Bond University

Legal Education Review

Volume 30 Issue 1

2020

Student Evaluations: Pedagogical Tools or Weapons of Choice?

Warwick Fisher
Southern Cross University

John Orr
Southern Cross University

John Page
Southern Cross University

Alessandro Pelizzon
Southern Cross University

Helen Walsh
Southern Cross University

STUDENT EVALUATIONS: PEDAGOGICAL TOOLS, OR WEAPONS OF CHOICE?

WARWICK FISHER, JOHN ORR, JOHN PAGE, ALESSANDRO PELIZZON AND HELEN WALSH[1]

I INTRODUCTION

This paper is an output of a Scholarship of Teaching and Learning reading group ('the Group') at Southern Cross University's School of Law and Justice (the 'SLJ'). Over the past three years, the Group has been meeting monthly to discuss scholarly articles and books relating to teaching and learning. More recently, our reading focus shifted to literature on student feedback surveys, often referred to as 'student evaluations of teaching' (or 'SETs'). It became apparent to us that it was imperative to discuss the use, impact, and implication of SETs within the SLJ, and to articulate how a tension exists, in our teaching practices, between SETs' original purpose as tools to inform pedagogical practice, and their current (mis)use as performance markers. Since much of the existing literature in this field comes from Europe and North America, we considered it necessary to contextualise this literature within an Australian higher education setting, and its increasingly neoliberal culture. Eventually, we resolved to reduce our wide-ranging discussions to written form, based on our interpretations of the literature, our first-hand experiences of SETs at the SLJ, and in some cases, a corporate memory of their introduction and evolution extending back 15 or more years.

SETs have been used for decades in higher education, often with the explicit intention of providing ways to improve pedagogy through feedback directly collected via student responses. In this sense, SETs can be described as pedagogical tools: that is, as formative feedback used to help with teachers' reflections, to evaluate teaching effectiveness, and thus, ultimately, to improve overall teaching and learning. The origin of SETs in Australia has been traced back to 1993, through the delivery of what was then called the 'Course Experience Questionnaire',[2] and SETs today play a more predominant

[1] School of Law and Justice, Southern Cross University, Australia. The authors' names are in alphabetical order.

[2] Simon Barrie, Paul Ginns and Rachel Symons, 'Student Surveys on Teaching and Learning', *University of Sydney* (Report, June 2008)

role in Australia than they do in many other countries, partly due to the 'centrality of student evaluations of teaching to both institutional and national quality assurance strategies' and partly due to 'the shift in the sector towards seeing students as "clients" and "consumers" of higher education "services"'.[3] It appears, however, that their original intent has been radically altered over the years, leading to a host of unintended detrimental consequences, both pedagogically and professionally.

As the starting point of our investigation of the existing literature, we used a 2012 paper by Lyn Alderman, Stephen Towers and Sylvia Bannah,[4] a comprehensive review of the literature of what the authors term 'student feedback systems'. This paper already identified a number of inherent flaws and inconsistencies in the student survey model. These are summarised as follows:

1. A lack of consistency of standards — which extended from the superficial (the nomenclature adopted to describe these surveys), to the profound (the lack of national standards or sector-wide criteria that set minimum benchmarks).[5]
2. The systemic failure to implement survey outcomes to improve the student learning experience, or effect pedagogical reform.[6]
3. The paucity of theory to substantiate the practice of student surveys, in what the authors said was 'a lack of explicit theoretical basis'.[7]
4. The tendency of student evaluations to operate in isolation, as stand-alone benchmarks that should properly be integrated into what the authors described as 'a broader approach to evaluation'.[8]
5. Lastly, that unacknowledged structural biases, such as gender bias, subverted the integrity of the survey tool.[9]

Going forward from 2012 to 2019, recent literature suggests that little has changed — despite the centrality of the topic to teaching and learning. These trends may suggest little appetite for reform — in a

---

<https://www.itl.usyd.edu.au/cms/files/Student_Surveys_on_Teaching_and_Learning.pdf>.

3 Ibid 3.

4 Lyn Alderman, Stephen Towers and Sylvia Bannah, 'Student Feedback Systems in Higher Education: A Focused Literature Review and Environmental Scan' (2012) 18 *Quality in Higher Education* 261.

5 See ibid 27. The authors note 'to a large extent these surveys remain idiosyncratic institutional practices… operating independently of any national system and usually without reference to each other'. Such idiosyncrasy takes form as 'considerable variation in question topics, wording and rating scales and ways the information is gathered, interpreted and acted upon'.

6 See ibid 264. The authors note that there 'is little evidence that study findings are being used to change or improve the student learning experience'.

7 Ibid 270.

8 Ibid.

9 Ibid.

context where the model is considered to generally 'work'. Or, it may suggest that critiques of student evaluation systems, such as they exist, are falling on deaf, or at least uninterested ears.

In 2017, Henry Hornstein for example, described SETs as 'an inadequate assessment tool for evaluating faculty performance'. [10] Hornstein argues that they simply operate as blunt instruments of 'summative evaluation that "sum up" overall performance to decide about promotion, and tenure.' [11] He reiterates that surveys have 'evolved [since the 1970s] into the dominant and in many cases sole indicator of teaching competence.' This has occurred in what he describes as a neoliberal context, where students are encouraged to 'see themselves as customers/consumers of education.'[12]

Hornstein notes that SETs do have value in gauging certain student experiences (such as measuring 'the audibility of the instructor, [the] legibility of instructor notes and availability of the instructor for consultation outside of class'), but otherwise, students lack the ability or experience to validly assess teaching competence. Their use to assess teaching competence beyond such experiential factors renders the student survey, in Hornstein's view, both 'invalid' and 'illegal'.[13]

There is a strong focus in this recent literature as to the profound ways in which the uncritical use of SETs continues to entrench structural disadvantage. MacNeil et al,[14] Boring et al,[15] and Boring,[16] report on gender bias and stereotypes, and the impact student perceptions have on survey results. These papers find that 'student evaluations of teaching are biased against female instructors by an amount that is large and statistically significant' and that these biases are 'stronger than any connection they might have with [teaching] effectiveness.'[17] Their single-minded use for 'promotions of tenure-track academics and contract renewals of adjunct professors' means that 'female professors… spend more effort on time-consuming dimensions of teaching… in an attempt to increase their SET scores [with the] opportunity cost of less time for research, which in turn hinders 'women's chances for promotions.'[18]

---

10 Henry Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool for Evaluation Faculty Performance' (2017) 4(1) *Cogent Education* 1.

11 Ibid 2.

12 Ibid.

13 Ibid 3.

14 Lillian MacNell, Adam Driscoll and Andrea Hunt, 'What's in a Name: Exposing Gender Bias in Student Ratings of Teaching' (2015) 40(4) *Innovative Higher Education* 291.

15 Anne Boring, Kellie Ottoboni and Philip Stark, 'Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness' (2016) (1) *ScienceOpen Research* 1.

16 Anne Boring, 'Gender Biases in Student Evaluations of Teaching' (2017) 145 *Journal of Public Economics* 27.

17 Boring, Ottoboni and Stark, 'Student Evaluations of Teaching' (n 15) 1.

18 Boring, 'Gender Biases in Student Evaluations of Teaching' (n 16) 35.

Structural racial bias may also be entrenched by the uncritical use of SETs. Basow et al[19] find that students pay more attention to the normative white professor than her African American colleague, and student ratings may be 'a more sensitive indication of race and gender biases' than student learning or teaching effectiveness. Collectively, this later literature forcefully argues that the use of student surveys should be treated with caution when it comes to assessing teaching performance, and only acted upon within strict guidelines and limitations.

In 2019, Vicci Lau argued that (law) students see little value in end of semester SETs. This lack of confidence manifests in low response rates and unreliable, untruthful or false answers given to survey questions.[20] Lau attributes this to a lack of student incentive, the perception that the surveyed cohort will not benefit from any reforms that may ensue in following semesters. Students need to see a 'tangible immediacy to the [SET] results,'[21] otherwise the process is 'undermined.'

In 2012, Alderman and her colleagues saw value for both the formative and summative use of student surveys. As to the former, they should be 'diagnostic feedback for academics about the effectiveness of their teaching', a 'component for use in quality assurance processes' and guides to 'students to use in the selection of units of study and teachers'. As to the latter, they provide a 'measure of teaching effectiveness for decisions regarding appointment and promotion'.[22] As SETs continue to evolve, it would seem that little has been done to advance the former. Nor has time stemmed the disproportionate dominance of the latter, as simplistic measures of summative teaching performance that enforce systemic biases and stereotypes.

As Boring et al surmise,

> Universities generally treat SET as if they primarily measure teaching effectiveness or teaching quality… it is not a foregone conclusion that they do. Indeed, the best evidence so far shows that they do not: they have biases that are stronger than any connection they might have with effectiveness.[23]

Recent litigation in Canada involving SETs exemplifies what this literature argues.[24] In April 2018, a dispute between Ryerson University, a public research university in Toronto, and its Faculty Association centred on the 'live issue' that was the University's

---

19 Susan Basow, Stephanie Codos, and Julie Martin, 'The Effects of Professors' Race and Gender on Student Evaluations and Performance' (2013) 47(2) *College Student Journal* 352.

20 Vicci Lau, 'How to Encourage Student Voice: Effective Feedback from Law Students in Course Evaluation' (2019) 29 *Legal Education Review* 1, 2.

21 Lau proposes mid-term assessments as an alternative model, see ibid 4.

22 Alderman, Towers and Bannah, 'Student Feedback Systems' (n 4) 263.

23 Boring, Ottoboni and Stark, 'Student Evaluations of Teaching' (n 15) 3.

24 *Ryerson University v Ryerson Faculty Association*, 2018 CanLII 58446 (ON LA).

reliance on 'Faculty Course Surveys (FCS) for employment related decisions such as promotion and tenure'.[25] The Faculty Association demanded that the University immediately stop its longstanding use of FCS *averages* to evaluate teaching effectiveness. The Association argued these results were 'skewed by bias and their use quite possibly contravened the [Canadian] Human Rights Code'.[26]

The Arbitrator's Award was scathing of the University's narrow use of student evaluation averages to measure teaching effectiveness. While such tests have 'some value', such as 'raising flags' and 'providing data about many things such as the instructor's ability to clearly communicate, missed classes made up, assignments promptly returned, the student's enjoyment and experience of the class, and the difficulty or ease, of overall engagement',[27] that was their effective limit. Significantly, while they were 'easy to administer and have an air of objectivity, appearances are somewhat deceiving'.[28] Relying on expert testimony evidence and available peer-reviewed literature, the Arbitrator concluded:

1. SETs are 'imperfect at best and downright biased and unreliable at worst'.
2. Biases such as 'race, gender, accent, age and "attractiveness"' skew results and 'it is almost impossible to adjust for bias and stereotypes'.
3. There are differences between SETs completed online and in class, and these differences 'need to be understood'.
4. 'The lower the response rate, the less reliable the results'.
5. Results cannot be 'extrapolated and applied to non-responders'.
6. Questions that seek to evaluate a teacher's knowledge and scholarship are 'highly problematic' since it is 'far from clear whether the students have the expertise to comment'.
7. Their timing 'may influence their reliability'.
8. And finally, these issues were non-exhaustive. As the arbitrator pithily noted, 'the list goes on'.[29]

In sum, it was concluded that 'the evidence is clear, cogent and compelling that averages establish nothing relevant or useful about teaching effectiveness. Averages are blunt, easily distorted and inordinately affected by outlier/extreme responses. Quite possibly their very presence results in inappropriate anchoring'.[30]

Notwithstanding the overwhelming evidence against the usefulness of SETs as currently construed, administered, and ultimately conceived, SETs remain a staple of academic life.

---

25 Ibid 2.
26 Ibid.
27 Ibid 4–5.
28 Ibid 5.
29 Ibid 5–7.
30 Ibid.

Emblematic of this apparent paradox is the Ryerson University case. The main argument offered by the University for its ongoing use was one of continuity, based on the suggestion that a 'rapid and radical change' would have been detrimental. However, as the case proved, and as we will argue below, we believe the opposite is the case, and that, instead, the continuation of the status quo in the particular funding regime of the tertiary sector is likely to significantly *decrease* teaching effectiveness. Contrary to what Ryerson University administrators argued, radical and rapid change is, in fact, needed.

II SETs IN THE NEOLIBERAL UNIVERSITY

As Hornstein observes,[31] the literature also concentrates on the contextualisation of SETs within an all-pervasive neoliberal paradigm in universities. For instance, Australian universities are said to suffer from an excessive audit culture:

> Within [such] an audit culture, university staff are to meet output targets and be outcome oriented. There is a demand to constantly 'produce evidence' that one is acting correctly... [and] in such a culture, it is ignored that research and teaching are qualitative and thus cannot be measured easily.[32]

In this output driven culture, measures of student satisfaction and teaching effectiveness derived from SETs become ipso facto important indicators of quality of teaching and courses. While we do not seek to trivialise the importance of measures of student satisfaction to core issues of teaching quality, campus experience, or future employability,[33] SET scores alone are insufficient. The inappropriate interpretation of student satisfaction as a measure of 'customer' satisfaction is a valid concern. The rhetoric of 'students as consumers' is an unfortunate outcome of the so-called New Public Management (NPM) discourse in higher education, 'coupled with a faith in the power of [economic] markets to have their needs met.'[34] In this neoliberal context:

---

31 Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool' (n 10).

32 Megan Kimber and Lisa Ehrich, 'Are Australia's Universities in Deficit? A Tale of Generic Managers, Audit Culture, and Casualisation. (2015) 37(1) *Journal of Higher Education Policy and Management* 83, 88 citing Megan Kimber, 'The Australian Public Service under the Keating Government: Managerialism Versus Democracy' (Unpublished PhD Thesis, University of New England, 2000).

33 See generally 'Overall Student Satisfaction – Victorian University Rankings', *Deakin University* (2018) <https://www.deakin.edu.au/life-at-deakin/why-study-at-deakin/student-satisfaction>.

34 See Michael Wallengren Lynch, 'Teachers' Experiences of Student Feedback: A View from a Department of Social Work in Sweden' (2019) 31(2) *Aotearoa New Zealand Social Work* 58 citing Maria De Lourdes Machado-Taylor, Virgilio Soares and Ulrich Teichler (eds) *Challenges and Options: The Academic Profession in Europe* (Springer, 2017) 236.

> Market-driven rewards cancel out the ethical imagination, social responsibility, and the pedagogical imperative of truth telling in favor of pandering to the predatory instincts of narrow-minded individual awards and satisfactions.[35]

Implied in SETs is that they measure the 'quality' of academic programs and teaching effectiveness. However, as noted, there is a growing body of research that identifies numerous issues of validity and bias within SETs results.[36] The literature paints a picture of a 'perennial debate...concerning the validity'[37] and reliability of student ratings. The concept of reliability refers to the ability to replicate the measure of the student evaluation score if the survey were to be repeated.

What do SETs measure, then? SETs, particularly those conducted online, are treated by universities as qualitative measures of teaching effectiveness — although they are designed as quantitative surveys that capture feedback on students' perception of teaching effectiveness. Student surveys are about the collective views of students regarding their experiences in an academic subject. As such, they are, first and foremost, 'student perception data'.[38] They are not valid quality evaluations and are not measures of student learning,[39] teaching effectiveness or academic quality of subjects. The 'perception of effective teaching' is not a measure of 'effective teaching'. Nonetheless, these student perceptions have become *de facto*,[40] measures of the quality and effectiveness of teaching in the university setting.

Like the arbitrator's findings in the Ryerson case,[41] Stark and Freishtat question the validity of using and comparing *average* scores.

---

35 Henry Giroux, 'Once More, with Conviction: Defending Higher Education as a Public Good' (2011) 20(1) *Qui Parle: Critical Humanities and Social Sciences* 117, 121.

36 See, eg, Alderman, Towers and Bannah, 'Student Feedback Systems' (n 4), Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool' (n 10), Boring, Ottoboni and Stark, 'Student Evaluations of Teaching' (n 15). See also John Lawrence, 'Student Evaluations of Teaching are Not Valid', *American Association of University Professors* (2018) <https://www.aaup.org/article/student-evaluations-teaching-are-not-valid#.Xc-dZHv-vRY>; Philip Stark and Richard Freishtat, 'An Evaluation of Course Evaluations' (2014) *Science Open Research* 1.

37 Angela Linse, 'Interpreting and Using Student Ratings Data: Guidance for Faculty Serving as Administrators and on Evaluation Committees' (2017) 54 *Studies in Educational Evaluation* 94 citing Michael Theall and Jennifer Franklin, 'Creating Responsive Student Ratings Systems to Improve Evaluation Practice' (2000) 83 *New Directions for Teaching and Learning* 95.

38 See Linse, 'Interpreting and Using Student Ratings Data' (n 37).

39 Ibid 95. See also Alex Tabarrok, 'ASA Against Student Evaluations', *Anglophone Economic Leaders Blog* (Blog Post, 10 September 2019) <https://leaders.economicblogs.org/tyler-cowen/2019/tabarrok-asa-student-evaluations/>; Boring, Ottoboni and Stark, 'Student Evaluations of Teaching' (n 15).

40 Stark and Freishtat, 'An Evaluation of Course Evaluations' (n 38).

41 *Ryerson University v Ryerson Faculty Association*, 2018 CanLII 58446 (ON LA).

SET scores are ordinal categorical variables without comparable values. According to the authors, the scores do not represent a categorical value that is comparable between the values given and are meaningless without the distribution of those scores.[42] Furthermore, the score allocated by one student does not necessarily translate to the same value as the same score given by another student and inter-rater reliability has no base measure from which to gain meaning. Since these scores are based on human judgment, which can vary significantly depending on the students' mood or time of day, other measures are required to maintain validity and reliability.[43]

Despite notoriously low response rates for online surveys,[44] their ease of administration and data availability have ensured their continued use in the neoliberal university.[45] However, low response rates may adversely impact SET representativeness,[46] particularly those within increasingly diverse student cohorts.[47] Furthermore, students who respond are not necessarily representative of the whole student cohort, and biases that may have motivated students to respond will not be evident from the SETs results.

Research indicates that there is a positive correlation between the perception of academic performance and response rates. Students who perceive they have earned a higher grade are more likely to complete evaluations online.[48] Inexperienced teachers may fear punishment by students through low SET scores,[49] and thus inflate student grades to

---

42 Stark and Freishtat conclude that 'scatter matters': Stark and Freishtat, 'An Evaluation of Course Evaluations' (n 38).

43 Stark and Freishtat, 'An Evaluation of Course Evaluations' (n 38).

44 Response rates are reported to be around 29 per cent: see Lynch, 'Teachers' Experiences of Student Feedback' (n 35) 60. See also, for example, where response rates are 30 per cent questions arise about the purpose validity and value of student evaluations: Alderman, Towers and Bannah, 'Student Feedback Systems' (n 4).

45 See Heidi Anderson, Jeff Cain and Eleanora Bird, 'Online Student Course Evaluations: Review of Literature and a Pilot Study (2005) 69(1) *American Journal of Pharmaceutical Education* 34.

46 Stark and Freishtat, 'An Evaluation of Course Evaluations' (n 38).

47 MacNell, Driscoll and Hunt, 'Exposing Gender Bias in Student Ratings of Teaching' (n 14); Boring, 'Gender Biases in Student Evaluations of Teaching' (n 16); Boring, Ottoboni and Stark, 'Student Evaluations of Teaching' (n 15). See also Friederike Mengel, Jan Sauermann and Ulf Zölitz, 'Gender Bias in Teaching Evaluations' (2019) 17(2) *Journal of the European Economic Association* 535; Yanan Fan et al, 'Gender and Cultural Bias in Student Evaluations: Why Representation Matters' (2019) 14(2) *PLoS ONE* 2; Karen Kozlowski, 'Culture or Teacher Bias? Racial and Ethnic Variation in Student-Teacher Effort Assessment Match/Mismatch' (2015) 7(1) *Race and Social Problems* 43; Ivo Arnold and Iris Versluis, 'The Influence of Cultural Values and Nationality on Student Evaluation of Teaching' (2019) 98 *International Journal of Educational Research* 13.

48 Anderson, Cain and Bird, 'Online Student Course Evaluations' (n 45).

49 Lynch, 'Teachers' Experiences of Student Feedback' (n 35).

compensate such fears.[50] Therefore, student satisfaction scores alone are insufficient to drive good pedagogy, as Giroux notes:

> Within this framework of simply giving students what they want, the notion of effective teaching as that which challenges common sense assumptions and provokes independent, critical thought in ways that might be unsettling for some students as well as requiring from them hard work and introspection is completely undermined.[51]

Quality and teaching effectiveness are difficult concepts to measure. Nevertheless, university administrators and managers presume these concepts can be adequately measured by SETs, 'yet, the research is abundantly clear in concluding that student evaluations are unreliable indicators of teacher performance.'

> Such an approach does no more than reinforce a neoliberal notion of students as customers paying for a service, turning faculty teaching into a form of entertainment that plays to what Cary Nelson... calls 'the applause meter'.[52]

In summary, drawing from the extensive body of national and international literature discussed above, we identify the following contentious elements as the lens through which to observe — and, if necessary, problematise — the experience of SETs, both in general and specifically in our experience at the SLJ:

1. *Definitory uncertainty*: as Hornstein notes, there is, at present, 'no consensus' among scholars concerning the definition of 'effective teaching' or 'teaching competence'.[53]
2. *Questionable statistical validity of the samples*: the low response rates typical of most SETs (well below 50 per cent of the students being surveyed) are a consequence of a number of issues including 'overall satisfaction with instruction, apathy, absence from class, technical problems, perceived lack of anonymity, [and] lack of importance'.[54] A corollary of this problem is the marked difference between online student cohorts (with far lower numbers of respondents), and on-campus participation (dependent on actual class participation, compulsory or not). Additionally, there is no way to ensure that the participating sample is representative of the whole cohort.
3. *Individual bias*: individual responses are influenced by a host of factors that cannot be accounted for. Students, therefore,

---

50 Prashant Tarun, and Dale Krueger, 'A Perspective on Student Evaluations, Teaching Techniques and Critical Thinking' (2016) 12(2) *Journal of Learning in Higher Education* 3.

51 Giroux, 'Once More, with Conviction' (n 36) 121.

52 Cary Nelson was the president of the American Association of University Professors: Giroux, 'Once More, with Conviction' (n 36) 121.

53 Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool' (n 10) 3.

54 Ibid 4.

are unlikely to be dispassionate evaluators of teachers' performance.

4. *Gender and race bias*: also highly described in the literature.
5. *Competence assessment paradox*: students lack the ability to evaluate the *content* of the unit of study undertaken (otherwise they wouldn't be in a student/teacher relationship).

> Students can reliably speak about their *experience* in a course, including factors that ostensibly affect teaching effectiveness such as audibility of the instructor, legibility of instructor notes, and availability of the instructor for consultation outside of class … they cannot evaluate *outside their experience*, i.e. how can they assess course pedagogy? By what valid criteria are they able to determine how "knowledgeable" an instructor is about his/her subject area?[55]

6. *Inverse relationship between student performance and student satisfaction*: paradoxically, 'the lower the evaluations, the better that student performance tends to be because the instructor has required students to expend significant effort in order to achieve better grades, and students dislike expending effort'.[56]
7. *Pressure to manipulate the scores*: due to the more-than-pedagogical value attributed to SETs (for promotion application and tenure), the 'onus is on the faculty member being evaluated to justify "low scores"' and thus members of faculty will do 'what they can to achieve the highest possible ratings, especially for junior faculty'.[57]
8. *Inherent problem with median scores*: there is an inherent mathematical problem with average satisfaction among teachers as an indicator of where a single teacher is located. It is indeed possible for an entire faculty to achieve excellent results, but, given that the faculty is measured against an overall median score, half of the faculty will always be, by mathematical definition, below that score, and 'when 90% of teachers at a university are rated "excellent", but … 50% are still below the median rating, the consequence is de-motivation and demoralization'.[58]
9. *Students view themselves as customers/consumers of education*: ultimately, the 'average of students' ratings appear objective simply because they are numerical and SETs are a 'measure of popularity… rather than *bona fide* measures of teaching capability'.[59]

---

55 Ibid 3.
56 Ibid 5.
57 Ibid 3–4.
58 Ibid 6.
59 Ibid 4.

III CASE STUDIES AT SOUTHERN CROSS UNIVERSITY

Southern Cross University (SCU) commenced a systematic approach to student feedback in 2004, with the introduction of formal online student evaluation of units. Academic Schools and Colleges had a choice of opting in or out of that process, however, in 2006 the University chose to make online student evaluation mandatory for every coursework unit in every study period. The next and most significant development was in 2009, when student evaluation was expanded to include the collection of feedback on teachers as well as units.[60]

In 2014, an internal Teaching Quality Processes Project (TQPP) reviewed 'all aspects of collecting student feedback, including the instrument, underpinning systems, and processes for review and reflection at a unit level'. While the TQPP working group was abandoned before submitting its final report, it had completed extensive research into staff attitudes to SETs. This 2014 review identified four broad areas of staff concern:

1. The instrument itself: its validity, the topics addressed by the questions, and the optional question bank;
2. The low response rates resulting in invalid data;
3. The quality of student comments, and
4. Disconnected evaluation and review processes.[61]

The review also noted that '[s]tudent feedback data is used as an input to both individual teacher performance and for evaluating a specific unit. These points of action are where, for many staff, issues of data validity became significant. For some, concerns about student feedback representing little more than a popularity contest were reflected in their response.'[62]

Despite strong reservations, SETs continue to be mandatory for all coursework units in every study period. Conducted in the final weeks of the teaching session and sourced from central university systems, SETs provide feedback based on unit location and against individual teaching staff. In addition to forming part of the Course Performance Metrics (discussed in Case Study 1), the results are distributed to the unit assessors, lecturers, tutors, and to School Management (Deans, Directors of Teaching and Learning, and Course Coordinators) and are used to evaluate academic performance as part of the performance review process. More recently, SETs are used to identify underperforming units in terms of success rates and/or unit

---

60 'Staff FAQs about Unit Feedback', *Southern Cross University* (2020) <https://www.scu.edu.au/staff/planning-quality-and-review/student-feedback/unit-feedback-survey/staff-faqs-about-unit-feedback/>.

61 Southern Cross University Office of Planning Quality and Review, Teaching Quality Processes Project Report (2014) 22.

62 Ibid 24.

satisfaction as part of the University's internal quality review process.[63]

In addition to SETS administered by SCU, students and graduates are asked to complete a number of national surveys administered by the Quality Indicators for Learning and Teaching (QILT).[64] Since 2015, QILT has administered the Student Experience Survey (SES), its primary purpose is the collection of data used to measure the quality of teaching and learning and the support provided to students. All students at participating universities are invited to complete the SES, which normally is available for one month around the middle of the year. The SES measures six aspects of the student experience: teaching quality, learner engagement, learning resources, student support, skills development and overall quality of the education experience. The results from the SES are made available to participating universities and published on the QILT website,[65] where users may search by institution or study area to help inform their study plans.

QILT also administers the Australian Graduate Survey, which comprises the Course Experience Questionnaire (CEQ). The CEQ is sent to all graduates of Australian universities four months after graduation. Designed to measure the overall level of satisfaction with their course, the questionnaire uses a Likert scale,[66] to establish graduates' perceptions on teaching quality, goals, assessment, workload and generic skills.[67] The results of the CEQ are also made available to participating universities and to the public on the QILT website enabling institutional comparison of courses and teaching quality.[68]

---

63 The Internal Quality Indicators in learning and Teaching (iQILT) Process reviews all AQF level 7 units as part of an 'accountability cycle where action is taken on student feedback received' (see Part B – Unit Monitoring and Review –iQILT Process – SCU Course and Unit Accreditation Policy).

64 Quality Indicators for Learning and Teaching (QILT) is funded by the Australian Government Department of Education and its website is maintained by the Social Research Centre. QILT is a 'suite of government endorsed surveys for higher education' designed to increase accountability and quality control in the higher education sector. See 'Quality Indicators for Teaching and Learning' (2020) <https://www.qilt.edu.au/>.

65 'Quality Indicators for Teaching and Learning' (2020) <https://www.qilt.edu.au/>.

66 Likert scales are commonly used to allow a respondent to express how much or little they agree with a statement. A typical Likert scale has a five (or seven) point scale e.g. strongly disagree, disagree, neutral, agree, strongly agree.

67 Beatrice Tucker et al, 'Online Student Evaluations Improves Course Experience Questionnaire Results in a Physiotherapy Program' (2008) 27(3) *Higher Education Research and Development* 281; Alderman, Towers and Bannah, 'Student Feedback Systems' (n 4).

68 'QILT Surveys' (2020) <https://www.qilt.edu.au/qilt-surveys>.

A Case Study #1: Australian Graduate Survey & Course Experience Questionnaire Student Experience Survey

The first case study draws on data from the SCU Course Performance Metrics (CPM), which is a summary of course statistics, [69] (encompassing demand, enrolments, EFTSL, student profile, student success and completion rates) and the results of SETs, the First Year Experience Survey [70], and the national surveys administered by QILT. The main purpose of the CPM is to monitor and review courses against national standards through a quality assurance process.

Using the CPM data from 2012 to 2018, the focus of this case study is the SLJ Bachelor of Laws (LLB). This course is delivered through two law degrees (a four-year undergraduate LLB and a three-year graduate LLB). The two degrees are selected for this case study as they provide an opportunity to compare the responses to the various surveys of two different cohorts exposed to the same program. That is, there is no difference between the mode of delivery, teaching methods, syllabus, level of support and assessment in each program. The only difference is the number of units studied to fulfil the requirements of the award. Students complete 32 units of study to fulfil the requirements of the undergraduate LLB and 24 units of study to complete the graduate LLB. Core units remain constant and students in these programs follow the same progression for the first two years. The difference is manifested in the number of electives, which are generally studied later.

In regard to the two cohorts, the CPM indicates that, in terms of demographic variables, the majority of students in both cohorts are female, aged in their thirties, and studying online. Furthermore, the success rates are only marginally different between the two cohorts (Table 1).

Table 1
Demographics of undergraduate and postgraduate students

	LLB Undergraduate	LLB Graduate
Study mode	74.3% online	81.8% online
Gender	75% female	63.5% female
Age	31 years (median age)	37 years (median age)
First in family	63%	49%
Success rate	76.5%	79.5%

Note - Data in table 1 is the average for the period 2012 to 2018.

[69] Here ‘Course’ refers to the degree or program that comprises Units of study (the individual subjects that make up a Course).

[70] The data from the First Year Experience Survey has not been used as students in the graduate program do not complete the survey.

In order to compare the responses of the two cohorts, this case study looks at two broad indicators found in all surveys — overall satisfaction (with the unit/course) and overall satisfaction with teaching. Figures 1 and 2 compare the undergraduate LLB response with the graduate LLB response for these two broad indicators in the SCU SETs,[71] for the period 2015 to 2018.[72]

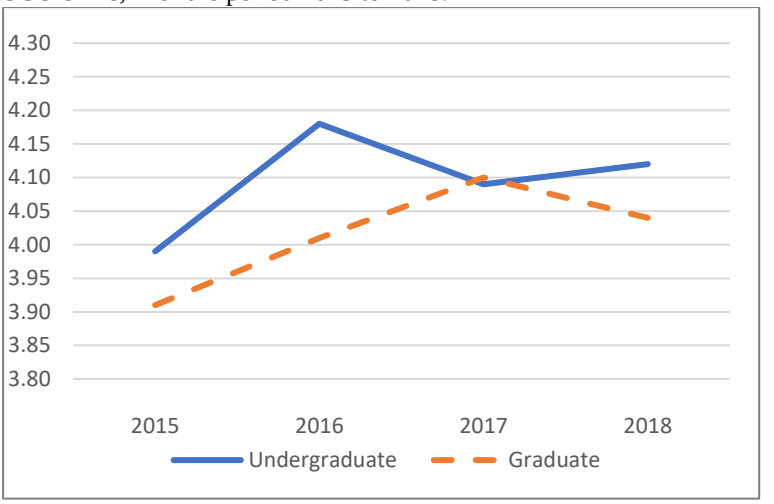

Figure 1
Individual Unit Survey - Overall Satisfaction with Units/Course

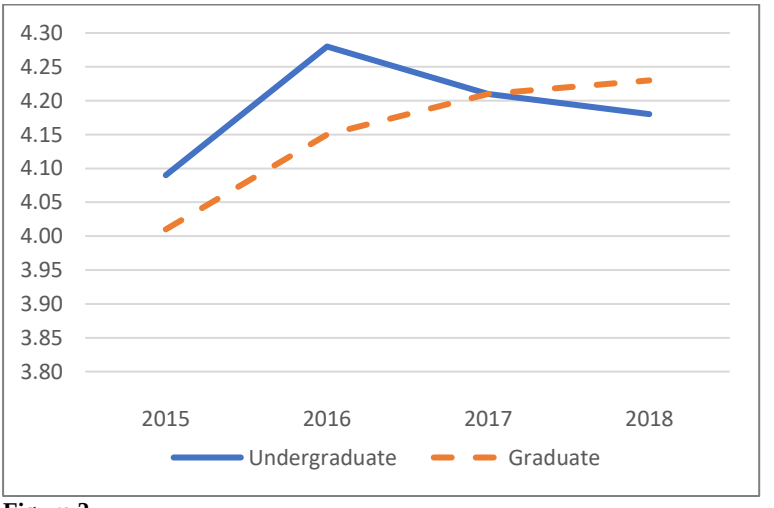

Figure 2
Individual Unit Survey - Overall Satisfaction with Teaching

---

[71] SCU SETs have standard sets of core questions relating to the unit of study and the teaching. This case study looks at the final statement in each set of core questions: 'Overall, I am satisfied with this unit' and 'Overall, I am satisfied with the teaching in this unit'. Students are asked to respond on a five-point Likert scale (Strongly Disagree, Disagree, Average, Agree and Strongly Agree).

[72] Individual Unit Survey data was not available at a course level pre-2015.

The feedback indicates that over a period of four years, undergraduate LLB students generally rated their overall satisfaction with the unit of study higher than students admitted to the graduate entry LLB. The same undergraduate cohort also rated their satisfaction with the quality of teaching higher or the same in three of the four years, and only slightly lower than the graduate cohort in 2018.

Figures 3 and 4 compare the responses of undergraduate LLB students with those of graduate LLB students to the SES for the two broad indicators, overall satisfaction with the course and overall satisfaction with teaching,[73] for the period 2014 to 2018.[74]

Unlike the SCU SET results, the results of the Student Experience Questionnaire (SEQ) indicate that, for a five-year period (2014–2018), graduate LLB students are generally more satisfied with both their educational experience in their course and the quality of the teaching.

Finally, graphs 5 and 6 compare the undergraduate LLB student responses with those of the graduate LLB students for the same two broad indicators in the CEQ,[75] for the period 2012 to 2018.

---

[73] The Student Experience Questionnaire (SEQ) which collects the data for the SES has 46 questions relating to five areas of the higher education experience (teaching quality, learner experience, student support, learning resources and skills development). This study focuses on just two questions contained in the Teaching Quality domain: 'Thinking about your <course> overall how would you rate the quality of your entire educational experience this year?' and 'thinking about this year, overall at <institution> how would you rate the quality of the teaching you have experienced in your course?'. Students are asked to respond on a four point Likert Scale (Poor, Fair, Good, Excellent).

[74] Survey data for 2012 was not available due to insufficient responses.

[75] The CEQ surveys recent graduates on three areas: overall satisfaction with their course, experience with good teaching and improved generic skills. This case study focuses on overall satisfaction with the course and the good teaching indicator, which is based on the average of graduates' responses to six statements relating to teaching practices. The six statements can be found at: 'Graduate Satisfaction', *Quality Indicators for Learning and Teaching* (2019) <https://www.qilt.edu.au/qilt-surveys/graduate-satisfaction>.

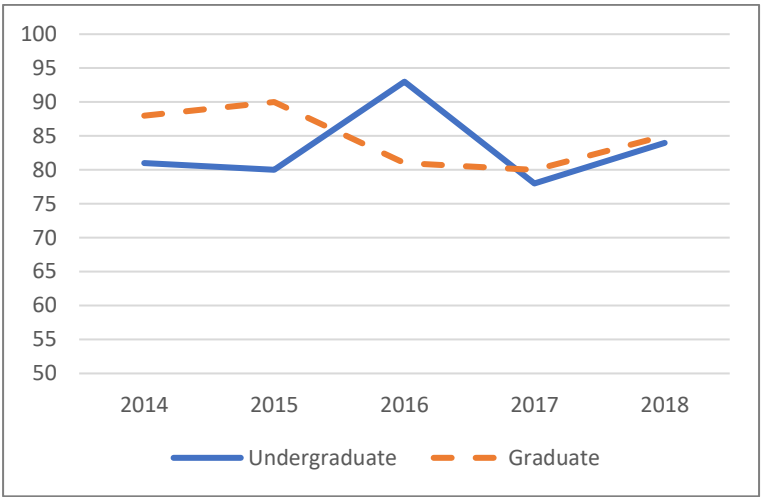

Figure 3
SES - Overall Satisfaction with Educational Experience

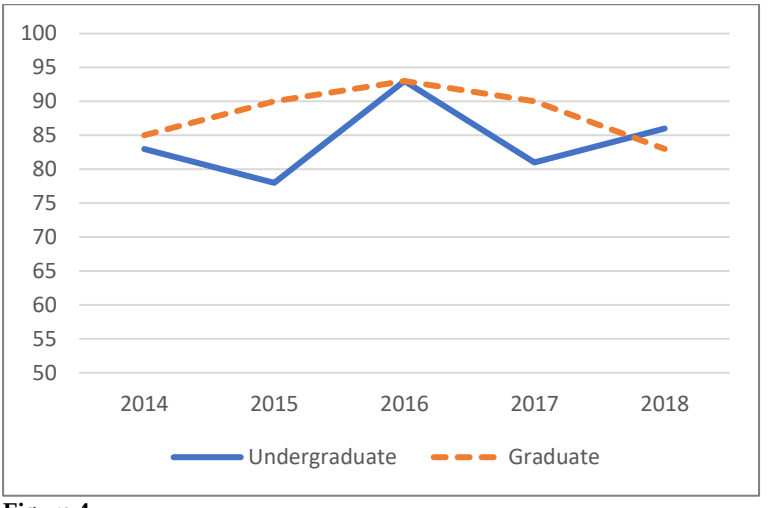

Figure 4
SES - Overall Satisfaction with Teaching

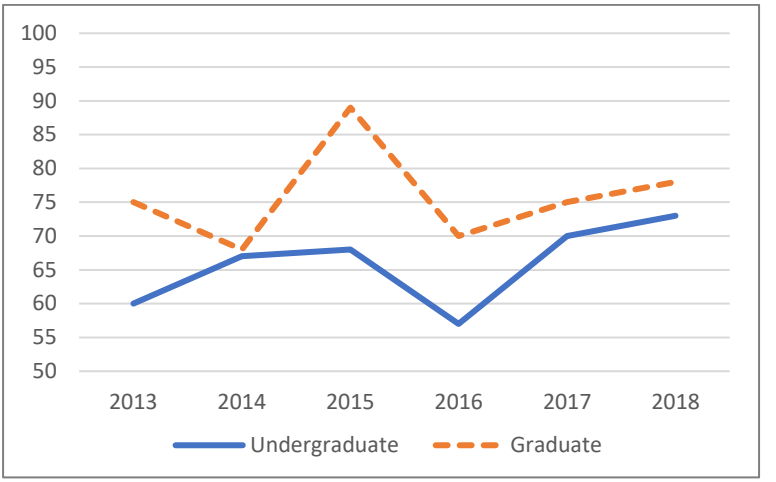

Figure 5
CEQ - Overall Satisfaction with Teaching

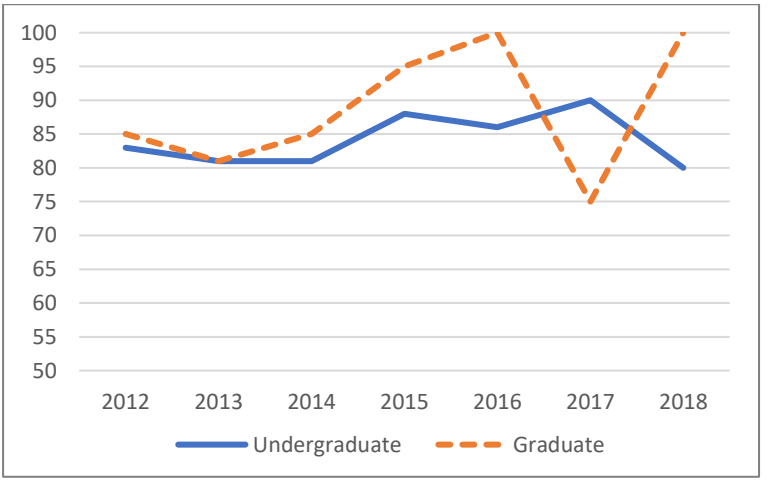

Figure 6
CEQ - Overall Satisfaction with Course

Unlike SET results, the CEQ results show that over a period of seven years, students admitted to the graduate LLB generally (and rather consistently) rated their experience with good teaching practices higher than students admitted to the undergraduate program. Apart from one outlier in 2017, the same graduate cohort was also generally more satisfied with the quality of the course over the seven-year period.

The fact that the different cohorts displayed different responses to identical circumstances in terms of both teaching and unit/course quality suggests the responses are a reflection of some characteristic of the cohort, rather than the quality of the unit or teaching. One possibility is that undergraduate students rate both teaching and units higher than their graduate counterparts when they are engaged in their study and preparing for exams (when SETs are administered), while

there is a tendency for lower ratings both in terms of teaching and the course as a whole during session break (when SES is administered) and after graduation (when CEQ is deployed).

Research conducted elsewhere suggests there is no difference between surveys conducted at the beginning and surveys completed at the end of a study period.[76] However, in the Ryerson University case, it was accepted that 'the timing of the administration of the SET may influence its reliability'.[77] Lau's 2019 study of mid-term assessments at Hong Kong University also agrees that survey timing has implications.[78] Our case study shows that both cohorts' perception of the quality of teaching and the course changes from the time when they are engaged in the unit to four months after graduation. One explanation why graduate LLB students rate teaching higher in the SES (which is normally completed after grades from the first study period have been published) could be that graduate students generally achieve higher grades, which could conceivably influence their ratings. A further explanation may be found in the graduate outcome data. A higher proportion of students graduating from the graduate entry LLB are in full time employment at the time of completing the CEQ. Consequently, the improved ratings (and more positive comments) in the CEQ data may reflect the graduates' circumstances and job satisfaction, rather than the teaching and course quality.

The difference between graduate and undergraduate responses may also be attributed to educational experiences, expectations and how these influence the perception of quality. For graduate students, who are completing a subsequent bachelor degree (normally with a view to changing careers), a high quality course may involve flexible delivery modes, with little or no expectation for interaction and the ability to fast track their degree. Undergraduate students, however, experiencing higher education for the first time, may instead rate the quality of the degree and the teaching on the basis of the educational experience and development of knowledge and skills that will improve grades and lead to employment upon graduation.

The conclusions drawn from this case study suggest that the administration of SETs — including the time they are deployed,

---

[76] Stephen Benton and William Cashin, 'Student Ratings of Teaching: A Summary of the Research and Literature', (IDEA Paper No 50, 2011) citing Larry Braskamp and John Ory, *Assessing Faculty Work: Enhancing Individual and Institutional Performance* (Jossey-Bass, 1994). See also the earlier work of K A Feldman who published widely on the impact of time, student characteristics and circumstances on evaluation of teachers. For example: K A Feldman, 'Grades and College Students' Evaluations of their Courses and their Teachers' (1976) 4 *Research in Higher Education* 1; K A Feldman, 'The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data from Multisectional Validity Studies' (1989) 30 *Research in Higher Education* 583.

[77] *Ryerson University v Ryerson Faculty Association*, 2018 CanLII 58446 (ON LA).

[78] Lau, 'How to Encourage Student Voice' (n 20).

together with the biases, characteristics and circumstances of the students, prior higher educational experience and employment outcomes — conceivably affect expectations and influence their perception of the quality of the teaching, the unit of study, and the course as a whole, thus making the overall use of SETs far more uncertain than its numerical results alone may *prima facie* suggest**.**

B Case Study #2: The Philosophy of Law (LAW00520)

The unit of study 'The Philosophy of Law' is a mandatory core unit for all students enrolled in a LLB degree at the SLJ. Students generally complete this unit in their second or third year of study, with around 150 to 200 enrolments every year. As is the case in most units in the SLJ, the student cohort is comprised of primarily online students (around 81 per cent), with the remainder distributed between the Lismore and Gold Coast campuses.

The unit is relatively unique, as it is one of the most noticeable departures from the overall type of core units generally expected of law students. Its critical and theoretical nature is seen as either rewarding or irrelevant for future lawyers. There is often a high degree of student resistance to the unit's content, at least at the onset of each teaching period.

Importantly, the unit introduced, unique among all units of study within the SLJ's LLB awards, a closed book exam in 2013. This closed book exam consisted of 13 broad questions,[79] communicated to students at the onset of the semester, and covered in detail throughout the entirety of the teaching session. Strong alignment between tutorials and assessment was deeply embedded in the pedagogical design of the unit. In the exam, students were asked to answer five questions out of seven, randomly selected among those 13.[80]

Given the negative response from students to the introduction of a closed book exam in 2013, an open book exam was introduced the following year, in 2014. The change, however, was also thoroughly monitored, to measure and evaluate whether the *perception* of difficulty in relation to a closed book exam, strongly voiced by students, was indeed reflective of an *effective* increase in difficulty, and related decrease in student results.

In order to reduce the variables, and secure the reliability of the observation, the only element to change within the unit was a shift from a closed book exam to an open book exam. While the questions were necessarily changed (and could not be communicated to students at the unit's onset), they were nonetheless close to the overall details discussed in tutorials. While it is clear that some degree of difference occurs in the structure and content of the exam papers, all efforts were made to align them as closely as possible, thus minimizing the

---

79 Such as 'describe the main lineaments of legal positivism'.

80 Additionally, students always have the opportunity to complete an *optional* essay in week 8. Generally, only half of the student cohort elect to do so.

differences among the two.[81] Furthermore, all other elements of the unit remained identical: lectures and tutorials were unchanged and their delivery was conducted in exactly the same manner. Moreover, the delivery was conducted by the same teaching team as the previous year, to exclude any personal bias based on different teaching personnel. As a result, the only two changed conditions from 2013 to 2014 were the closed book/open book exam format, and the student cohort. In order to isolate the exam issue, the student cohort was monitored in all other units of study undertaken at the same time as The Philosophy of Law. No significant changes were detected in all other units of study (whereby cohorts performed in a pattern consistent with previous — and subsequent — years), and thus, it was possible to isolate a direct correlation between the exam conditions and student results. As a corollary of this, it was also possible to measure any variance in student surveys. Once the observations were complete, the unit reverted, in the following year, to a closed book exam, and the results were again closely monitored for two further years, under identical conditions.

As Table 2 shows, when faced with an open book exam, the amount of fail grades increased significantly (by almost 50 per cent), whereas the amount of overall distinctions and high distinctions decreased equally significantly (again, by around 50 per cent). It was apparent that student *results* were not only disconnected, but also actually inversely related to student *perceptions* about the closed/open book nature of the exam. Significantly, student perceptions appeared to be more relevant than actual overall results insofar as SETs were concerned. The overall student satisfaction with the unit (the most important of the SETs questions) changed significantly when the open book exam was offered to students, leading to the only result above four (out of five), compared against the relatively stable 3.6 average in all other instances.[82]

---

[81] Naturally, some degree of difference remains in terms of the exam questions, to account for the closed book/open nature of the exam format. However, these differences were carefully addressed to minimize the variance. It must be noted that a degree of variance occurs among *all* exam papers within units, unless exam papers are identical in every exam period, and thus some inference must be able to be made when differences are appropriately accounted for.

[82] Again, this data is drawn from an analysis of answers to the standard Question 7 in SCU SETs.

Table 2
Results from open book versus closed book exams

		2013 closed book	2014 open book	2015 closed book	2016 closed book
Percentage of respondents		49%	37%	40%	38%
SETs results		3.74	4.06	3.64	3.54
Student results	Fails	23%	33%	22%	21%
	Distinctions/HDs	27%	11%	22%	24%

While further research (such as, for example, the administration of an open book/closed book exam randomly allocated to the same cohort) would deepen the inferences drawn from this case study, the analysis already suggests an inverse relationship between student perception and related satisfaction, as measured by SETs and actual student performance. Moreover, since the initial change away from a closed book exam had been driven by student arguments that placed sufficient pressure upon the School to demand the change to an open book exam, the observations also indicated the very relevant problem of how a consumerist mentality is privileged in driving pedagogical choices, at the demonstrable cost of actual student results and teaching effectiveness.

C Case Study #3: Introduction to Business Law, and the impact of international students on SETs

International students are both an important source of revenue and a significant source of diversity. International students introduce new perspectives, foster a diverse campus environment, enrich the learning environment with different cultural perspectives as well as creating significant income opportunities for the students themselves.[83]

Nyland et al point out that Australia has been a major force within this international student market and has been very successful in recruiting international students as supported through neoliberal policies and agendas set by the government and higher education institutions. [84] SCU has been a significant participant in this international student market with the percentage of international students increasing from 13.7 per cent of its total student population in 2014 to 28 per cent in 2018.[85]

---

83 Ravichandran Ammigan and Elspeth Jones, 'Improving the Student Experience: Learning from a Comparative Study of International Student Satisfaction' (2018) 22(4) *Journal of Studies in International Education* 283.

84 Chris Nyland et al, 'International Student Workers in Australia: A New Vulnerable Workforce' (2009) 22(1) *Journal of Education and Work* 1.

85 Southern Cross University Office of Planning Quality and Review, 'SCU at a Glance 2014-2019' (2019) <https://www.scu.edu.au/media/scueduau/staff/planning-quality-and-review/SCU-At-A-Glance-2014-19.pdf>.

Throughout these five years, the unit 'Introduction to Business Law' has experienced numbers of international students much greater than the overall SCU cohort due to it being a mandatory unit in both the Bachelor of Business and Bachelor of Tourism awards. These courses attract larger numbers of international students than most undergraduate programs. Table 3 shows the significant difference between the percentage of international students enrolled in Introduction to Business Law, and SCU overall.

Given the insignificant numbers of international students enrolled in its courses, the table also provides, from the SLJ's perspective, a rare insight into the effect of the university sector's increased focus on recruitment of international students.

Table 3
Comparison of enrolments

Year	LAW00150 Total Enrolment	LAW00150 International Students (%)	SCU International Students (%)	LAW00150 SETs results All students
2009-13	308 (mean)	28 (mean)	No data	4.72 (mean)
2014	755	28	14	4.48
2015	725	38	15	4.32
2016	925	52	18	4.29
2017	819	60	21	4.38
2018	856	62	28	4.40
2019	862	63	27*	4.33

This data shows that student unit satisfaction has decreased in the unit as the percentage of enrolled international students has increased. In the five years prior to 2014, the Overall Satisfaction mean score was 4.72 — with international students making up 28 per cent of the unit cohort. In 2019, the Overall Satisfaction mean score had decreased to 4.33 while the percentage of international students had increased to 63 per cent.[86] While the difference may indeed be statistically insufficient to prove any causal relationship, we believe that a correlation is nonetheless visible. Indeed, such correlation becomes even more apparent when multiple survey questions are considered (Table 4).

[86] The Overall Satisfaction with the Unit (or 'Question 7') is the concluding survey question that asks students to rate overall their satisfaction with the unit/teacher. It is the SET question that figures almost exclusively in yearly performance review interviews.

Table 4
International and domestic student satisfaction

Question	2016		2017		2018		2019	
	Domestic	International	Domestic	International	Domestic	International	Domestic	International
I am satisfied with the assessment tasks in this unit	4.27	3.81	4.40	4.04	4.53	4.07	4.49	3.81
I am satisfied with the way this unit was taught/delivered	4.37	3.76	4.49	4.09	4.49	4.14	4.53	3.93
Overall, I am satisfied with this unit	4.37	3.92	4.43	4.14	4.50	4.10	4.61	3.88

To explain why such differences may occur, Picker et al have identified 'legal cultural issues' as significant.[87] Among these legal cultural issues, the authors suggest that the alien terrain (students coming from different legal systems), the different role of courts and government, social context, and different religions, ideologies and culture may adversely affect student performance, and hence satisfaction.[88] The authors also note the significance of logistical issues involving visas (ongoing bureaucratic demands can prove debilitating), emotional and psychological issues arising from their distance from familial support, and also mundane issues such as dealing with banks, mobile phone providers, and universities themselves.[89]

Do these myriad issues facing international students influence their overall learning experience and, of particular relevance to this case study, the international students' unit satisfaction? Empirical evidence in the form of student comments in this unit's SETs suggest that there are different views on the learning experience depending on whether the responder is an Australian resident/citizen, or an international student. Whereas the resident/citizen comments were generally highly positive, those of international students were often highly critical.

Examples from the former group include:

- I don't think much improvement is needed, I found it very understandable and clear.

---

87 Colin Picker et al, 'Comparative Perspectives on Teaching Foreign Students in Law: Pedagogical, Substantive, Logistical and Conceptual Challenges' (2017) 26(1) *Legal Education Review* 161.

88 Ibid 167–171.

89 Ibid 179–182.

- The lecturer brilliantly breaks things down into easy to understand words and terms.
- This unit will help me both in my future career and in life. So relevant.

Examples from the latter group include:

- Many things were not cleared in this unit. Lecturer was confused me.
- Let the language be simple, because there are many people who came from different places, and we cannot understand.
- As a Chinese student this subject is meaningless to me. Why should I have to study Australian legal system?

The comments were analysed to identify the most common themes, with five emerging more clearly throughout the majority of the surveys. Once identified, the surveys were further analysed to determine the positive or negative responses provided by students in relation to each of those themes. The results, presented in Table 5, seem to confirm the above hypothesis.

While certainly unable to prove any causative relationship, the data nonetheless suggests that Unit Assessors may have no control over the decline in a key performance indicator because of a University-wide effort to increase the number of international enrolments.

Table 5
Positive and negative responses from students

Common themes in student responses		2016		2017		2018		2019	
		Domestic	International	Domestic	International	Domestic	International	Domestic	International
Assessment	positive	87%	15%	78%	33%	95%	77%	90%	15%
	negative	13%	85%	22%	67%	5%	23%	10%	85%
Workload	positive	81%	27%	85%	30%	86%	69%	87%	33%
	negative	19%	73%	15%	70%	14%	31%	13%	67%
Teaching	positive	92%	53%	91%	45%	93%	64%	94%	50%
	negative	8%	47%	9%	55%	7%	36%	6%	50%
Good unit	positive	94%	31%	85%	38%	84%	54%	79%	38%
	negative	6%	69%	15%	62%	16%	46%	21%	62%
Language understanding	positive	94%	25%		5%		0%	90%	0%
	negative	6%	75%		95%		100%	10%	100%

IV CONCLUSIONS

The three case studies generally align with the numerous concerns raised by the literature. SETs are affected by unavoidable — and, more importantly often unidentifiable — bias. Student and faculty, gender and race, educational level, course characteristics (elective versus compulsory), class sizes, quantitative versus qualitative courses, traditional (face to face) versus online teaching, are among the most important factors for which SETs are incapable of adjustment. John Lawrence submits that it is not possible to compare 'apples and oranges', as 'it makes no sense to compare SETs scores of very different classes, such as a small physics course and a large lecture class on Shakespeare and hip-hop'.[90] The overall measure of teaching 'quality' is still profoundly vague in the case of SCU SETs. And thus, although '[t]he concept of quality is primarily that of fitness for purpose',[91] SETs are still beset by a fundamental definitory problem, in that quality is *inferred* rather than pre-determined, and then appropriately measured. The Group concurs with Alderman et al, in noting that '[m]any [SETs] are poorly conceived and designed; and generate data sets that cannot be validated, are used for inappropriate purposes...or are ignored by those who could benefit from...the feedback'.[92]

Furthermore, bias — certainly individual, and possibly influenced by gender and race considerations — inform the totality of the surveys investigated by the Group. As Goos and Salomon observe, 'the signal SETs provide on teacher quality is contaminated by noise.'[93] Hornstein had already suggested that:

> [s]tudent satisfaction is a complex phenomenon influenced by a number of variables … image and tradition as well as the availability of adequate facilities, classrooms and resources at postsecondary institutions significantly contribute to overall student satisfaction.[94]

We agree with Hornstein in asserting that '[t]hese findings…suggest that teaching competence is not a component of its assessment'.[95]

All case studies clearly show that surveys are, at best, an overview of students' *opinions* of teaching, rather than a valid form of assessment of teaching capabilities. The emphasis on statistical — and, particularly, median — results creates further confusion, since '[a]verages of students' ratings appear objective simply because they

---

90 Lawrence, 'Student Evaluations of Teaching are Not Valid' (n 38).

91 Chenicheri Sid Nair, 'Evaluation of Subject, Teaching and Research' (Conference Paper, HERDSA, 2002) 482.

92 Alderman, Towers and Bannah, 'Student Feedback Systems' (n 4).

93 Maarten Goos and Anna Salomons, 'Measuring Teaching Quality in Higher Education: Assessing Selection Bias in Course Evaluations' (2017) 58(4) *Research in Higher Education* 341, 343.

94 Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool' (n 10) 4.

95 Ibid.

are numerical'.[96] After all, '[i]f you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference.'[97] The Group agrees that SETs are a measure of 'popularity and liking (utility) rather than *bona fide* measures of teaching capability'.[98]

At least one of the case studies (#2) shows very powerfully the inverse relationship between student performance and student satisfaction. Student results were significantly higher, albeit clearly less popular, when SETs results were markedly lower. Furthermore, in that instance, pressure on the part of students directly forced the change in assessment regimes in the unit, showing unequivocally the power that SETs exert on pedagogical choices. However, in this instance, it also became apparent how such pressure is often a result of students' perceptions, rather than being based on objective data. Unfortunately, pedagogical choices ultimately beneficial to students (as indicated by the case study results) may often be suppressed by student voices, since 'Universities generally treat SETs as if they primarily measure teaching effectiveness or teaching quality', [99] thereby, at the same time, allowing them to drive pedagogical choices and reinforcing the articulation of students as customers/consumers of education, rather than co-creators of knowledge.

While keen to further explore the issue, possibly through the additional research engendered by the case studies themselves (such as the administration of a closed/open book exam to the same cohort), the Group nonetheless already finds that 'student evaluations, with all the biases they embrace, put pressure on faculty to go slow and not rock the boat',[100] with the perverse effect of 'turning faculty teaching into a form of entertainment that plays [to what is called] "the applause meter".'[101] This may very well be because, as Hornstein suggests, 'administration wants to retain students and prefers a low-cost system to monitoring faculty that looks "objective"',[102] or it may be an unintended consequence of a well-intended measuring tool. Whichever the case, there can be no doubt that, 'if the objects in the evaluation instrument are unclear and the criteria measuring those objectives are vague, there will be an unsatisfactory payoff for [all concerned].'[103]

From a tool initially designated for teachers to reflect on their pedagogical practices, SETs have become, willingly or not, explicit

---

96 Ibid 2.
97 Darrel Huff, *How to Lie with Statistics* (Penguin Books, 1991) 72.
98 Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool' (n 10) 4.
99 Boring, Ottoboni and Stark, 'Student Evaluations of Teaching' (n 15).
100 Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool' (n 10) 6.
101 Giroux, 'Once More, with Conviction' (n 36) 121.
102 Hornstein, 'Student Evaluations of Teaching are an Inadequate Assessment Tool' (n 10) 5.
103 Tarun and Krueger, 'A Perspective on Student Evaluations' (n 52).

parameters for managers to determine promotion applications and tenure.[104] Now, SETs are becoming tools to define tertiary funding (and thus, tertiary education policy). As soon as SETs and their results are removed from the control of academic teachers, they are manipulated to leverage, albeit indirectly, the success rate of any particular unit. In this way SETs have been 'weaponised'.

While SETs may appear as secondary tools in the overall landscape of pedagogical and political considerations within which universities are enmeshed, their impact is disproportionate. On the one hand, they determine individual academic careers, albeit not necessarily as a reflection of the individual's actual teaching competence. On the other hand, as the case studies show, they disproportionately influence pedagogical choices, often detrimentally. This problem has now been exacerbated by the linkage of additional federal university funds to 'satisfaction numbers' as indicated by current results, leading to the inevitable conclusion that the overall quality of Australian tertiary education is likely to decrease.

One may wonder what consequences can be drawn from such a negative assessment of SETs by both the literature and our case studies. An immediate question is, indeed, whether SETs should be altogether abandoned, and, if so, whether they should be replaced by other measures to monitor, reflect, and ultimately improve teaching capabilities and student experience (a corollary of such an alternative is the question as to what data should SETs collect). Two corollary questions, irrespective of the answer given to the above, are *who* should be looking at the data being collected, and *how* this (re-defined) data should be used.

These questions are not unique to SCU, and a number of interesting answers were provided to the Group upon presentation of these findings at the 2019 Australasian Legal Academics Association conference. Some universities have proposed less frequent surveys, while others have emphasised more 'teaching-oriented' qualitative questions. The University of Auckland has substituted SETs altogether with staff-student consultative committees and 'select student representatives'. Solutions certainly abound, and the literature appears unanimous in suggesting that 'teaching evaluation should be used for formative purposes, to help faculty improve teaching, and not merely for summative decisions...', [105] and that the ultimate 'development of [valid] measures of teaching effectiveness...would lead to enhanced teaching quality'.[106]

As a result of our preliminary evaluations, we propose four main avenues of reform:

1. If SETs are to be retained, they should revert to their originally intended purpose as a tool for self-reflection. To

[104] SCU Promotions Application Policy is an example of this, whereby their use is explicitly stated.

[105] Anderson, Cain and Bird, 'Online Student Course Evaluations' (n 45).

[106] Fan et al, 'Gender and Cultural Bias in Student Evaluations' (n 47).

avoid SETs being used for *summative* purposes (such as performance reviews), and rather as *formative* tools (as originally intended), numerical data should be abandoned, or at least significantly de-emphasised, and, instead, precise qualitative questions should be designed.

2. If numerical data is to be included, it should be limited to *objective* questions (such as audibility of the instructor, legibility of notes, etc.), which could lead to actionable choices to be balanced against a host of other pedagogical considerations.
3. Both these re-defined quantitative results and qualitative answers should be discussed as a faculty (either by being de-identified or through small group workshops) or, at least, in peer review teams, since 'ideas of pedagogical well-being and emotional well-being are interlinked'.[107]
4. Finally, students should be involved, either throughout the teaching session or at a later stage, though a selected student representative voice.

These proposals would have the effect of better aligning SETs with both the literature and teaching practice, as one of several non-exclusive means to measure student satisfaction. In this way, SETs may fulfil their original ambition — to be tools that inform and shape good pedagogical practice, rather than blunt weapons of simplistic and often flawed application.

[107] Lynch, 'Teachers' Experiences of Student Feedback' (n 35).