Studies in Educational Evaluation 81 (2024) 101342

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Performance evaluation in teaching: Dissecting student evaluations in higher education

Steve Cook [a], Duncan Watson [b,*], Robert Webb [c]

[a] School of Social Sciences, Swansea University, UK
[b] School of Economics, University of East Anglia, UK
[c] School of Management, Stirling University, UK

ARTICLE INFO

Keywords:
Student evaluations of teaching
Pedagogical innovation
Student experience
Teaching quality

ABSTRACT

Numerous studies have highlighted the significant role of Student Evaluations of Teaching (SETs) as a key metric for assessing teaching quality in Higher Education (HE). Building upon these insights, our study introduces an innovative four-tiered model, derived from diverse research, to examine the reliability of SETs. This model addresses biases associated with SETs, delving into both statistical anomalies and cognitive biases, with particular emphasis on often-overlooked hidden context and timing factors. We reveal that these biases can distort SET scores, leading to potentially inaccurate representations of both individual and comparative academic performances. The implications of our research are significant for those influencing HE policy-making and performance evaluation. We echo previous calls for a more expansive approach to teaching effectiveness, essential for genuine insight into teaching quality. By adopting this perspective, HE can design better-informed strategies, ensuring policies and practices reflect the diverse nature of teaching excellence.

1. Introduction

The initiatives implemented by Higher Education (HE) institutions to improve teaching provision have undeniable importance to the sector and stem from not only an intrinsic commitment to the provision of excellent education but also the presence of significant external pressures. For the UK, this external encouragement is exemplified by government educational policy where there have been a series of substantive commissioned reports. These include reports from: Robbins (Committee on Higher Education, 1963); Hale (University Grants Committee, 1964); Dearing (National Committee of Inquiry into Higher Education, 1997); and Browne (Browne, 2010); as well as the Future of Education White Paper (Department for Education and Skills, 2003). These all call for substantive improvements in teaching practices and the elevation of the status of teaching within HE.

The policy implications are clear: these reports have not only identified a pressing need for pedagogical enhancement, but also a demand for greater esteem to be placed upon teaching activities. Prompt action is recommended, with policy measures suggested to incentivise and reward superior teaching, and to better recognise its impact. Such urgency is underlined by the formation of prominent bodies such as the Staff Educational Development Association, the Society for Research in Higher Education, the Quality Assurance Agency, The Higher Education Academy, and Advance HE. These organisations exemplify the drive in the sector to elevate teaching practice and, more broadly, its pivotal role in education policy-making.

The existence of these reports and the establishment of such organisations indicate a clear focus upon teaching enhancement, so it is unsurprising that effort is devoted to the assessment of teaching provision at an institutional level. Here, Student Evaluations of Teaching (SETs) dominate the internal monitoring of teaching effectiveness in HE institutions, as reflected in their description as "*the most common measure used by most universities*" (Gourley and Madonia, 2020, p.75) and "*the primary data used to evaluate teaching effectiveness*" (Peterson et al., 2019, p.1). Similar comments on the dominance of SETs are not difficult to find with, for example, Becker et al. (2012, p.332) noting that "*the evaluation of teaching at almost all schools still relies heavily and almost exclusively on SETs*". Typically, SETs not only assess broader elements such as the overall quality of instruction and the fairness of grading procedures but also delve into lecturer-specific issues. These can include the clarity and organisation of lectures, the instructor's knowledge of the subject, availability for consultation, and the ability to stimulate interest in the course. Therefore while occupying a central role in the development of teaching strategies, SETs also are prominent in the evaluation of faculty

* Corresponding author.
E-mail address: duncan.watson@uea.ac.uk (D. Watson).

https://doi.org/10.1016/j.stueduc.2024.101342
Received 23 August 2023; Received in revised form 23 January 2024; Accepted 8 February 2024
Available online 14 February 2024

and the design of human resource policies.

Viewed as providing a crucial feedback mechanism on academic performance at the level of the individual member of staff, SETs have been noted as central to human resource decisions on annual appraisal, recruitment, promotion and career progression (see, for example, Stroebe, 2020). However, when inspected from a pedagogical standpoint, their reliability as a performance metric is questionable. From the pioneering work of Remmers and Brandenburg (1927) through to more recent studies, SETs have been researched so extensively that they have been referred to as "*the most researched topic in higher education*" (Linse, 2017, p.95). Unfortunately, while voluminous in nature, the resulting SETs literature contains a negative tone, with critics of SETs concluding that they misrepresent teaching quality and are often misapplied and misinterpreted. This calls into question their efficacy as the primary tool for assessing teaching proficiency.

This paper provides a novel means to assess the policy value of SETs as a measure of teaching quality and hence a valid management tool to measure academic performance. The paper is structured as follows. Section 2 synthesises evidence from a breadth of research across disciplines such as education, statistics, and psychology. This analysis forcefully demonstrates that not only can SETs provide fundamentally flawed information, but can also introduce discrimination and incentivise poor practice. Section 3 examines the many complexities of SETs, introducing the Duhem-Quine Thesis to highlight how biases can distort SETs scores, thus impacting their accuracy in measuring teaching quality. Section 4 uses these insights to critique the operational efficiency and hidden costs of SETs in higher education, advocating for a transition to more qualitative, holistic methods of faculty evaluation. Section 5 concludes by summarising the limitations of SETs as a measure of teaching effectiveness. It acknowledges that while SETs data does hold some value, it is not necessarily useful or reliable (as a single source) for evaluating teaching quality, indicating a need to decouple the data from such evaluations and use multiple and varied assessments.

2. The misleading nature of SETs

Overall, if we consider SETs as a method of evaluating staff performance, any interpretation of the past research remains both inconclusive and frustrating. In an extensive meta-analysis of this issue, Uttl et al. (2017, p. 40) conclude, "*.simple scatterplots as well as more sophisticated meta-analysis methods indicate that students do not learn more from professors who receive higher SET ratings"*. Stroebe (2020) agrees with this pessimistic conclusion, stating that "*SETs are unrelated to teaching effectiveness*" (p.283). Furthermore, Galbraith et al. (2012, p. 353) observe that high SETs scores are "*associated with significantly lower levels of student achievement*".

With SETs seemingly unrelated or even negatively related to current achievement, some researchers have examined whether they might be linked to future learning. Unfortunately, findings in this area have also proved disappointing. The relationship between SETs and long-term learning has been questioned by Weinberg et al., (2009), argued to have '*limitations*' by Kornell and Hausman (2016), and in some cases, even found to be negative (see Carrell and West, 2010; Braga et al., 2014). In this context, the use of SETs is often viewed as counterproductive. For instance, SETs might merely demonstrate a teacher's ability to act as a salesperson or to generate shallow '*customer satisfaction*' (Dowell and Neal, 1983; Galbraith et al., 2021).

Given this discouraging landscape, we assess the reliability of the information provided by SETs. Achieving this necessitates a deeper understanding of the sources of bias, which we define as the extent to which a metric deviates from a true representation of teaching quality. Though this seems like a straightforward task, its complexity is increased by the multidisciplinary nature of the analysis. Studies considering psychological or cognitive issues, for example, often focus on cognitive biases, or "*any selective or non-veridical processing of emotion-relevant information*" (Mineka and Tomarken, 1989). In contrast, studies adopting a more mathematical perspective typically consider issues that impact the statistical foundations of SETs. Our overall objective is not to entirely dismiss the role of SETs, but to acknowledge any intricate flaws and, in turn, seek to appropriately recalibrate evaluation methods to account for known complexities.

We divide the properties of cognitive biases into two categories distinguished by their severity and, consequently, the proposed solutions. The first set of biases results in '*noise*'. As an example, we can consider the anchoring effect, where preliminary information subconsciously influences response outcomes. Hitczenko (2013) describes how Likert-scale questioning can lead to '*sequential anchoring*', where responses to one question determine subsequent responses, theoretically diminishing the quality of the information provided. SET analysis also refers to the existence of '*halo effects*', suggesting that all separate dimensions of teaching effectiveness can be swayed by an overarching impression of the teacher (see Pike, 1999; Feeley, 2002; Cannon and Cipriani, 2022). Another form of cognitive bias emerges in the research of Andersen and Hjortskov (2016) that discusses '*dual process theories*'. These theories posit that reported satisfaction is shaped by both intuitive and reflective thought processes. As the former operates more swiftly than the latter, opinions can be influenced by intuitive factors such as mood, with bias arising when responses are provided before (and not allowing for) moderation of more sober reflection.

The second group of biases poses a far greater concern, given they suggest that SET metrics can drastically deviate from objective measures of learning quality. For instance, Merritt (2008, p. 239) explores the influence of a lecturer's non-verbal mannerisms, concluding that "*many of the non-verbal behaviors that influence teaching evaluations are related to race, gender, and other immutable characteristics; they stem from physiology, culture, and habit*". Consequently, SETs might introduce a layer of discriminatory outcomes. For example, studies have indicated that general factors such as instructor fluency can inflate scores independent of learning outcome gains (Carpenter et al. 2016; Toftness et al. 2018; Carpenter et al. 2020a) and these discriminatory outcomes are regrettably apparent in the literature. Gender bias is arguably the most prominent of these, with Boring et al. (2016) finding female instructors receive lower SET scores than their less effective male counterparts, and Peterson et al. (2019) concluding that gender effects influence the mean SET score by 0.5 marks on a 5-point Likert scale. Similar negative impacts are observed for race, with ethnicity and an academic's first language influencing SET scores (Fan et al., 2019; Heffernan, 2002).

In addition to objective demographic factors like age, numerous subjective factors have also been argued to affect SETs. To accurately assess teaching quality, existing literature indicates a broad range of factors need to be taken into account when considering SETs scores: the age, ethnicity, gender, likeability, attractiveness, perceived enthusiasm, engaging manner, tenure status, and organised nature of instructors; perceptions of required workloads for a course; academic discipline; gender of students; level of study; maturity/age of students; the phrasing of SET documents; anticipated and received grades; and the size, elective/mandatory, and quantitative/non-quantitative nature of courses (see, among others, Clayson (1999); Gray and Bergmann, 2003; Pounder, 2007; Felton et al., 2008; Koper et al., 2015; Carpenter et al., 2016; Uttl and Smibert, 2017; Toftness et al., 2018; Fan et al., 2019; Peterson et al., 2019; Carpenter et al. 2020a; Curby et al. 2020; Gurung, 2020; Lamb et al., 2020; Berezvai et al., 2021; Gourley and Madonia, 2021; Heffernan, 2022).

In addition, there exists a strand of literature exploring issues of attendance and engagement (Babcock and Marks, 2011). Perceived problems in these areas have been noted, with studies reporting low attendance levels (Lam et al., 2020; Emahiser et al., 2021), debating the impact of lecture capture (Chai and Guest, 2017; Huyssen 2018; Edwards and Clinton 2019), and examining the effects of a growing number of '*commuter students*' (Thomas and Jones, 2017). This research fundamentally questions students' ability to fully understand the teaching methods being used and evaluate their quality without

succumbing to bounded rationality.

Considering broader statistical issues, there is a notable focus on concerns regarding response rates (Gerbas, 2015; Chapman and Joines, 2017; Young et al., 2019). This concern has prompted research into strategies designed to boost the completion of SETs (see, for instance, Lipsey and Shepperd, 2021). This prompts a particular research perspective: namely, what response rates are necessary to ensure the reliability of the information provided by SETs? A prominent example of research in this field is Nulty (2008, p. 301), which examines SETs response rates capable of delivering "*adequate evidence for accountability and improvement purposes*". The analysis conducted in this research hinges upon varying assumptions regarding sampling error and confidence levels, resulting in the suggestion of required response rates under '*liberal*' and '*stringent*' conditions for different sample sizes. For instance, for a class of 100 students under stringent conditions, an exacting response rate of 87% is suggested to ensure reliability. A more recent approach, offered by He and Freeman (2021), assesses the accuracy or reliability of SETs data via simulation-based analysis. In this context, samples are drawn from artificially generated population data designed to replicate the characteristics of actual SETs series. By comparing the statistical properties of the population and samples, the authors recommend required response rates for various class sizes.

While these studies help shed light on issues related to response rates, they tend to overlook more fundamental statistical problems associated with SETs data. Two issues, in particular, stand out: SETs generate ordinal, rather than interval, data; and SETs are not produced using random sampling. Although these issues are acknowledged in some studies (McCullough and Radson, 2011; Stark and Freishtat, 2014), their implications generally do not receive enough consideration. For instance, consider the ordinal data provided by a 5-point Likert scale from '*Strongly Disagree*' to '*Strongly Agree*' for the statement: "*I would recommend this teacher to other students*". A student choosing a score of 4 instead of 3 is expressing a higher level of satisfaction, as the order is informative. However, since SETs data are ordinal, the computation of a mean value is indisputably invalid. For example, McCullough and Radson (2011, p. 187) state that taking the mean value of ordinal data is "*meaningless*". Similarly, the use of standard errors and subsequent attempts at comparison via '*equality of means*' testing suffer a similar critique. Nonetheless, institutions continue to use mean SETs scores for evaluation purposes, with specific mean values serving as benchmarks for assessment and promotion. Despite the noted issues with the 'mean', we will continue with a discussion of mean SETs scores, given their common use within the sector, and show that deficiencies extend beyond this seldom-recognised issue of calculating a mean value for ordinal data.

Our second concern, random sampling, emphasises the need for more scrutiny of low response rates. SETs do not represent a randomly drawn sample from an underlying population but instead represent an incomplete set of information. As a result, the tabulated response rates presented in Nulty (2008, p.309) are invalidated, a fact acknowledged within the study itself, where it is recognised that the estimates given are "*based on the application of a formula derived from a theory that has random sampling as a basic requirement. With teaching and course evaluations, this requirement is not met*". Therefore, any SETs analysis based on mean values is immediately plagued by non-response bias as well. To illustrate this bias, consider the following simple problem:

$$(1) \quad y = 0.4(4.5) + 0.6x$$

Suppose the value of x is unknown, but you are asked to state the value of y. An accurate response is to acknowledge that, without knowledge of x, the value of y cannot be stated: while $y = -4.2$ if $x = -10$, $y = 13.8$ when $x = 20$, and y will take a range of other values as x is varied further. While this is obvious, such simple arithmetic tends to be forgotten when considering SETs. Suppose y becomes the overall view of students on a module, '0.4' represents the 40% of students who complete SETs and '4.5' represents the average score that they return. Our focus is on 'y' as we wish to know the views of all students on a module, but we do not view this and are instead presented with '4.5' based on the views of a subset of students. Knowledge of the value of y requires receipt of the views of the remaining 60% of students. As the value 'x' is unknown, the SETs score based on the views of all students for this example could be as high as 4.8 or it could be as low as 2.4 assuming a 5-point Likert scale.[1] This is an extraordinarily wide range which could be crudely viewed as quality outcomes running from: 'you need to attend a training session to *receive* support' to 'you should be running training sessions to *provide* support'.

Intuitively, as the only information on the difference between the '40%' and '60%' groupings of students is that, for whatever reason, they occupy opposing extreme positions on the issue of returning scores (i.e. one group does, the other does not), we might assume intuitively that x significantly differs from 4.5. While Stark and Freishtat (2014) raise this issue in a general context, support for this assumption can also be found in the work of Estelami (2015), who observed different SETs scores between early and late responders, and Reisenwitz (2016), who noted underlying attitudinal differences between those who do and do not respond to SETs.

Considering the issues associated with the use of means, some research promotes the consideration of proportions, or distributions, of scores achieved at each level (McCullough and Radson, 2011; Stark and Freishtat, 2014). However, we argue that the use of proportions is challenging given non-response bias and the existence of cognitive biases. This can be exemplified by re-expressing equation (1) as:

$$(2) \quad y = 0.4(4.5 + a) + 0.6x$$

Rather than just the unknown x created by non-response, we now also have the unknown a. Representing the summation of the '*hidden context*' associated with the 40% of students who respond, this can be defined as any factor which divorces a student's subjective score from the '*true*' effectiveness of the instructor. Distributional outcomes, therefore, depend on the extent of this hidden context. For instance, apart from the biases already discussed above related to age, gender, and ethnicity, we can include perceived leniency in grading and grade inflation (Stroebe, 2020; Berezvai et al., 2021). This issue is so significant that Koper et al. (2015) recommend adjusting SET scores based on grades to generate '*real SETs*'.

Importantly, without such adjustments, perverse incentives for instructors can be created. It can be argued that to avoid low SET scores being used when making personnel decisions, an incentive to grade leniently emerges. This is ultimately just one example in the literature of how the hidden context within SETs can foster harmful practices. Such a pessimistic conclusion is supported by Kornell (2020, p.166): "*[SETs] influence teaching. Teachers often try to give the students what they want. The problem is, students do not always want what is best for their learning.*"

In addition to these arguments, consider that if 'x' represents the 'true' score of the non-responding 60% in equations (1) and (2), cognitive biases may not allow this actual value to be reported. Instead, a modified score of '$x + b$' is likely observed in practice. Analogous to how 'a' captures the cognitive bias of responders, 'b' represents a similar bias for non-responders. The previously noted differences between responders and non-responders suggest that $a \neq b$. Consequently, we face a problem: as response rates increase (i.e., non-responders begin to provide scores), the value of 'b' might compound the inaccuracies already introduced by 'a' in the SETs scores.

Specific issues raised in relation to less desirable practices include: a superficial appearance of reduced difficulty (Bjork and Bjork, 2011); the avoidance of more beneficial active learning approaches in favour of didactic methods that produce higher SET scores (Carpenter et al., 2020); the promotion of the "*watering down of content*" (Gray and

[1] Stark and Freishtat (2014) present a similar example of the unknown extremes that can underlie SETs scores based on incomplete returns.

Bergmann, 2003); and the tendency towards spoon-feeding and softening of content and assessment (Simpson and Siguaw, 2000). Complementing this discussion is a literature on the underlying motivations of students. For instance, the National Union of Students (2008) report categorises students into '*toe dippers*', '*next steppers*', '*option openers*', and '*academics*'. While some students are deemed '*outcome maximisers*', Allgood (2001) also refers to '*grade targetters*' who are motivated to achieve a predetermined outcome with the least amount of effort. Clearly, these different student types may return varied scores based solely on their underlying traits or objectives when presented with the same teaching methods and resources. This leads us to conclude that we may well expect the '*grade targetter*', when presented with a didactic, overly-supportive form of delivery and assessment, to return a score of '5' while the '*academic*' seeking an intellectual challenge to prepare for advanced study and future employment returns a score of '3'.

In our discussion, we have seen that even if questioning of the statistical validity of the frequently used mean SET score is overlooked, various underlying biases introduce a disconnect between SETs and teaching effectiveness. Such a disconnect reflects terms presented within the SETs literature such as '*illusions of learning*', '*subjective impressions of learning*', and '*misjudgement of learning*' (see Carpenter et al., 2020). In addition, the collation of findings from a wide-ranging SETs literature make the more extreme experimental results concerning 'fictitious characters' in some studies less surprising.

Within this particular area of research, four key figures are prominent: Dr Myron L Fox, Chris Miller, Kim Phillips and Pat Turner. Rather than being ground-breaking pedagogical researchers deserving of the positive SETs scores they received, they are all non-existent individuals. Myron, for example, appears in a classic experiment by Naftulin et al. (1973). Played by a paid actor, he provides a game theory lecture "*with an excessive use of double-talk, neologisms, non sequiturs, and contradicting statements*" (p.631). Exchanging intellectual content for humorous delivery, Myron was still capable of seducing the professional audience into providing positive reviews. The other characters appear in Uijtdehaage and O'Neal (2015). Despite also being fictious and never delivering a single teaching session, Chris, Kim and Pat still received feedback associated with effective levels of teaching provision. While extreme in nature, these findings can be related to the above review of SETs biases with, for example, issues including likeability, perceived enthusiasm and an engaging manner being obvious potential factors underlying positive reviews observed for Myron.

Rather than just being flawed, we have seen that SETs can be discriminatory across teacher careers and counterproductive for attempts to improve pedagogical practice. However, the extent of their illegitimacy needs further consideration. Given the importance of student feedback and the prominence of SETs, should we just advocate for an outcome where comparisons of mean values are ignored? Or, do we expect such significant levels of hidden context that even using the distribution of SETs outcomes can be counterproductive? We now turn to answering these questions.

3. Decoding SETs

Our critique of SETs as indicators of teaching effectiveness is encapsulated in a four-tiered analysis, each tier uncovering further limitations of SETs as measures of teaching performance:

1. *Objective Assessment of Teaching Effectiveness*: At our foundational level, we postulate the existence of an objective assessment of teaching effectiveness. This presupposes the availability of an unbiased evaluation of an instructor's actual teaching effectiveness. We consider two instructors for our analysis: an '*innovator*' instructor, who is objectively superior in teaching effectiveness compared to their '*satisficer*' counterpart.
2. *Impact of Limited Information*: Progressing to the second tier, we explore how limited information creates a disparity between objective and subjective measures of teaching quality. This gap introduces 'noise' into the findings, with students employing proxy measures for teaching effectiveness. These may include student grades or feedback, which, being imperfect indicators, can lead to a skewed understanding of teaching quality.
3. *Distorted Information on Proxies*: In this stage, we address distortions in information due to student biases, underscoring the significance of the hidden context. These biases can misrepresent the proxies, leading to an inaccurate portrayal of teaching effectiveness. For instance, a teacher's style or recent classroom events could bias student evaluations, potentially misrepresenting their effectiveness.
4. *Response Rate Issues*: Finally, we recognise that response rates for SETs typically fall below 100%, resulting in data from a self-selecting, smaller sample. We accordingly adjust SET outcomes to account for this distortionary effect.

In Tier 3 of our analysis, we delve into how the diverse characteristics of instructors can adversely affect Student Evaluation of Teaching (SETs), signalling a need for more comprehensive discussion. Beyond the context addressed in Section 2, this section introduces the significant role of '*timing effects*'. We explore these effects in the context of the Duhem-Quine Thesis (Duhem, 1906/1954; Quine, 1951; Harding, 1976; Søberg, 2005), which posits that our underlying assumptions—about instructor characteristics, student perceptions, and evaluation criteria—are interconnected. These assumptions, according to the thesis, cannot be individually confirmed or disproved, suggesting a complex interplay that influences SET outcomes. The thesis emphasises that what we measure with SETs is not just teaching effectiveness in isolation, but a complex amalgamation of factors, biases, and assumptions, challenging the use of SETs as a straightforward metric for assessing teaching performance. Therefore, for SET outcomes to robustly test teaching quality, the following auxiliary assumptions must hold:

- *Zero Immediacy Effects Assumption*: This assumes that student evaluations are not swayed by immediate reactions to a teacher's style or personality. However, the reality of student evaluations often contradicts this, as immediate perceptions can significantly influence scores.
- *Zero Short Run Effects Assumption*: This implies evaluations are not affected by recent events, such as grades or specific classroom incidents. However, these short-term factors can have a disproportionate impact on student perceptions.
- *Zero Long Run Effects Assumption*: This suggests students evaluate based on long-term benefits of teaching (e.g. impact of teaching experience for employability outcomes), a perspective that may not be prevalent in all evaluations.

To demonstrate the consequences of overlooking these unrealistic assumptions, Tier 3 of our analysis systematically examines the varied characteristics of our instructors. This exploration is designed to shed light on both the effects of the standard hidden context and the significance of timing effects.

We are now in a position to demonstrate how the characteristics of academics might affect SET outcomes. We consider two academics: Lecturer A, our '*innovator*', who should objectively outperform their counterpart in SET outcomes. She is an ethnic minority woman with a limited research reputation, but a strong commitment to pedagogical investments. Professor B, our '*satisficer*', is a white male who focuses on research and minimally invests in teaching excellence, doing just enough to fulfil his teaching duties. To test whether Lecturer A will outperform Professor B, we require details of our student respondents, allowing us to hypothetically simulate distributions of SET scores over a 5-point scale for a class of 100 students. To simulate the impact of bias, we adopt two distinct methodologies:

- *Distribution Bias Method*: This method mechanically adjusts student responses to address two response issues: neutral response bias and extreme response bias. The former creates a '*conservative*' distribution, where the validity of auxiliary assumptions alters some students' scores by 1 point, e.g., a score of 4 could become 3 or 5. The latter offers a '*polarised*' distribution, reflecting more extreme student reactions, pushing students away from their initial views to the endpoints of the distribution at 1 and 5.
- *Predicted Average Method*: This approach utilises a quasi-natural experiment methodology, informed by the impact of remote teaching during the Covid pandemic, to simulate SET score distributions. It provides a statistically robust insight into how external factors can influence student evaluations, with distributions determined through random number generation using a Poisson distribution.

Our findings, presented in Tables 1 and 2, illustrate how biases and timing effects, stemming from the failure of our stated assumptions, can significantly distort SET outcomes. For Lecturer A, an adept and innovative teacher, the analysis suggests a potential reduction in SET scores due to these biases. Conversely, for Professor B, a research-focused academic with lesser emphasis on teaching, our approaches suggests that SETs scores are inflated above the '*true*' measure.

These methodologies highlight two primary concerns regarding the use of SETs as a measure of teaching quality. Firstly, they challenge the consistency of SETs as a reliable measure, since effective teaching does not necessarily result in higher scores. Secondly, our analysis, segmented by periods, emphasises the statistical asymmetry in SETs, where biases may disproportionately favour certain types of teachers.

Table 1
SETs Distribution for Lecturer A.

Tier 1:	
The teaching on a module is truly excellent. A score of 5 is therefore appropriate.	
Starting Distribution: {0, 0, 0, 0, 100}	
Tier 2:	
Inherently imprecise mapping from the elusive notion of teaching effectiveness to the SETs framework is compounded by the wording of the specific document employed. The impact depends on the extent of more polarised effects.	
Distribution Bias (Conservative)	{0, 0, 0, 40, 60}
Distribution Bias (Polarised)	{10, 0, 0, 0, 90}
Predicted Average	{0, 3, 6, 29, 62}
Tier 3:	
Period 1: The learning resources are different to the norm and some students mistakenly take this as evidence that the lecturer is not well-prepared. Low attendance accentuates these problems, with students failing to see how materials excellently build through the course of the term and their relation to assessment.	
Distribution Bias (Conservative)	{0, 0, 20, 40, 40}
Distribution Bias (Polarised)	{20, 0, 0, 0, 80}
Predicted Average	{1, 3, 14, 32, 50}
Period 2: There is no early assessment for students to check their progress without effort. Formative assessment exercises are provided, but they are underutilised. There are rumours that the Lecturer, compared to other staff, is a harsh marker and that the assessment is overly demanding.	
Distribution Bias (Conservative)	{0, 10, 30, 30, 30}
Distribution Bias (Polarised)	{30, 0, 0, 0, 70}
Predicted Average	{2, 4, 19, 36, 39}
Period 3: The quality of resources and approaches taken by the lecturer is underestimated. In time when their positive impact on employability and/or further study become apparent. However, they are not recognised now.	
Distribution Bias (Conservative)	{10, 10, 30, 30, 20}
Distribution Bias (Polarised)	{40, 0, 0, 0, 60}
Predicted Average	{2, 11, 19, 42, 26}
Hidden Context: The Lecturer is a female member of staff from an ethnic minority. Delivery of the quantitative material occurs in a direct and modest manner without overemphasising their enthusiasm for pedagogical know-how. They are seen as less successful than other staff as they have a shorter track field in publishing academic research. Dual process theory becomes relevant as negative affective factors are unmoderated by reflection and hence influence the returns of some students.	
Distribution Bias (Conservative)	{20, 20, 20, 20, 20}
Distribution Bias (Polarised)	{50, 0, 0, 0, 50}
Predicted Average	{6, 16, 17, 41, 20}
Tier 4:	
The module has a response rate of 50%. This leads the SETs 'average score' range to be between:	
Distribution Bias (Conservative)	{1.8, 4.2}
Distribution Bias (Polarised)	{1, 5}
Predicted Average	{2.7, 4.4}

Table 2
SETs distribution for Professor B.

Tier 1:	
The teaching on a module is quite ordinary, with the Professor focusing his time on academic research. There are areas that can be improved, but there are some good elements. A score of 3 would be appropriate.	
Starting Distribution: {0, 0, 100, 0, 0}	
Tier 2:	
Inherently imprecise mapping from the elusive notion of teaching effectiveness to the framework of a SETs document.	
Distribution Bias (Conservative)	{0, 0, 40, 60, 0}
Distribution Bias (Polarised)	{40, 0, 0, 0, 60}
Predicted Average	{0, 0, 63, 25, 12}
Tier 3:	
Period 1: The resources provided by the lecturer are 'off-the-shelf' materials derived in conjunction with a standard textbook. Familiarity leads to overestimated perceptions of quality. Low attendance makes such familiarity an even more attractive property to the students.	
Distribution Bias (Conservative)	{0, 0, 40, 40, 20}
Distribution Bias (Polarised)	{20, 0, 0, 0, 80}
Predicted Average	{0, 0, 51, 34, 15}
Period 2: Regular short multiple-choices assessments, again provided by the textbook publisher, creates an inflated self-efficacy. Students exaggerate learning gains and the value of the learning experience for a more challenging final examination.	
Distribution Bias (Conservative)	{0, 0, 20, 40, 40}
Distribution Bias (Polarised)	{30, 0, 0, 0, 70}
Predicted Average	{0, 0, 44, 34, 22}
Period 3: The quality of resources and approaches adopted by the Professor, being familiar, are overestimated. Perceptions of a positive impact on employability and/or further study are more easily determined.	
Distribution Bias (Conservative)	{10, 10, 30, 30, 20}
Distribution Bias (Polarised)	{20, 0, 0, 0, 80}
Predicted Average	{0, 0, 36, 35, 29}
Hidden Context: The Professor is a male member of staff who is seen as at the pinnacle of the discipline. While they invest no time into pedagogical research, they foster this impression through extensive (and University-recognised) research contributions. Dual process theory becomes relevant as negative affective factors are unmoderated by reflection and hence influence the returns of some students.	
Distribution Bias (Conservative)	{0, 0, 0, 20, 80}
Distribution Bias (Polarised)	{5, 0, 0, 0, 95}
Predicted Average	{0, 0, 30, 33, 37}
Tier 4:	
The module has a response rate of 50%. Attitudinal differences between responders and non-responders result in the return of SETs by student awarding scores of 4 and 5 only. This leads the SETs 'average score' range to be between:	
Distribution Bias (Conservative)	{4.5, 5}
Distribution Bias (Polarised)	{5}
Predicted Average	{4.7, 4.9}

4. Policy implications

The initial appeal of SETs in HE lies in their operational efficiency and scalability. However, this efficiency is not without hidden costs, such as undermining faculty morale and perpetuating inequities. Our paper, drawing upon the Duhem-Quine Thesis which highlights the complexity of isolating hypothesis in empirical research, suggests that SETs should not be the sole measure in career progression. Consequently, the pivotal question arises: do SETs still hold value when applied appropriately?

The Duhem-Quine Thesis, with its emphasis on the multifaceted nature of empirical research, does not advocate for the elimination of SETs. Instead, it underscores the necessity of recognising multiple factors that influence student responses. Consequently, this thesis suggests the importance of combining SETs with various sources of pedagogical assessment. By integrating SETs into a broader evaluative framework, we can better understand and enhance teaching effectiveness, acknowledging the complexity of factors at play.

This changed outlook then allows for specific recommendations over

the redesign of SETs to prompt more insightful and reflective responses. In particular, rather than directly rating the instructor's perceived teaching quality, the focus should be on guiding students to assess the effectiveness of various teaching methodologies. For example, SETs should include questions that encourage students to critically analyse how certain teaching approaches, like the '*flipped classroom*', aided their understanding of course objectives. A question could be, "*How did the 'flipped classroom' method enhance your comprehension of the key concepts?*". Furthermore, it is important for SETs to ask students to contemplate the synergy of different course elements and how they collectively influenced their learning journey. For example, a question might be, "*How did the combination of readings, multimedia resources, and interactive activities contribute to achieving your learning objectives?*".

Such queries steer students to think about the overall educational experience and the efficacy of teaching strategies, rather than focusing solely on the instructor's performance. This approach aligns with the broader aim of SETs to gather feedback that is instrumental in refining teaching practices and course design, rather than serving as a direct assessment of an instructor's teaching quality.

In summary, our paper advocates for a redefinition of the role and usage of SETs in HE. This redefinition, informed by the insights of the Duhem-Quine Thesis, recognises the limitations of using SETs in isolation and underscores the need for a more balanced and comprehensive approach to evaluating teaching effectiveness. The proposed changes aim not just for minor adjustments but a significant transformation in how SETs are conceptualised and implemented within educational assessment, contributing to a more accurate understanding of teaching effectiveness.

5. Concluding remarks

Our study on SETs in Higher Education underscores their considerable value in understanding specific aspects of the educational process. SETs provide insights into students' perceptions of teaching methods, the impact of timing on learning outcomes, and potential biases, including discriminatory reactions. This information is crucial for refining institutional strategies through a deeper understanding of student feedback. Consequently, our research does not advocate for the complete abandonment of SETs; rather, it highlights their role in capturing a specific dimension of the student learning experience.

Despite the utility of SETs in certain contexts, our study identifies the limitations of using SETs as a sole measure of overall teaching effectiveness. Applying the Duhem-Quine Thesis in our four-tier demonstration, we highlight how various non-teaching factors and inherent biases can significantly skew SETs scores. These distortions can result in evaluations that do not accurately represent an instructor's teaching ability. Our findings emphasise the need for caution in using SETs as the primary criterion in critical academic decisions like faculty promotions, tenure, and remuneration.

In summary, our research underscores the necessity of moving beyond simplistic, quantitative measures like SETs for evaluating teaching effectiveness in Higher Education. The reliance on these cost-effective measures is insufficient and is likely to be misleading. We advocate for a transition towards portfolio-based qualitative assessments, which offer a more thorough and multifaceted view of teaching performance. Embracing this approach promises a more equitable and effective method for assessing teaching quality, essential for maintaining high standards of educational excellence and fairness.

CRediT authorship contribution statement

Watson Duncan: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Webb Robert:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Cook Steve:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

References

Allgood, S. (2001). Grade targets and teaching innovations. *Economics of Education Review, 20*, 485–493.

Andersen, S., & Hjortskov, M. (2016). Cognitive biases in performance evaluations. *Journal of Public Administration Research and Theory, 26*, 647–662.

Babcock, P., & Marks, M. (2011). The falling time cost of college: Evidence from half a century of time use data. *Review of Economics and Statistics, 93*, 468–478.

Becker, W., Bosshardt, W., & Watts, M. (2012). How departments of economics evaluate teaching. *Journal of Economic Education, 43*, 325–333.

Berezvai, Z., Lukáts, G., & Molontay, R. (2021). Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching. *Assessment and Evaluation in Higher Education, 46*, 793–808.

Bjork, E., & Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. Gernsbacher, R. Pew, L. Hough, & J. Pomerantz (Eds.), *in Psychology and The Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 56–64). New York: Worth Publishers.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research* (Vol. 0,(0), 1–11. DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71–88.

Browne, J. (2010). Securing a sustainable future for higher education. *Report of the independent review of higher education funding and student finance*. London: Department for Business, Innovation and Skills.

Cannon, E., & Cipriani, G. (2022). Quantifying halo effects in students' evaluation of teaching. *Assessment and Evaluation in Higher Education, 47*, 1–14.

Carpenter, S., Mickes, L., Rahman, S., & Fernandez, C. (2016). The effect of instructor fluency on students' perceptions of instructors, confidence in learning, and actual learning. *Journal of Experimental Psychology: Applied, 22*, 161–172.

Carpenter, S., Northern, P., Tauber, S., & Toftness, A. (2020a). Effects of lecture fluency and instructor experience on students' judgments of learning, test scores, and evaluations of instructors. *Journal of Experimental Psychology: Applied, 26*, 26–39.

Carpenter, S., Witherby, A., & Tauber, S. (2020). On Students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition, 9*, 137–151.

Carrell, S., & West, J. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*, 409–432.

Chai, A., & Guest, R. (2017). Exploring the links between online lecture recordings, cramming and academic performance. *Australasian Journal of Economics Education, 14*, 1–30.

Chapman, D., & Joines, J. (2017). Strategies for increasing response rates for online end-of-course evaluations. *International Journal of Teaching and Learning in Higher Education, 29*, 47–60.

Clayson, D. (1999). Students' Evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education, 21*, 68–75.

Curby, T., McKnight, P., Alexander, L., & Erchov, S. (2020). Sources of variance in end-of-course student evaluations. *Assessment and Evaluation in Higher Education, 45*, 44–53.

Department for Education and Skills. (2003). *The future of higher education*. London: HMSO.

Dowell, D., & Neal, J. (1983). The validity and accuracy of student ratings of instruction: A reply to Peter A. Cohen. *Journal of Higher Education, 544*, 459–463.

Duhem, P. (1906/1954). *The aim and structure of physical theory*. Wiener, P. trans. Princeton: Princeton University Press.

Emahiser, J., Nguyen, J., Vanier, C., & Sadik, A. (2021). Study of live lecture attendance, student perceptions and expectations. *Medical Science Education, 31*, 697–707.

Estelami, H. (2015). The effects of survey timing on student evaluation of teaching measures obtained using online surveys. *Journal of Marketing Education, 37*, 54–64.

Fan, Y., Shepherd, L., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS ONE, 14*, Article e0209749.

Feeley, T. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education, 51*, 225–236.

Felton, J., Koper, P., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on Ratemyprofessors.com. *Assessment and Evaluation in Higher Education, 33*, 45–61.

Galbraith, C., Merrill, G. B., & Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education, 53*, 353–374.

Gourley, P., & Madonia, G. (2021). The impact of tenure on faculty course evaluations. *Education Economics, 29*, 73–104.

Gray, M., & Bergmann, B. (2003). Student teaching evaluations: Inaccurate, demeaning, misused. *Academe, 89*, 44–46.

Harding, S. (1976). *Can theories be refuted? Essays on the Duhem-Quine Thesis*. Dordrecht-Boston: Reidel.

He, J., & Freeman, L. (2021). Can we trust teaching evaluations when response rates are not high? Implications from a Monte Carlo simulation. *Studies in Higher Education, 46*, 1934–1948.

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment and Evaluation in Higher Education, 47*, 144–154.

Hitczenko, M. (2013). *Modeling anchoring effects in sequential Likert scale questions. Working Papers, No. 13-15*. Boston, MA: Federal Reserve Bank of Boston,.

Koper, P., Felton, J., Sanney, K., & Mitchell, J. (2015). Real GPA and real SET: Two antidotes to greed, sloth and cowardice in the college classroom. *Assessment and Evaluation in Higher Education, 40*, 248–264.
Kornell, N. (2020). Why and how you should read student evaluations of teaching. *Journal of Applied Research in Memory and Cognition, 9*, 165–169.
Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings? *Frontiers in Psychology, 7*, 570. https://doi.org/10.3389/fpsyg.2016.00570
Lamb, S., Chow, C., Lindsley, J., Stevenson, A., Roussel, D., Shaffer, K., & Samuelson, W. (2020). Learning from failure: how eliminating required attendance sparked the beginning of a medical school transformation. *Perspectives on Medical Education, 9*, 314–317.
Linse, A. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation, 54*, 94–106.
Lipsey, N., & Shepperd, J. (2021). Examining strategies to increase student evaluation of teaching completion rates. *Assessment and Evaluation in Higher Education, 46*, 424–437.
McCullough, B., & Radson, D. (2011). Analysing student evaluations of teaching: Comparing means and proportions. *Evaluation and Research in Education, 24*, 183–202.
Merritt, D. (2008). Bias, the brain, and student evaluations of teaching. *St. John's Law Review, 82*, 235–288.
Mineka, S., & Tomarken, A. J. (1989). The role of cognitive biases in the origins and maintenance of fear and anxiety disorders. In T. Archer, & L. Nilsson (Eds.), *Aversion, avoidance, and anxiety: Perspectives on aversively motivated behavior* (pp. 195–221). Hillsdale, NJ: Erlbaum.
Naftulin, D., Ware, J., Jr, & Donnelly, F. (1973). A paradigm of educational seduction. *Journal of Medical Education, 48*, 630–635.
National Committee of Inquiry into Higher Education. (1997). *Higher Education in the Learning Society [The Dearing Report]*. London: HMSO.
National Union of Students (2008). NUS Student Experience Report. Available at: ⟨http://www.nus.org.uk/PageFiles/4017/NUS_StudentExperienceReport.pdf⟩.
Nulty, D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education, 33*, 301–314.
Peterson, D., Biederman, L., Andersen, D., Ditonto, T., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLoS ONE, 145*, Article e0216241. https://doi.org/10.1371/journal.pone.0216241
Pike, G. (1999). The constant error of the Halo Effect in educational outcomes research. *Research in Higher Education, 40*, 61–86.
Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60*(1), 20–43.
Remmers, H., & Brandenburg, G. C. (1927). Experimental data on the purdue rating scale for instruction. *Educational Administration and Supervision, 13*, 519–527.
Simpson, P., & Siguaw, J. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*, 199–213.
Søberg, M. (2005). The Duhem-Quine thesis and experimental economics: A reinterpretation. *Journal of Economic Methodology, 12*, 581–597.
Stark, P., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research* (Vol. 00,, 1–7. DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1.
Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology, 42*, 276–294.
Thomas, L., & Jones, R. (2017). *Student engagement in the context of commuter students*. London: The Student Engagement Partnership.
Toftness, A., Carpenter, S., Geller, J., Lauber, S., Johnson, M., & Armstrong, P. (2018). Instructor fluency leads to higher confidence in learning, but not better learning. *Metacognition Learning, 13*, 1–14.
Uijtdehaage, S., & O'Neal, C. (2015). A curious case of the phantom professor: mindless teaching evaluations by medical students. *Medical Education, 49*, 928–932.
University Grants Committee. (1964). *Report of the Committee on University Teaching Methods [The Hale Report]*. London: H.M.S.O.
Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ, 5*, Article e3299. https://doi.org/10.7717/peerj.3299
Uttl, B., White, C., & Wong Gonzalez, D. (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.
Weinberg, B., Fleisher, B., & Hashimoto, M. (2009). Evaluating teaching in higher education. *Journal of Economic Education, 40*, 227–261.