer in the context of university policy.

In the case of the question of work returned timeously, if all marked assignments are returned within the university-stipulated window for grading, all students should provide the highest score of 'always' or 'strongly agree' in the SET. While students appreciate prompt feedback on their assignments, a lecturer should not be penalised for returning work within the window of time stipulated by their university, nor incentivised to grade work more quickly than this as it would compromise both the accuracy of marking and their wellbeing. As in the case of questions on lecturer punctuality, a lecturer would be able to self-evaluate how promptly they have returned work to the class, which can be independently monitored.

Reflecting on our SET scores, our university policy requires marked coursework assessments to be returned to students within two weeks of submission. Across five different SET scores, the average for this question was consistently below 9.40/10 and, in one instance, as low as 7.07/10. This is a departure from a score of 10/10, indicating that students 'strongly agree' that 'work is returned timeously'. In this instance, the departure could represent a bias within the student group and unrealistic expectations of what constitutes a 'timeous' return of work.

This benchmarking question does not only apply to lecturers who return work within the university-stipulated timeframe. Those who return work within the stipulated period 50% of the time should have an average and mode score of 5/10 or in the 'neutral' category, respectively. Those who never return work within the stipulated period should receive an average and modal score within the lowest range and category.

For the question on availability during consultation hours, the same applies. If the member of staff adheres to their university requirements in terms of providing listed consultation hours with sufficient slots to meet demand, and is present in their office (or during COVID-19 lockdown conditions online on a pre-determined platform), student responses to questions on availability should be

ISSN: 2788-9114

framed within these requirements rather than a student's preference for an unattainable continual availability of staff. As for the question on timeous return of work, a lecturer cannot be expected to be available to their student for consultation at all times and require anything but the university-regulated number of consultation hours to obtain better scores, assuming that students are accurately reflecting that they want more engagement and are not just down scoring lecturers as part of the spillover effect (Mengel et al., 2019), presents perverse incentives to cut into time that should be spent on other academic activities (Boring, 2017). Any low-graded responses for the case where a lecturer fulfils their university obligation would suggest that a student is either unaware of the university policies, or does not perceive these to be acceptable. This should not be to the disadvantage of the member of teaching staff who is meeting their contractual obligations, and so, similar to inaccurate responses on arrival time to class, this question could be used in evaluating the level of fairness in received responses and to remove those deemed inaccurate or unfair.

If staff are available when they say they would be, and if this is consistent with university policy, they too should receive the score classifications of 'always' or 'strongly agree'. Again, reflecting on our own scores, this is not the case, with a maximum score of 8.25/10 attained for any of our courses in any given year, despite being available in excess of the university requirement.

The benchmarking can be used for staff who cannot avail themselves as often as required or cannot make it to their scheduled consultation times. As for the benchmarking questions on punctuality for lectures and on return of work, if staff are never available for consultation, a representative student evaluation score would have an average and modal value in the lowest range and category; while those who avail themselves for half of the period of time would have values in the mid-range.

Using the Benchmarking Questions: Calibration of Review Scores and Samples

Having identified these three classes of benchmarking questions, we propose methods for their use in addressing bias in SET responses and aggregated

SET scores.

For all three of these benchmarking questions, lecturers would know whether they meet the criterion in question all of the time, some of the time, or not at all, and could self-evaluate SET responses accordingly in a way which is not possible in the more subjective questions on lecturing quality. More importantly, each could be objectively and independently measured, particularly during remote or blended learning, where meeting slots, lecture upload dates and times, and grade returns are logged digitally. Punctuality to lectures and availability for consultation could be verified through records of the timing of card access to the university and into respective buildings, CCTV camera records, and the login time on teaching laptops. Where lectures and consultations are held online, this is even more simple through the time stamps captured on Teams, Zoom, and other online teaching platforms. Assessments are increasingly marked online, which again would automatically carry these time stamps. Where traditional hard-copy marking is used, dating the mark and/or signature would capture the return date. This information can be used in interpreting the validity of a SET score output on the part of the lecturer and, more comprehensively, in improving the SET score sample group by the university administration (Boysen, 2015). We do not propose a system whereby these time stamps, login dates and card access data are monitored continuously. Instead, where a lecturer's score obtained on a SET is not representative of their performance, they would be able to motivate for an adjustment on the basis of both self-evaluation and more robust empirical evidence.

Where scores differ from performance, there are two avenues for correcting the revealed bias. Both derive from the role of the spillover effect presented by Mengel et al. (2019)

The first involves the removal of individual SET responses that, for these benchmarking questions, demonstrate a deviation in scores from the observed practice of the lecturer. This could occur through three approaches:

1. The lecturer identifies those responses with scores that are two or more categories different from the empirical performance on one or more of the benchmarking questions, removes those from the sample, and re-calculates their average and modal scores from the rest of the sample.
2. The lecturer brings those

ISSN: 2788-9114

responses to the attention of the evaluations office, together with evidence of their performance in the relevant categories, and the evaluations office removes those responses and re-calculates the lecturer's average and modal scores, providing a revised, official report.

3. Where the majority of responses indicate a particular category or value in relation to these benchmarking questions, and a few responses provide markedly different answers, those could be handled by the evaluations office before returning SET evaluation scores to the lecturer. This could be done by automatically removing the responses with outlier answers in these categories or requesting supporting evidence from the lecturer.

This approach is particularly useful in cases where the unrepresentative scores in some SET responses are hypothesised to result from a lack of information rather than an inherent bias. This could be where class attendance is very low or inconsistent or where very few students have attended consultation slots. If a student does not attend lectures or engage with resources sufficiently frequently to accurately answer those benchmarking questions, it can be argued that they would also be less able to provide meaningful responses to the more subjective questions.

The second avenue involves a statistical re-weighting of the SET scores for the lecturer based on the magnitude of the difference between the score that would align with empirical performance, and the score averaged from students' responses. For example, if a lecturer attends every contact session early but receives an average score of 7.8 for that question, a deviation of 2.2 would be calculated. This deviation would represent the discrepancy between the empirical and perceived performance in this category. Using the logic of the spillover effect (Mengel et al., 2019), it could be anticipated that a similar level of deviation would apply to all questions and the overall SET score. The difference could then be scaled to each response, and to the mean score (Equation 1). To strengthen this approach, the deviation for all three questions could be calculated and averaged, and this average deviation applied to the remaining questions and the average score.

Equation 1

$$New\ score = original\ question\ score + \left(\frac{\sum_{i=0}^{n}(Evidence\ based\ score - SET\ score)}{n}\right)$$

where n = number of questions for which evidence can be gathered.

Similar to the approach of removing individual responses, this approach could be applied through three approaches:

1. The lecturer reflects on the teaching term and assigns a score relating to their performance for these three benchmark questions. They then calculate the deviation between their self-assessed score and the score on their evaluation for each question. The lecturer averages these three scores and either adds or subtracts the difference from all scores in their evaluation accordingly.
2. If the lecturer finds a discrepancy between the SET score and their self-evaluated scores, they could bring this to the attention of the evaluations office, together with evidence of their performance in the relevant categories. The evaluations office could then apply the adjustment through the calculated mean deviation and provide a revised, official report.
3. In cases where longitudinal data for a particular lecturer reveal consistent performance in one or more of these three domains and for a particular year, the responses in these categories differ from the long-term mean. The deviation could be handled by the evaluations office before returning SET evaluation scores to the lecturer by requesting empirical evidence from the lecturer, calculating the appropriate score, and adjusting by the deviation.

Score adjustment, rather than the removal of individual responses, is a quicker approach to take and arguably one that is more sensitive in addressing bias rather than a lack of knowledge. It is, however, only feasible where average scores are being used and would not translate well to the use of modal values. We acknowledge the statistical issues relating to the use of averages on Likert-style questions (Hornstein, 2017), but while this remains a widely used practice, this approach may assist in yielding more representative averages.

ISSN: 2788-9114

The three possible routes under each approach represent differing levels of control from the lecturer through to the evaluations office. In instances where the SET is used primarily in informing improvements in teaching and providing insight to the lecturer on avenues to focus on (Makondo & Ndebele, 2014), there should be no problem in a lecturer self-evaluating and adjusting their scores. Indeed, this would reflect the type of critical engagement with SET scores that universities encourage. Where SET scores are used more summatively, as in the case of probation, confirmation and promotion (Hornstein, 2017; Zabaleta, 2007), individual self-assessment and adjustment of scores may be less well received. In these cases, the second or third approach would be more suitable. In cases where new lecturers are receiving their first few reports, unfairly negative and inaccurate responses can be very damaging to the self-esteem and career development of the individual (Hay & Van der Merwe, 2007; Mengel et al., 2019). In these instances, the third approach would be most appropriate, pre-screening the results for bias before the lecturer is presented with them. While the second and third approaches would result in a greater administrative load on the evaluation offices, they would be disseminating more accurate and representative evaluation reports, fulfilling their core responsibilities.

Discussion

The greatest value of SETs, where they form part of a university's mode of lecturer evaluation, is in providing information that lecturers can use in improving their teaching technique and their course offerings (Van der Merwe, 2012; Makondo & Ndebele, 2014; Montshiwa & Moroke, 2014). To do so, they need to be able to discern the valuable SET responses, compiled by students who regularly attend class, have realistic expectations of their learning environment, and who do not harbour ill intent or bias towards their lecturer, such that both positive and negative comments can be understood as realistic and valuable avenues toward the improvement of teaching (Clayson, 2009; Merritt, 2008; Simpson & Sigauw, 2000; Sojka et al., 2002; Steyn et al., 2019). While benchmarking questions that can be used in identifying biased and unrepresentative feedback do not solve all of these problems, for the individual lecturer, these benchmarking questions, and importantly a greater suite of statistics including the mode and the distribution functions of the

scores for those, provide an overview of the percentage of the class scores that can be deemed to be reliable and useful. This is particularly valuable given the impact of SET scores on lecturer confidence, particularly in the early years of their careers (Hay & Van der Merwe, 2007; Mengel et al., 2019).

Importantly, in all three of these questions, it should be noted that a lecturer who is fulfilling their conditions of service should be attending all classes on time, be available during their consultation times, and return work within stipulated timeframes. Elevation of scores to represent this performance would, under ideal circumstances, be applied to all members of lecturing staff. However, this is not always the case. When staff do not always meet these criteria, this could also be monitored and quantified, providing a measure of the inferred distance between the ordinal points on the Likert-style questionnaire (Hornstein, 2017) and a quantification where a more suitable mode and range are included. More importantly, those should be flagged for engagement with the lecturer to improve the degree to which they do fulfil their obligations (Hay & Van der Merwe, 2007; Makondo & Ndebele, 2014). The evaluation of these questions therefore serves a dual benefit, and would provide concrete areas for monitoring and improvement *if needed*. While there is concern about the inclusion of questions in SETs which lecturers know that they achieve as these would artificially inflate scores (Neath, 1996), we would argue that a lecturer deliberately trying to 'play the system' (Hornstein, 2017) by being punctual, returning work on time, and being available for consultation is only to the advantage of the university and to the students.

Moreover, if these are compulsory questions, they would equally up-weight the evaluation scores of all lecturers who are diligent in delivering on their primary duties of service. This does not suggest that these are the most important aspects of teaching, but the dual benefit of encouraging punctuality, timeous return of work, and availability to students would undoubtedly be of value to the university and the students.

It has been suggested that some of the issues inherent in the SET process could be addressed by providing feedback to students, both before and after surveys, regarding their purpose, their implementation and the findings (Montshiwa & Moroke, 2014; McClain et al., 2018). Following our proposed approach, whereby

ISSN: 2788-9114

either the lecturer or the evaluations office either removes unrepresentative evaluations, or re-scores the whole sample, feedback could be given to a class on the objective measure of that question. Communicating to the class on the proportion of students who had answered with responses that were not in alignment with the empirical evidence, the margin of deviation in scores, and an indication that there are correction mechanisms may aid in improving honesty and fairness in future SETs completed by those students (McClain et al., 2018).

Conclusion

Despite the numerous issues with SETs that have been documented in the literature over recent decades, and the empirical evidence for these issues, they remain used in evaluating lecturer performance. The benefits to management in being able to evaluate, quantify and compare lecturer performance are sufficiently large that these will likely remain in use for the foreseeable future. When used appropriately and mindful of inherent biases, these SETs can facilitate a degree of valuable self-reflection and improvement in teaching approach and delivery. It is important, therefore, to identify avenues that could lead to improved accuracy in the tool and its evaluation.

This study highlights two important categories of benchmarking questions that could be used to identify and correct for bias. The first critiques the accuracy of responses in identifying a measurable and quantifiable component of performance and punctuality to lectures. Regardless of whether inaccurate reflection on this question relates to a misunderstanding of the Likert scale, poor attendance in class to evaluate punctuality, or the spillover effect in a disgruntled student, they are valuable in flagging questionnaire responses that are likely to be inaccurate or unfair on the whole. The second category includes questions that would identify whether student expectations are in line with university policy. Again, regardless of the reason for inaccurate or unfair responses, those should not be included in evaluations if the lecturer followed the policy requirements. Removing those questionnaires that are not fair or accurate or re-weighting all scores would thus be a more truthful representation of the performance of the lecturer in staffing-based decisions and would encourage meaningful professional development. While this would not address all forms of bias that come through in SETs, it

makes a contribution to addressing this issue by identifying the spillover effect (Mengel et al., 2019).

The identification of these questions, and the removal of problematic questionnaires or re-weighting of results based on them, do not address all of the issues relating to the use and analysis of SETs and their scores. It does, however, make a contribution to improving accuracy and validity. We would encourage the use of these benchmark questions. We would recommend self-evaluation in correcting for them and possibly for institutional correction (through the administrative office that facilitates SETs) for the benefit of a less biased assessment of teaching performance, especially where job security and salary are implicated. We believe this allows for the most meaningful self-development and improvement of teaching pedagogy and techniques. Concurrent use of a range of alternate approaches, including peer review and ongoing training, is not negated through this. The system of SETs is underpinned by a sense of honesty. Because of this, the system is open to abuse by lecturers and students alike. Unless there is an active filter to remove dishonest questions and answers, this bias will remain.

Author Bios

Jennifer Fitchett is a Professor of Physical Geography in the School of Geography, Archaeology and Environmental Studies at the University of the Witwatersrand. Her research and teaching is situated in the field of biometeorology, exploring climate change and the impacts on plants, animals and people. She currently serves as the President of the Society of South African Geographers (2022-2024).

Craig Sheridan is a Professor in the School of Geography, Archaeology and Environmental Studies at the University of the Witwatersrand. He conducts research into water and waste water treatment. He holds the Claude Leon Foundation Chair in Water Research.

Competing interests

The authors have no competing or conflicting interests, financial or otherwise, to declare..

Funding

This work is not funded by any project grant.

ISSN: 2788-9114

References

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, 74, 170–179. https://psycnet.apa.org/doi/10.1037/0022-0663.74.2.170

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

Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of difference in teaching evaluation means irrespective of statistical information. *Teaching of Psychology*, 42(2), 109–118. https://doi.org/10.1177/0098628315569922

Chisadza, C., Nicholls, N., & Yitbarek, E. (2019). Race and gender biases in student evaluations of teachers. *Economics Letters*, 179, 66–71.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of marketing education*, 31(1), 16–30. https://doi.org/10.1177/0273475308324086

Davies, M., Hirschberg, J., Lye, J., Johnston, C., & McDonald, I. (2007). Systematic influences on teaching evaluations: The case for caution. *Australian Economic Papers*, 46(1), 18-38. https://doi.org/10.1111/j.1467-8454.2007.00303.x

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment and Evaluation in Higher Education*, 29(5), 611–623. https://doi.org/10.1080/02602930410001689171

Hay, H. R., & Van der Merwe, B. C. (2007). The role of student evaluation in improving the quality of teaching and learning practices at the Central University of Technology, Free State: a case study. *South African Journal of Higher Education*, 21(5), 468–487.

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

Kogan, V., Genetin, B., Chen, J., & Kalish, A. (2022). *Students' grade satisfaction influences evaluations of teaching: Evidence from individual-*

level data and experimental intervention. EdWorking Paper 22-513, Annenberg Institute Brown University, Providence. https://doi.org/10.26300/spsf-tc23

Machingambi, S., & Wadesango, N. (2011). University Lecturers' Perceptions of Students Evaluations of their Instructional Practices. *The Anthropologist*, 13(3), 167–174.

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

Makondo, L., & Ndebele, C. (2014). University lecturers' views on student-lecturer evaluations. *The Anthropologist*, 17(2), 377–386. https://doi.org/10.1080/09720073.2014.11891447

McClain, L., Gulbis, A., & Hays, D. (2018). Honesty on student evaluations of teaching: effectiveness, purpose, and timing matter! *Assessment and Evaluation in Higher Education*, 43(3), 369–385. https://doi.org/10.1080/02602938.2017.1350828

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economics Association*, 17(2), 535–566. https://doi.org/10.1093/jeea/jvx057

Merritt, D. J. (2008). Bias, the brain, and student evaluations of teaching. *St Johns Review of Law*, 82(1), 235-288. https://heinonline.org/HOL/P?h=hein.journals/stjohn82&i=239

Montshiwa, V. T., & Moroke, N. T. (2014). Assessment of the reliability and validity of student-lecturer evaluation questionnaire: A case of North West University. *Mediterranean Journal of Social Sciences* 5(14), 352–364.

Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. *Psychological Reports*, 78, 1363–1372. https://doi.org/10.2466/pr0.1996.78.3c.1363

Simpson, P. M., & Sigauw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, 22(3), 199–213. https://doi.org/10.1177/0273475300223004

Sojka, J., Gupta, A. K., & Deeter-Schmelz, D. R. (2002). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching*, 50(2), 44–49. https://doi.org/10.1080/87567550209595873

Steyn, C., Davies, C., & Sambo, A.

ISSN: 2788-9114

(2019). Eliciting student feedback from course development: the application of a qualitative course evaluation tool among business research students. *Assessment and Evaluation in Higher Education*, 44(1), 11–24.

Subbaye, R., & Vithal, R. (2017). Teaching criteria that matter in university academic promotions. *Assessment and Evaluation in Higher Education*, 42(1), 37–60

Unger, R. K. (1979). Sexism in teacher evaluation: the comparability of real life to laboratory analogs. *Academic Psychology Bulletin*, 1, 163–170. https://psycnet.apa.org/record/1981-02140-001

Van der Merwe, D. C. (2012). The usefulness of Student Evaluations for Enhancing the Effectiveness of Teaching Financial Accounting Students at a South African University. *Accounting in Africa*, 12, 107–126.

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

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12(1), 55–76. https://doi.org/10.1080/13562510601102131