What student evaluations are not: Scholarship of Teaching and Learning using student evaluations

Ameera Ali
York University, Canada, aliamira@yorku.ca

Joseph Crawford
University of Tasmania, Australia, joseph.crawford@utas.edu.au

Leela Cejnar
Royal Holloway, University of London, l.cejnar@outlook.com

Kristyn Harman
University of Tasmania, Australia, Kristyn.Harman@utas.edu.au

Kwong Nui Sim
Auckland University of Technology, New Zealand, kwongnui.sim@aut.ac.nz

Follow this and additional works at: https://ro.uow.edu.au/jutlp

What student evaluations are not: Scholarship of Teaching and Learning using student evaluations

Abstract

In this Editorial, we stay committed to the objective of the Journal of University Teaching and Learning Practice regarding sharing, evaluating, and developing stronger evidence-based practice papers by focusing on the topic of national and institutional student evaluations. We create an important theoretical and practical foundation for authors considering publishing with our Journal on studies that utilise student surveys as their primary method of data collection. The editorial begins by providing a comprehensive overview of the history and emergence of student evaluations dating back to medieval times, we trace the evolution of student evaluations to present day looking at the rationale behind the induction of such tools. Following this, we discuss the validity of student evaluations through an exploration of factors such as student satisfaction, the timing of when student surveys are administered, and the idiosyncrasies regarding paper-based and online evaluations. We then further discuss the reliability of student evaluations by contextualising what student evaluations do not say and uncover how various forms of bias can influence the ways student evaluations are both completed and interpreted. Through this we assert that due to confounding factors of bias that influence the results of student evaluations, they cannot always be thought of as wholly objective data collection tools. This then leads into our discussion of the contemporary social contexts within which student evaluations are situated and how both micro and macro dynamics influence student experiences of teaching and learning, where we contest that broader external factors experienced by students can skew the ways teaching is both perceived and evaluated. We conclude our Editorial by critically envisioning a new direction for future manuscript submissions to our Journal. We assert that although the use of student evaluations as evidence of teaching practice may be inherently flawed, there nonetheless remains merit in their use following critical and reflexive engagement throughout the research process. As such, we are hopeful that our critical review of student evaluation-based scholarship may be utilised to leverage higher quality research output.

Practitioner Notes

1. Student evaluations are often used as primary tools of data collection to offer significant findings pertaining to student experiences of teaching, yet studies have shown that data produced in such a way can have issues with validity, reliability, and bias.

2. Broadly speaking, researchers who utilise evidence drawn from student evaluations of teaching in their SOTL publications tend to so without critiquing these tools.

3. Although the use of student evaluations as evidence of teaching practice may be flawed, there remains merit in their use considering that critical and reflexive engagement is observed throughout the research process.

4. Remaining methodologically reflexive and explicitly acknowledging confounding factors throughout the data analysis procedure can help to directly address complexities involved as well as enhance future research trajectories for others.

5. JUTLP invites contributions that deconstruct the ways student surveys are used within SoTL while exploring ways to respond to these challenges.

Keywords

Introduction

Student experience and teaching evaluation surveys are a veritable norm among higher education institutions globally. Their role in curriculum evaluation activities is typically embedded in institutions and local contexts through quality improvement, quality assurance, continuous improvement, assurance of learning, and/or teacher as action researcher processes. Since their inception, legislation that mandates specific tools be deployed has been introduced and enforced (e.g., Bertaccini et al., 2019), with a growing volume of scholarly publications documenting their use (e.g., Patterson et al., 2020).

The survey tools deployed by teachers and institutions at the end of each study period, and at nationally signposted times, afford the opportunity to identify and implement improvements in teaching such as formative feedback and to benchmark and assure attainment of learning outcomes through summative assessment and feedback. Scriven (1967: cited in Taras, 2005) first distinguished between formative and summative feedback in the program evaluation context. These terms are perhaps more familiar to those implementing formative and summative assessments, yet they share conceptual similarities. Institutionally introduced teaching evaluation surveys were designed to provide insight to those administering them on the respective attainment of learning for students, and the quality of teaching and support systems that facilitate such learning (Davies et al., 2010). However, and with greater frequency, there are calls of critique for these blunt instruments and what they have been designed to achieve. This Editorial reflects on the affordances of these surveys in the contemporary context of the scholarship of learning and teaching (SoTL), and does so with a focus on the types of manuscripts we review, reject, and publish in the Journal.

We receive manuscripts claiming *significant* findings pertaining to decontextualised and broad stroke student evaluation and student experience surveys. We find these, at times, are analysed rigorously to share evidence of innovative teaching practices, some aligned to high quality curricula and co-curricular activities. However, some manuscripts write descriptively to paint a picture the data may not support. Such articles may state, for example, differences in mean scores across gender, or similar, assuming that the raw data differences must be significant. While significance might exist, the research may also fall short if the analysis draws on a skewed sample, does not engage with statistical significance, or occupies a context whereby the student population has unique differences across demographic characteristics. In a good practice example, O'Dea et al. (2018) highlight that with carefully and targeted analysis, student evaluation with appropriate significance analysis, data can be used to reshape our understanding of theory and practice.

As the *Journal of University Teaching and Learning Practice* progresses towards sharing, evaluating, and developing stronger evidence-based practice papers (see Percy et al., 2021), we see clarification of national and institutional student evaluations as an area requiring review. The aim of this Editorial is to create an important theoretical and practical foundation for authors considering publishing with our Journal on studies that adopt such surveys as their primary method of data collection. We believe this is a crucial step towards ensuring that the work we publish reflects the qualitative whole-of-student, institutional, and national contexts within which they are situated. This also means that we will seek to ensure that manuscripts progressing to review will include accounts of assessment using appropriate measurement tools and draw on related quantitative validity and reliability testing to ensure rigour .

We begin this critical Editorial with an account of the history of student evaluation of their learning and experiences. We explore the way in which evaluation tools seek to measure student learning,

student experience, and attainment of learning outcomes. We believe that to better understand what information student evaluations do not provide and their biases, we need to examine the contexts within which student evaluations are situated. We conclude by exploring what this means in the context of future manuscript submissions, and how to leverage this critical Editorial of student evaluation-based scholarship for high quality research output.

Why student evaluations

The rationale for universities producing regular student evaluations of learning has a long history dating back at least to medieval times. At medieval universities, student guilds had significant input into decisions taken about the tenure, remuneration, and working conditions of the masters who, at the time, were usually lawyers, doctors, and theologians (Powell, 1972). Student influence waned over time but re-emerged in a new form in a few American universities from the 1920s when systematic attempts to gather and utilise student feedback about quality of teaching were introduced (Maassen, 1997; Powell, 1972). This practice became increasingly widespread across, then beyond, the United States of America as the twentieth century progressed.

The global uptake of student evaluations of teaching can be attributed to several key drivers. Higher education in the latter half of the twentieth century was characterised by a substantial increase in tertiary enrolments as universities engaged in strategies of 'massification' (Dian-Fuh & Yeh, 2012). The numbers of higher education institutions increased as did staffing. New institutions were being founded, and new staff were being recruited and appointed across the burgeoning sector. At the same time, learning and teaching practices were changing rapidly with more sophisticated teaching pedagogies and technologies becoming available. These significant changes informed universities' increasingly widespread use of student feedback mechanisms from the 1960s onwards, initially as a lens through which individual lecturers could reflect on and improve their approaches to learning and teaching (Barrow & Grant, 2015).

Between the 1960s and 1990s, student feedback became increasingly institutionalised with universities centralising the tools used to gather feedback from students, making use of the resulting data to establish 'quality assurance' norms and to draw comparisons between perceived levels of teaching performance. The emphasis shifted from teaching faculty using student feedback to improve their practice, as shown in the work of Laurillard (2008) with her focus on the teacher as action-researcher, to utilising it in support of other processes, including applications for tenure or academic promotion (Barrow & Grant, 2015).

Following a systematic review (*n* = 1,964 journal articles), Pineda and Steinhardt (2020) identified the United States of America as being at the forefront of utilising student evaluations of teaching dating from the 1970s, with German-speaking countries adopting these instruments in the mid-1990s, and China and Latin America following in the early 2000s. Why student evaluations of teaching are used differs depending on different cultural contexts. For example, Pineda and Steinhardt (2020) discerned that while in the Americas student evaluations of teaching were seen predominantly as a managerial quality assurance tool, in Chinese and German-speaking countries there has been a greater emphasis on quality improvement in teaching practice. These researchers found that while student evaluations of teaching are now used globally, other countries or regions, including Japan, France, and Scandanavia had insufficient publications available from which to draw conclusions.

As increasing numbers of universities introduced student evaluations of teaching, governments began to see how they and their agencies might benefit from regulating the use of these tools to create sector-wide quality assurance measures which, following 'massification', had seen a substantial increase in the costs involved in its delivery. For example, in the United Kingdom and

Australia in the 1990s, governments and their agencies began utilising student evaluations of university teaching as a quality assurance tool, co-opting a practice that had previously been the sole province of individual institutions (Chalmers & Hunt, 2016). Likewise, in Italy, regularly obtaining students' feedback on their courses became the responsibility of universities from 1999 by ministerial decree and formed part of a wider auditing process (Bertaccini et al., 2019). In a similar vein, over recent decades the rise of neoliberal agendas in relation to the higher education sector has seen a rise in the use of national surveys to gather feedback from students about their levels of satisfaction with their courses. Such surveys include the UK National Student Survey, the Australian Student Experience Survey, and the Australian Course Experience Questionnaire, with similar surveys being carried out in Canada, Denmark, Finland, Ireland, Japan, the Netherlands, Norway, and the United States of America (Arthur, 2020; Winstone et al., 2021).

In the twenty-first century, the continuing use of student evaluations of teaching is tied to the extraordinary value to domestic economies of higher education as an export commodity. It is, for example, the fourth largest export sector in the United States of America and the third largest in Australia (Ewing, 2021). As Uncles (2018, p. 187) has explained, albeit in a pre-COVID-19 context, higher education across the globe has become "a terrific success story by most measures of industrial success" with the "explosion of data capture" enabling higher education institutions, governments, and other interested parties to take an evidence-based approach to measuring and evaluating the provision of higher education. Student evaluations of teaching fit within a much wider context within which data is captured at every stage of study from the application process through to campus heat maps and learning management system usage to graduation destination data. With many institutions being focused on providing student-centered educational experiences, the student voice as heard through evaluations of teaching remains core to evidencing teaching performance despite controversy around the efficacy of survey tools.

Broadly speaking, researchers who utilise evidence drawn from student evaluations of teaching in their Scholarship of Teaching and Learning (SOTL) publications tend to so in a descriptive way without necessarily critiquing the tool from which they are citing. Yet studies have shown that the data produced using these evaluative tools can have issues with validity and reliability. For example, Arasaratnam-Smith et al. (2021) provide a quality analysis of the qualitative comments attached to differences in the Australian national student experience survey. This work uses the generic national quality assurance tool and seeks to add some degree of depth. However, it does not use validity or reliability measures to identify if the differences projected can be relied upon. This is in contrast with Downie et al. (2021) who draw on significance testing to compare teacher and student technology-enhanced learning use. These examples highlight the importance of understanding historical climates for critical scholarship of teaching and learning that draws on contemporary survey tools.

How evaluations are not always valid measures

To go past the historical discussion, and into the rigour of student evaluation tools, we move on to discuss challenges with aggregation, temporal bias, and validity assessments. The lack of effective tools to measure student perceptions of learning is not a new issue in higher education (Wieman, 2015). Nor is it a new issue in the measurement sciences. Grafton et al. (2010) highlight that for performance evaluation measures to be effective, they should facilitate future employee and managerial decisions. In early twenty-first century organisational theory, Luthans (2002) challenged scholars to build more relevant core theoretical concepts that would lead to best practices. More recently, a national review of U.S. leadership development programs identified an evidence-practice gap identified with measurement tools typically developed in the absence of effective theory (Loew & Wentworth, 2013). Higher education is no exception to the challenges of effective program

evaluation, with recognised challenges which include the perceived validity of the tools used to assess and assure quality. For example, Clayson (2020) identifies that student evaluations are not supporting effective distinguishing of high- and low-quality teaching, particularly when the questions in the student evaluation form tends to fail the differences between high and low-quality teaching.

Aggregation is another 'weakness' of student evaluations. Kitto et al. (2019) argues generic aggregation of median values without assessment and adjustment based on sample- or scale-specific skewness can create less rigorous student evaluations. This is particularly true when the evaluation modality serves as an administrative quality assurance tool instead of providing meaningful quality improvement opportunities for teaching faculty. It can be argued that the intention of an evaluative measure is to understand how the programme went, and what could be improved in subsequent deliveries. Importantly, there is a distinct lack of engagement in validity measures to assure data rigour before application. Indeed, the use of quantitative Likert-scales rarely deploy factor analyses, structural equation modelling, or to examine the validity of the items against their underlying theoretical concepts. This is essential in the context of a research paper using student evaluations as primary or secondary data.

Evaluations in higher education contexts are typically used in post-period feedback measures, instead of mid-period concurrent or feedforward controls. While offering feedback, it neglects the opportunity to provide concurrent controls to enable learning and experiential improvements for cohorts within that delivery period. The shift from paper-based evaluation to online evaluation, coupled with end-of-semester student invitation emails, has a huge impact on student participation rate. Student participation rates drop dramatically when this process is not being undertaken within a physical space, perhaps challenged by the concept of joint attention (e.g., Tice et al., 2021). In addition, Ahmad (2018) identified that response rates are improved through high quality academic communication. The implications for these discrepancies in participation are important. If a student evaluation with a low response rate is a primary measure in teaching performance, managerial decisions may ineffectively assign performance for staff (Luo, 2020). When reflecting on the researcher implications, authors using these tools to make quantitative relationships between measures; consideration to the temporal nature of the delivery period and situated nature of students providing their evaluation is necessary.

As an aside, most student evaluations consist of both quantitative (e.g., scales of measurement) and qualitative (e.g., comments or suggestions) questions. When the quantitative data is used for research purposes, there is also benefit in considering the parallel qualitative data collected to help explore how data points change between deliveries (Palmer et al., 2019) or particular contextual challenges (see below). This data may also serve as a form of validity assessment, to seek qualitative unpacking of sentiment and how this aligns (or otherwise) to the underpinning theoretical concepts (Shah & Pabel, 2020). Bartkowiak-Theron et al. (2020) argues this data is important, and there may be a need to rethink student evaluations to be more meaningful and support greater alignment between the concept of 'teaching quality' and how it is measured.

What student evaluations do not say reliably

While validity is important, so too is the reliability of student evaluation data. Although student evaluations of teaching are often used by institutions as neutral, acceptable, and teaching-centric tools, there are idiosyncratic factors that prevent them from truly being so. While researchers ought to apply reliability tools (e.g., Cronbach alphas and composite reliability), reliability tests (e.g., split-half: see Drost, 2011), and ensure appropriate assessments of difference (e.g., Chi-square testing or *t*-tests), it is important to undertake more detailed analysis to understand how individual differences

may affect specific responses. The demographic makeup of both faculty and students in higher education institutions is increasingly diversified. It is necessary from a reliability perspective to consider how subjective, pre-conceived notions may influence students' perceptions of teaching. Student perceptions of their instructors can have significant influence on how student evaluations of teaching are completed. There are numerous factors that influence what information students include in their evaluations. Indeed, as we go onto discuss further, student evaluations contain forms of bias including, but not limited to race, gender, disability, ethnicity, and religion. These biases may be implicit or explicit, with implicit bias being far more difficult to detect.

Gender inequity is a prevalent issue in higher education (Allen et al., 2021), and it is not surprising that it is also a chief factor of bias in student evaluations. Research has found male teachers to receive significantly higher-rated student evaluations despite female instructors demonstrated to be equally efficient in their teaching (see Boring, 2016; Mitchell & Martin, 2018). Instructor age and perceived attractiveness have also been identified as factors affecting student evaluations of teaching such that students' ratings have been shown to be higher when an instructor is perceived to be attractive and young (see Sohr-Preseton et al., 2016; Murray et al., 2020; Reinsch et al., 2020). For example, Reinsch et al. (2020) found young attractive male teachers receive higher rated evaluations as compared to attractive female teachers. These biases render the concept of teaching effectiveness difficult to measure with reliability.

It is evident categories of identity may inadvertently influence student evaluations of teaching due to issues of implicit bias. Moreover, social identity and group membership can affect possible interpretations of data. Student and faculty intersectionality is further elucidated when considering race as a significant factor influencing student evaluations of teaching (see Reid, 2010; Smith & Hawkins, 2011). The limited empirical research investigating instructor race and student evaluation scores suggest that non-white instructors typically receive lower scores on teaching evaluations compared to their white counterparts. Race is also found to intersect with gender, with white male instructors typically outperforming others on student evaluation measures (Murray et al., 2020). Instructors who are noted by students as having an accent, tend to receive lower evaluations compared with those who are not noted as having an accent (see Murray et al., 2020). Although having a foreign accent is not indicative of one's knowledge, language proficiency, or teaching efficiency, this trait can influence the teaching evaluations an instructor receives. Again, a challenge to researchers seeking to analyse and report on reliable and usable research data.

Student populations have unique differences across demographic characteristics. For instance, in some geographical regions, to critique an instructor's teaching may not accord with socio-cultural norms and perceived imbalances of power, leading students from these backgrounds to evade negative or critical evaluations. Bias in student evaluations can also materialise through factors such as student interest in the course (Murray et al., 2020), instructor likability (see Clayson, 2021; Feistauer & Richter, 2018), cultural-linguistic contexts (Al-Issa & Sulieman, 2007), grades received in the course (Zabaleta, 2007), instructor sexual orientation (Osborne et al., 2007), and affective factors (Grimes et al., 2017). Additionally, many student evaluations of teaching tend to use Likert scales and opinion-based questions which often invoke spontaneous responses rather than deep and thoughtful reflections which can further induce bias in responses (Reinsch et al., 2020).

Our Journal receives manuscripts reporting significant findings from student evaluations of teaching which are used as evidence of effective teaching practice. The considerations outlined above may appear to render data collected through student surveys futile (invalid, unreliable). However, useful insights are possible and such data can be used to evoke critical consideration of research processes. Indeed, remaining methodologically reflexive and explicitly acknowledging confounding factors throughout the data analysis procedure can help to directly address complexities such as bias as well

as enhance future research trajectories for others. As such, we urge authors to be alert to underlying factors which may negatively impact their data analysis, to remain critical of the many contextual factors that may be influencing student responses, and to maintain rigour in the presentation of their data so that their interpretation of results accurately describes what the data suggests. Additionally, awareness of the ways evaluation design can influence the reliability of data when using student evaluations as a tool to measure instructor teaching is critical. Awareness of limitations will ensure any insightful findings are appropriately contextualised and not lost, and provide the basis for suggesting a way forward for future research endeavours.

Context: What contexts student evaluations are situated within

Carefully situated studies acknowledging and accommodating local and national contexts are rare. Context, however, is crucial for understanding how we may translate the qualitative description of any research. In what ways does the specific geographic, temporal, and spatial conditions affect the ability for others to understand the findings of a study? In an evidence-based practice paper this question is essential, including for student evaluations. Each student evaluation takes place in a particular landscape worthy of explication, so that those who read in the future can clearly examine the potential rigour of that study. To provide an illustrative example, we document an application of context within this Editorial, situated within the Australian national higher education landscape, although not all our co-author editors are based there; and within a COVID-19 pandemic landscape.

Over 2020 and 2021, COVID-19 has had a pervasive effect on higher education's educational, social, and financial sustainability. The mid-pandemic response returning to campus has seen an even greater emphasis on teaching as a valued activity, as universities work hard to encourage students back onto their campuses, to recoup losses. The 2020 Australian national Student Experience Survey, released by the Social Research Centre (2021), highlights undergraduate student ratings fell from 78 per cent in 2019 to 69 per cent in 2020, with some of the more established universities experiencing the largest decline in student ratings, regarding the quality of their entire educational experience. Thus, student evaluations in higher education carry a new sense of importance not only to the teacher, but also to the wider institution. Although evaluations have in the past reflected different personal and contextual factors, as well evolving social practice (Ryan, 2014), evaluations also reflect students' values about their learning. For example, as a source of information on how they have been impacted by the dynamics that manifest within learning contexts, as well as the opportunities they feel they have had to access and engage with their learning and with others during periods of uncertainty and challenge, emanating from the pandemic.

For example, when face-to-face teaching unexpectedly shifted to an emergency online teaching context, the transition was universally evaluated negatively (Garris et al., 2020). Although historically the literature has argued that paper-based evaluations and evaluations of face-to-face teaching yield higher overall outcomes than those that are administered online (Lin, 2008; Mau, 2012), students' evaluations of online teaching in the context of COVID-19 were impacted by their emotional well-being and safety, arising from the shock of the emergency transition. In the context of the pandemic, student evaluations therefore reflected students' struggles with the change: the technology-related anxiety, the decrease in ability to interact with others face-to-face, as well as students' perceptions of how effective and confident their teachers were in making the sudden transition to online learning (Garris et al., 2020). The unexpected and unprecedented change meant that new important predictors of evaluations surfaced, such as a student's ability to access the Internet and use technology at their location, how empathetic and responsive a teacher was to a student's situation, how interesting and enjoyable a teacher was able to make the online learning experience, and how connected and engaged students felt, as a result. At some universities, 2020 saw a move to include technology-based questions in their evaluation tools.

What has transpired from this context has been a wave of scholarship of learning and teaching in higher education, centred often exclusively on student experiences of, and responses to, learning during the pandemic (Crawford, 2021). This has raised the important question of how student evaluations might be used in ongoing research, to develop a 'best practice' framework for future responses to unexpected, emergency situations that might force sudden changes to teaching and learning approaches (e.g., Al-Naabi et al., 2021). To add, student evaluations provide a short and time-specific student perspective, and play a significant role in supporting universities to understand what is different as universities seek to reimagine and reform their approaches to teaching and learning. This is beyond the world of lockdowns, Zoom®, MS Teams®, WebEx® and other online learning platforms. Therefore, using these evaluations to reflect on and disseminate teaching and learning practices used before and during the pandemic will have its place in identifying what worked and what might have worked better. But meaningful impact from such research will come from the ability to use these evaluations to create and share new strategies, innovations and approaches, to address how we move beyond the pandemic into a more resilient future of higher education.

Conclusion

In closing, what does it mean to capture student experiences in SoTL research? Student experiences of teaching and learning are varied and are contingent upon various contextual and subjective factors. The methods and tools with which we investigate student perspectives must be situated in the context, and recognise that these elements may influence their validity and reliability. Engaging in SoTL research requires us to always remain reflexive in relation to our methods, tools, and data and the ways that these may be shaped by confounding dynamics. Thus, we invite contributions that recognise the complexities involved in using student surveys as evidence of the quality of teaching, while thoughtfully addressing these complexities and taken-for-granted understandings from evaluations of teaching and learning. We encourage critical discussions that provoke meaningful insights into the ways that student evaluations are used within SoTL, and implore authors to continue to explore ways to respond to the challenges of using student surveys as primary tools of data collection.

References

Ahmad, T. (2018). Teaching evaluation and student response rate. *PSU Research Review*, 2(3), 206-211. DOI: 10.1108/PRR-03-2018-0008

Al-Issa, A., & Sulieman, H. (2007). Student evaluations of teaching: Perceptions and biasing factors. *Quality Assurance in Education, 15(*3), 302-317. https://doi.org/10.1108/09684880710773183

Al-Naabi, I., Kelder, J. A., & Carr, A. (2021). Preparing teachers for emergency remote teaching: A professional development framework for teachers in higher education. *Journal of University Teaching & Learning Practice, 18*(5), 4. https://doi.org/10.53761/1.18.5.4

Allen, K., Butler-Henderson, K., Reupert, A., Longmuir, F., Finefter-Rosenbluh, I., Berger, E., Grove, C., Heffernan, A., Freeman, N., Kewalramani, S., Krebs, S., Dsouza, L., Mackie, G., Chapman, D., & Fleer, M. (2021). Work like a girl: Redressing gender inequity in academia through systemic solutions. *Journal of University Teaching & Learning Practice, 18*(3). https://ro.uow.edu.au/jutlp/vol18/iss3/03

Arasaratnam-Smith, L., Coetzee, N., & Hodson, C. (2021). The double-edged sword of 'best aspects' and 'needs improvement' in student experiences: A qualitative analysis. *Journal of University Teaching & Learning Practice, 18*(3), 06. https://doi.org/10.14453/jutlp.v18i3.6

Arthur, L. (2020). Evaluating student satisfaction – restricting lecturer professionalism: Outcomes of using the UK national student survey questionnaire for internal student evaluation of teaching. *Assessment & Evaluation in Higher Education, 45*(3), 331-344. https://doi.org/10.1080/02602938.2019.1640863

Barrow, M. & Grant, B. (2015). Changing mechanisms of governmentality? Academic development in New Zealand and student evaluations of teaching. *Higher Education*, *72*(5), 589-601. https://doi.org/10.1007/s10734-015-9965-8

Bartkowiak-Théron, I., McShane, A., & Knight, M. G., (2020). Departing from anonymous and quantitative student feedback: Fostering learning and teaching development through student evaluations. *Journal of Applied Learning & Teaching*, 3(1), 1-11. DOI: https://doi.org/10.37074/jalt.2020.3.s1.16

Bertaccini, B., Giusti, A., & Petrucci, A. (2019). Students' opinions on teaching and services provided by the Italian Universities: A proposal for a new evaluation scheme. *Quality & Quantity, 53*(2), 1037-1050. https://doi.org/10.1007/s11135-018-0801-y

Boring, A. (2016). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*(1), 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Chalmers, D., & Hunt, L. (2016). Evaluation of teaching. *HERDSA Review of Higher Education*, *3*, 27. https://www.herdsa.org.au/herdsa-review-higher-education-vol-3/25-55

Clayson, D. (2020). The Association Between Student Learning and Student Evaluations. In *A Comprehensive Critique of Student Evaluation of Teaching*. Routledge. https://doi.org/10.4324/9781003091462-7

Clayson, D. (2021). The student evaluation of teaching and likability: what the evaluations actually measure. *Assessment & Evaluation in Higher Education.* Advanced Online Publication. https://doi.org/10.1080/02602938.2021.1909702

Crawford, J. (2021). During and beyond a pandemic: Publishing learning and teaching research through COVID-19. *Journal of University Teaching & Learning Practice*, 18(3). https://doi.org/10.53761/1.18.3.2

Davies, M., Hirschberg, J., Lye, J., & Johnston, C. (2010). A systematic analysis of quality of teaching surveys. *Assessment & Evaluation in Higher Education, 35*(1), 83-96. https://doi.org/10.1080/02602930802565362

Dian-Fu, C., & Yeh, C. (2012). Teaching quality after the massification of higher education in Taiwan: A student perspective. *Chinese Education & Society, 45*(5-6), 31-44. https://doi.org/10.2753/CED1061-1932450503

Downie, S., Gao, X., Bedford, S., Bell, K., & Kuit, T. (2021). Technology enhanced learning environments in higher education: A cross-discipline study on teacher and student perceptions. *Journal of University Teaching & Learning Practice, 18*(4), 12. https://doi.org/10.14453/jutlp.v18i4.12

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and perspectives, 38*(1), 105-123.

Ewing, L. (2021) Rethinking higher education post-covid 19. In J. Lee & S. Han. (eds.) *The Future of Service Post-COVID-19 Pandemic* (pp. 37-53). Vol. 1. Switzerland: Springer Nature. https://doi.org/10.1007/978-981-33-4126-5_3

Feistauer, D., & Richter, T. (2018). Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation*, *59*(1), 168-178. https://doi.org/10.1016/j.stueduc.2018.07.009

Garris, C., & Fleck, B. (2020, October 19). Student evaluations of transitioned-online courses during the COVID-19 pandemic. *Scholarship of Teaching and Learning in Psychology*. Advance online publication. https://doi.org/10.1037/stl0000229

Grafton, J., Lillis, A., & Widener, S. (2010). The role of performance measurement and evaluation in building organizational capabilities and performance. *Accounting, Organizations and Society, 35*(7), 689-706. https://doi.org/10.1016/j.aos.2010.07.004

Grimes, A., Medway, D., Foos, A., & Goatman, A. (2017). Impact bias in student evaluations of higher education. *Studies in Higher Education, 42*(6), 945-962. https://doi.org/10.1080/03075079.2015.1071345

Kitto, K., Williams, C. & Alderman, L. (2019). Beyond Average: Contemporary statistical techniques for analysing student evaluations of teaching. *Assessment & Evaluation in Higher Education*, 44(3), 338-360, DOI: 10.1080/02602938.2018.1506909

Laurillard, D. (2008). The teacher as action researcher: Using technology to capture pedagogic form. *Studies in Higher Education, 33*(2), 139-154. https://doi.org/10.1080/03075070801915908

Lin, Q. (2008). Student views of hybrid learning, *Journal of Computing in Teacher Education*, *25*(2), 57-66. https://doi.org/10.1080/10402454.2008.10784610

Loew, L., & Wentworth, D. (2013). *Leadership: The state of development programs.* Delray Beach, FL: Brandon Hall Group. https://www.skillsoft.com/assets/research/research_bhg_leadership_development_programs.pdf

Luo, Mingchu Neal (2020) "Student Response Rate and Its Impact on Quantitative Evaluation of Faculty Teaching," *The Advocate*, 25(2). https://doi.org/10.4148/2637-4552.1137

Luthans, F. (2002). The need for and meaning of positive organizational behavior. *Journal of Organizational Behavior, 23*(6), 695-706. https://doi.org/10.1002/job.165

Maassen, P. (1997). Quality in European higher education: Recent trends and their historical roots. *European Journal of Education, 32*(2), 111-127.

Mau, R., & Opengart, R. (2012). Comparing Ratings: In-class (paper) vs. Out of Class (online) Student Evaluations. *Higher Education Studies*, 2(3). https://doi.org/10.5539/hes.v2n3p55

Mitchell, K. & Martin, J. (2018). Gender bias in student evaluations. *Political Science & Politics, 51*(3), 648, 651. https://doi.org/10.1017/S104909651800001X

Murray, D., Boothby, C., Zhao, H., Minik, V., Be´rube´, N., Larivière, V. & Sugimoto, C . (2020). Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLoS ONE 15*(6). https://doi.org/10.1371/journal.pone.0233515

O'Dea, R., Lagisz, M., Jennions, M., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications, 9*(1), 1-8. https://doi.org/10.1038/s41467-018-06292-0

Osborne, D., Duran, A., Davies, P.G., Wagner III, W., Rienzi, B. (2007). Does sexual orientation matter? An experimental assessment of student evaluations of a gay professor. *Educational Psychology Research Focus, 63*, 49-63.

Palmer, S. & Hall, W. (2019). Using qualitative student evaluation data to illuminate quantitative scale item ratings: seeing beyond the numbers. Australasian Association for Engineering Education Conference 2019 (AAEE).

Patterson, N., Schultz, M., Wood-Bradley, G., Lanham, E., & Adachi, C. (2020). Going Digital

to Enhance the Learning of Undergraduate Students. *Journal of University Teaching and Learning Practice, 17*(3), 6. https://doi.org/10.53761/1.17.3.6

Percy, A., Press, N., Andrew, M., & Pollard, V. (2021). Reframing theory of, and for, practice in higher education. *Journal of University Teaching and Learning Practice, 18*(4), 1. https://doi.org/10.53761/1.18.4.1

Pineda, P. & Steinhardt, I. (2020). The debate on student evaluations of teaching: global convergence confronts higher education traditions. *Teaching in Higher Education.* Advanced online publication. https://doi.org/10.1080/13562517.2020.1863351

Powell, J. (1972). Student evaluation of university teaching: A survey of research. *Research in Education*, *8*(1), 1. https://doi.org/10.1177/003452377200800101

Pratt, D. (1997). Reconceptualizing the evaluation of teaching in higher education. *Higher Education*, 34(1), 23-44. https://doi.org/10.1023/A:1003046127941

Shah, M. & Pabel, A. (2020). Making the student voice count: using qualitative student feedback to enhance the student experience. *Journal of Applied Research in Higher Education*, 12(2), pp. 194-209. https://doi.org/10.1108/JARHE-02-2019-0030

Smith, B., & Hawkins, B. (2011). Examining student evaluations of black college faculty: Does race matter? *Journal of Negro Education*, *80*(2), 149–162.

Social Research Centre. (2021). *2020 Student Experience Survey National Report*. Canberra, Australia: Department of Education, Skills, and Employment, Commonwealth Government of Australia. Accessed 9 November 2021. < https://www.qilt.edu.au/docs/default-source/default-document-library/2020-ses-national-report.pdf?sfvrsn=d1237953_5

Sohr-Preston, S., Boswell, S., McCaleb, K., & Robertson, D. (2016). Professor gender, age, and "hotness" in influencing college students' generation and interpretation of professor ratings. *Higher Learning Research Communications*, *6*(3). https://doi.org/10.18870/hlrc.v6i3.328

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Tice, D., Baumeister, R., Crawford, J., Allen, K-. A., & Percy, A. (2021). Student Belongingness in Higher Education: Lessons for Professors from the COVID-19 Pandemic. *Journal of University Teaching and Learning Practice, 18*(4), 2. https://doi.org/10.53761/1.18.4.2

Reid, L. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.Com. *Journal of Diversity in Higher Education*, *3*(3),137-152. https://doi.org/10.1037/a0019865

Reinsch, R., Goltz, S., & Hietapelto, A. (2020). Student evaluations and the problem of implicit bias. *Journal of College and University Law, 45*(1), 114-140.

Ryan, M. (2015). Framing student evaluations of university learning and teaching: discursive strategies and textual outcomes. *Assessment and Evaluation in Higher Education*, *40*(8), 1142-1158. https://doi.org/10.1080/02602938.2014.974503

Ulker, N. (2021). How can student evaluations lead to improvement of teaching quality? A cross-national analysis, *Research in Post-Compulsory Education*, *26*(1), 19-37. DOI: 10.1080/13596748.2021.1873406

Uncles, M. (2018) Directions in higher education: A marketing perspective. *Australasian Marketing Journal, 26*(2), 187-193. https://doi.org/10.1016/j.ausmj.2018.05.009

Wieman, C. (2015). A Better Way to Evaluate Undergraduate Teaching, Change. *The Magazine of Higher Learning*, 47(1), 6-15, DOI: 10.1080/00091383.2015.996077

Winstone, N., Ajjawi, R., Dirkx, K, & Boud, D. (2021). Measuring what matters: The positioning of students in feedback processes within national student satisfaction surveys. *Studies in Higher Education,* https://doi.org/10.1080/03075079.2021.1916909

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