Harnessing data to tackle some ‘myth-perceptions’ about Student Evaluations

Dr Richard O’Donovan
Monash University
Australia

APAC 2023

Background

- Student evaluations of teaching and units (SETU) have become central to good practice and quality assurance in higher education
- SETU now features in assessing the performance of academic staff and is integral to academic promotion & probation at many universities
- Trust in the integrity of SETU is essential for staff to take it seriously and for future students to benefit from current student feedback
- Academics have understandable concerns about SETU being influenced by factors beyond their control – some of these dominate perceptions about SETU, but perhaps these are just myths…?
- Today we'll explore some of these concerns using recent SETU data to see if they qualify as 'myth-perceptions'

Pulse question

What types of factors (beyond Academics' control) do you think - or have heard - might influence SETU results?

Please use Pulse to share your thoughts…

Overview

There's a lot of research (and mixed findings) around SETU. Today we'll use 2022 SETU Teaching data from Monash University to look at:

1. Strong negative feelings dominating SETU
   (e.g., Brenner & DeLameter, 2016; Katrompas & Metsis; 2021; Rap & Paxton, 2021)
2. 'Revenge reviews' from students who fail
   (e.g., Backer, 2012; Miles & House, 2015; Sullivan et al., 2023)
3. Unit size effects
   (e.g., Badri et al., 2006; Miles & House, 2015)
4. SETU during exams penalise units with exams cf. to those without exams
   (e.g., Wagenaar, 1995; Hejase et al., 2013)
5. SETU scores are impacted by 'racist' and 'sexist' attitudes of students
   (e.g., Laube et al., 2007; Boring, 2017; Esaray & Valdes, 2020; Sigurdardottir, 2022).

But first…a brief word on statistical significance & effect size

- *p* values < 0.05 give us **confidence** we found a **non-random** difference
- **BUT** Statistical significance ≠ Practical significance
- So we need to calculate Effect Size too such as Cohen's *d* & η^2 (Cohen,1969,1988)

- Small effect $d > 0.2$ $\eta^2 > 0.01$
- Medium effect $d > 0.5$ $\eta^2 > 0.06$
- Large effect $d > 0.8$ $\eta^2 > 0.14$

But first…a brief word on statistical significance & effect size

e.g., A weight loss study involved 13,000 people in two groups. One group lost more weight, with $p = .01$ (Bhandari, 2020)

But Cohen's $d < 0.1$ so it was only a trivial difference…

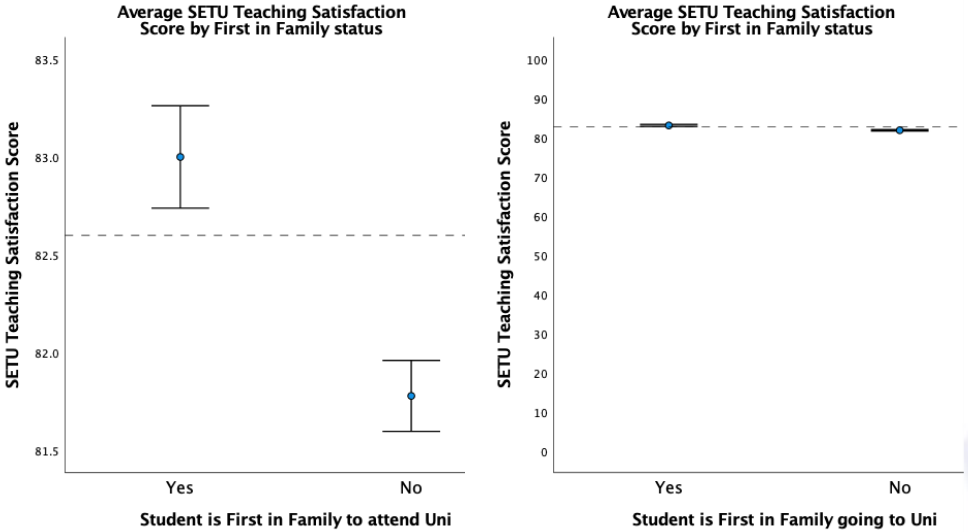
On average, the control group lost 10.6kg while the treatment group lost 10.7kg

The treatment worked, but was only *negligibly better than no treatment at all*

So while Statistical Significance tells us there probably *is a* difference
Effect Size captures the *how big* that difference is

…and a quick note on charts

- Another important consideration is how data are presented visually…
- Consider these charts of First in Family status:



- This looks like a big difference, but if we show the full scale…

  …we see a truer picture of this trivial effect size (Cohen's *d* = 0.06)

Pulse question

Do you think SETU results are dominated by students with strong negative feelings?

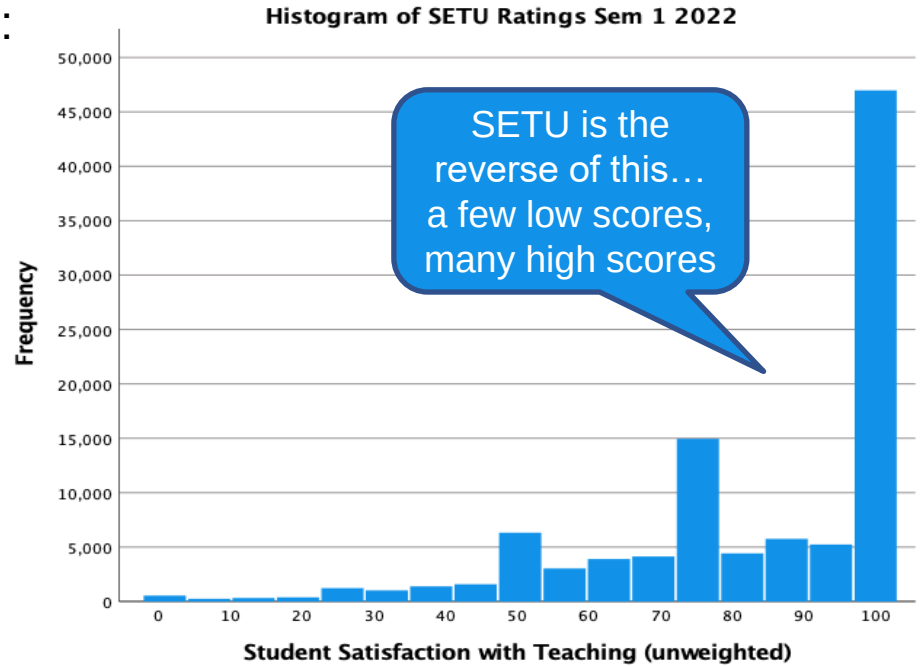
- ☐ Yes
- ☐ No
- ☐ Can't guess

Please use Bluepulse to share your thoughts…

Is SETU dominated by students with strong negative feelings?

One way to investigate this is to look at a chart of SETU responses…

If true, we might expect something like this:



CONCLUSION: ***No.*** If strong feelings are motivating students to complete SETU, they seem to be predominantly positive ones.

Pulse question

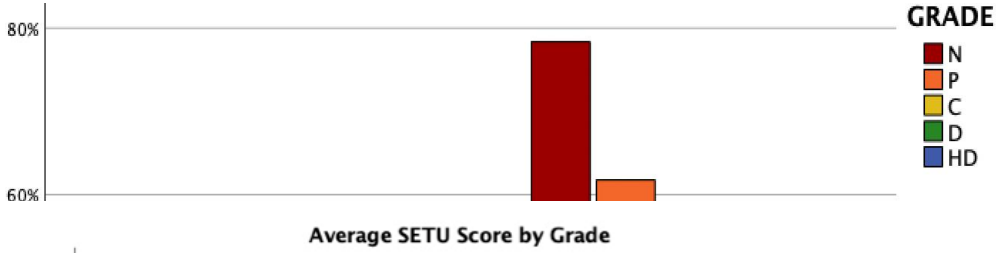
Do you think students who fail a unit are more likely to complete SETU than others and give 'revenge reviews'?

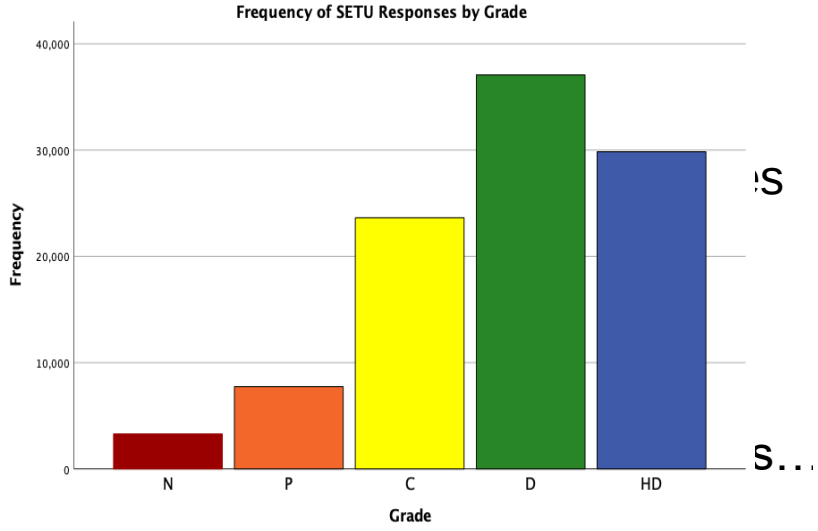
- ☐ Yes
- ☐ No
- ☐ Can't guess

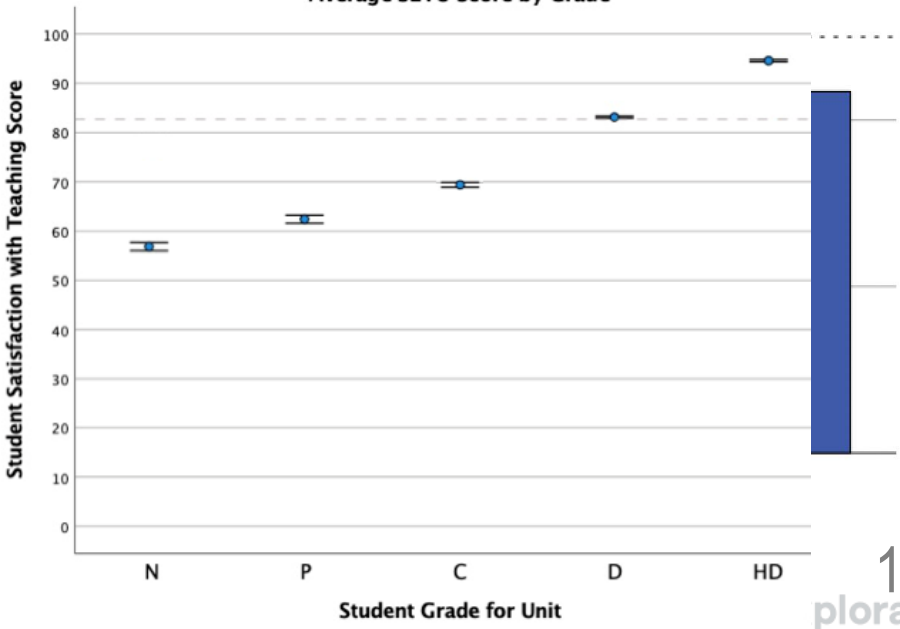
Please use Bluepulse to share your thoughts…

Are students who fail more likely to complete SETU and give *revenge reviews*?

Students who fail (i.e., get an 'N') are far less likely to complete SETU than other students (*nearly 80% don't give SETU feedback).* So even if very negative, they only account for 3.2% of responses.







Are students who fail more likely to complete SETU and give *revenge reviews*?

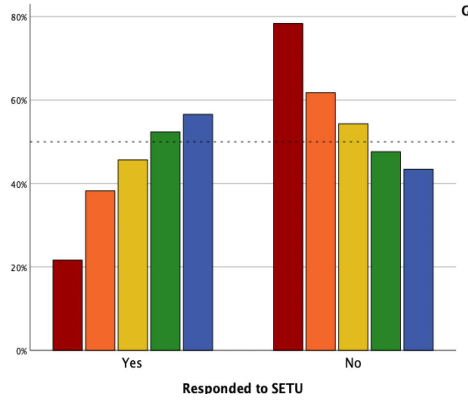
Students who fail (i.e., get an 'N') are far less likely to complete SETU than other students (*nearly 80% don't give SETU feedback*). So even if very negative, they only account for 3.2% of responses.

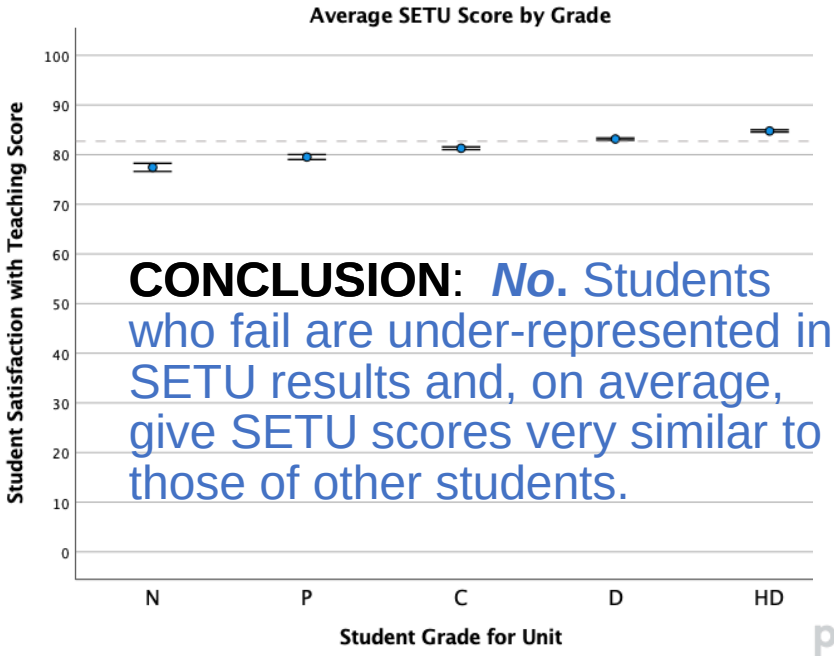
But there **is** a very statistically significant difference in SETU scores given by students with different grades (one way ANOVA)
($F_{4,101752}$=192.000, p<.001, η^2 = .007)

This is a tiny effect size

So we might expect a chart of SETU scores by Grade to looks a bit like this…

…but it actually looks like this





CONCLUSION: ***No.*** Students who fail are under-represented in SETU results and, on average, give SETU scores very similar to those of other students.

Pulse question

Do you think educators in larger units are rated lower than those in smaller units?

- ☐ Yes
- ☐ No
- ☐ Can't guess

Please use Bluepulse to share your thoughts…

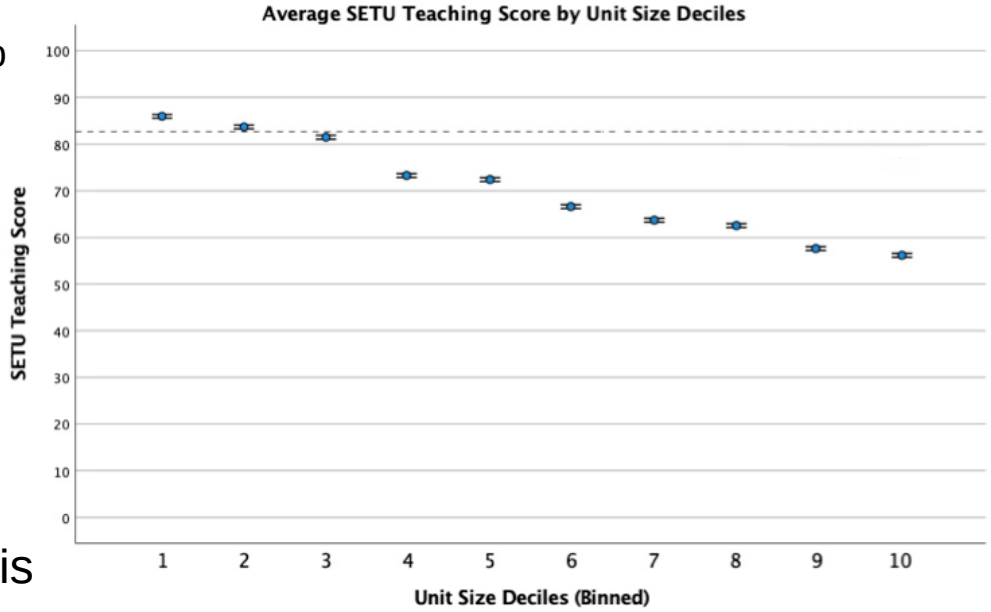
Are educator in larger units rated lower than those in small units?

Unit sizes varied enormously, from 1 to 1,419 (M=263.5; SD=255.0) so they were divided into groups of ~10% each (deciles).

If Unit Size has a large effect we might expect a chart like this:

ANOVA tests shows a significant result, but only a small effect size:
($F_{3,101753}$=279.014, p<.001, η^2 = .011)

…and actualy looks like this



CONCLUSION: ***Not really***. Further analysis shows the small effect size for most Faculties ***but not all***, so it may be due to teaching strategy, not size.

Pulse question

Do you think units with exams receive lower SETU scores than units without exams?

- ☐ Yes
- ☐ No
- ☐ Can't guess

Please use Bluepulse to share your thoughts…

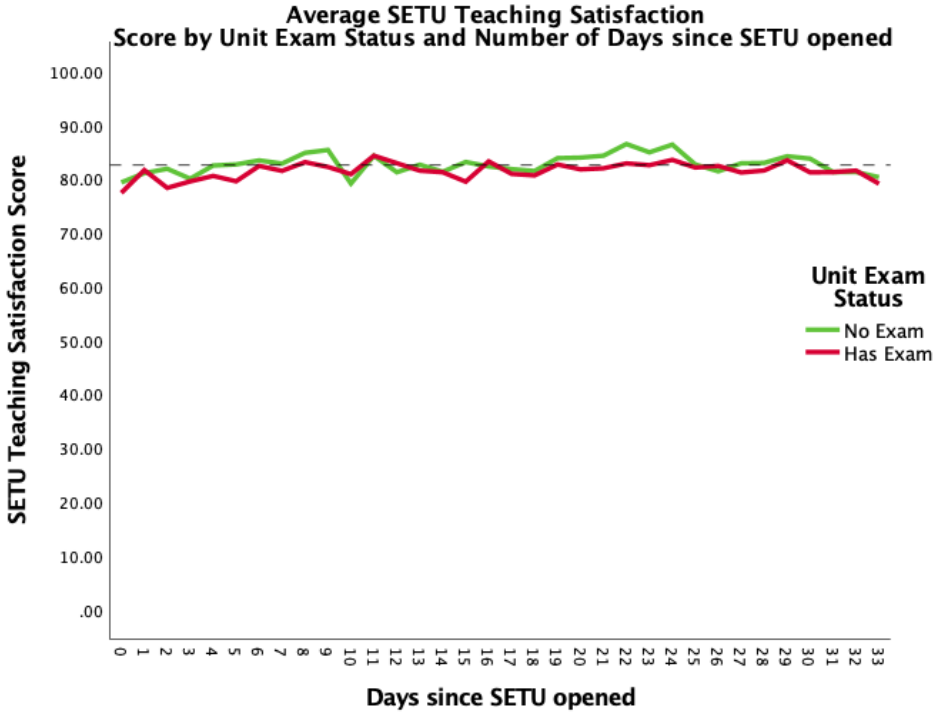
Are educators in units with exams rated lower than those without exams?

Academics are concerned that exams during SETU impact students' SETU Teaching Satisfaction scores.

A statistically significant difference exists between units with & without exams
[$t(112073)=10.82$, $p<.001$, Cohen's $d=.06$]
…but with a trivial effect size

In terms of timing, SETU feedback was open for 5 weeks, and the exam period ran during (roughly) the last 2 weeks of data collection

As is apparent in the line chart, there is no obvious difference between average SETU scores throughout the 5 weeks irrespective of whether exams were running or not and whether the units had exams or not.



CONCLUSION: ***No.*** There is a negligible effect size between units *with and without* exams, and no visible difference in the exam period. Exams have no impact.

Pulse question

Do you think sexism and/or racism have a significant impact on SETU results?

- ☐ Yes
- ☐ No
- ☐ Can't guess

Please use Bluepulse to share your thoughts…

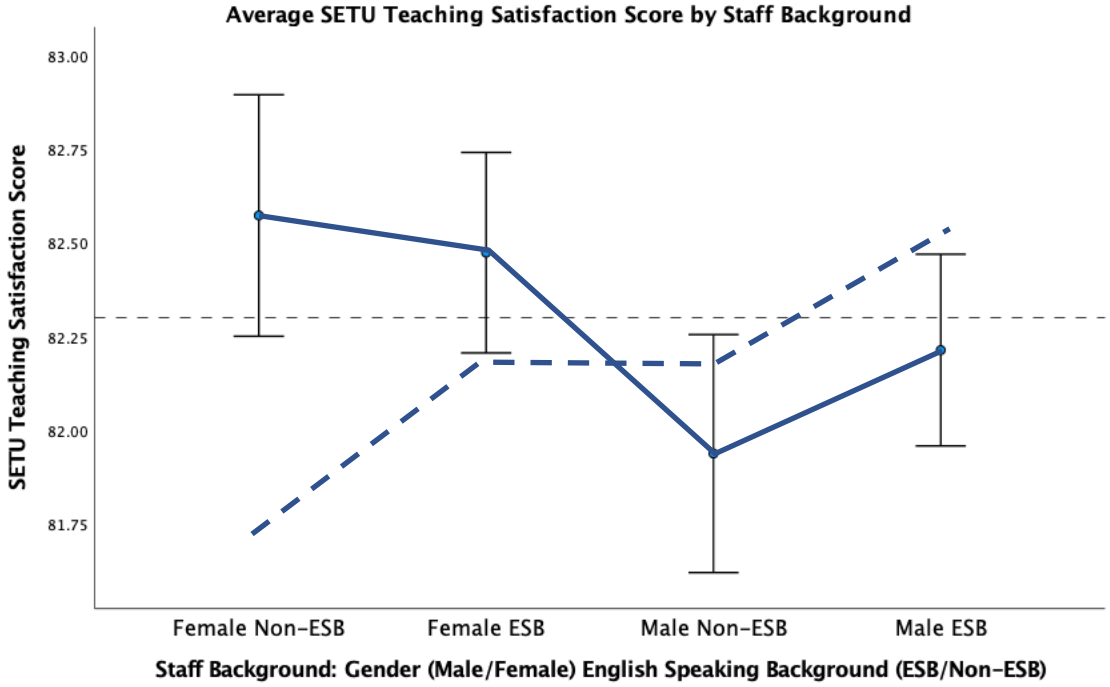
Are SETU teaching scores impacted by ‘racist’ and ‘sexist’ attitudes?

Many studies report concerns about bigotry influencing SETU. We could expect this to manifest as **lower** Teaching Satisfaction **scores** for **Women** and for educators from **Non-English Speaking Backgrounds** (as a proxy for race)

If racism dominates then NESB educators would be rated lower…

…if sexism dominates then Female staff would be rated lower.

At Monash, there is a significant difference ($F_{3,101753} = 3.401$, $p<.017$, $\eta^2 = .000$) but no detectable effect size.



CONCLUSION: ***No.*** At Monash, students do not differ in how they rate staff across language background or genders.
Even zooming in yields unexpected results…

Conclusion

Fortunately, these data show that many of the biases identified elsewhere are not playing out systematically at Monash

So academics can be confident that, overall, SETU is not plagued by bigotry or other biases – indeed students are overwhelmingly positive!

However, we have to remain vigilant, and there can always be local biases which can be detected and corrected

As a result, we have implemented a process of weighting to help adjust for such effects in our SETU reporting, which can perhaps be shared next time

Thank you for your attention and participation!

Any Questions?

REFERENCES

Arnold, I. & Versluis, I. (2019). The influence of cultural values and nationality on student evaluation of teaching, *International Journal of Educational Research, 98*, 13-24.

Andersen, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. PS: Political science & politics, 30(2), 216-219.

Backer, E. (2012). Burnt at the student evaluation stake - the penalty for failing students. *E-Journal of Business Education & Scholarship of Teaching, 6*(1), 1-13.

Badri, M., Abdulla, M., Kamali, M., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management, 20*, 43–59.

Bhandari, P. (2020). What is Effect Size and Why Does It Matter? https://www.scribbr.com/statistics/effect-size/

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

Brenner, P., & DeLamater, J. (2016). Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias, *Social psychology quarterly, 79*(4), *333–354. https://doi.org/10.1177/0190272516628298*

Cohen, J. (1969) *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, 101-105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Lawrence Earlbaum Associates.

Esarey, J., & Valdes, N. (2020). Unbiased, Reliable and Valid Student Evaluations Can Still Be Unfair. Assessment & Evaluation in Higher Education, 45 (8), 1106–1120. doi:10.1080/02602938.2020.1724875.

Hejase, A., Al Kaakour, R., Halawi, L., & Hejase, H. (2013). Students' Perceptions of Student Evaluation of Teaching (SET) Process. International Journal of Social Sciences and Education, 3(3), 565-575. Retrieved from https://commons.erau.edu/publication/307Hofstede, G. (1986). Cultural differences in teaching and learning, *International Journal of Intercultural Relations, 10(3), 301-320.*

Katrompas, A., & Metsis, V. (2021). Rate My Professors: A study of bias and inaccuracies in anonymous self-reporting. *2021 2nd International Conference on Computing and Data Science (CDS)*, 536-542, doi: 10.1109/CDS52072.2021.00098.

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do. NWSA Journal, 19(3), 87–104. http://www.jstor.org/stable/40071230

Miles, P. & House, D. (2015). The Tail Wagging the Dog: An overdue examination of student teaching evaluations. *International Journal of Higher Education, 4*(2), 116-126. http://dx.doi.org/10.5430/ijhe.v4n2p116

Rap, R., & Paxton, P. (2021). How Accurate Are Self-reports of Voluntary Association Memberships? *Sociological Methods & Research, 50*(2), 866–900. https://doi.org/10.1177/0049124118799384

Sigurdardottir, M. Rafnsdottir, G., Jónsdóttir, A., & Kristofersson, D. (2022). Student evaluation of teaching: Gender bias in a country at the forefront of gender equality. Higher Education Research & Development, DOI: 10.1080/07294360.2022.2087604

Sullivan, D., Lakeman, R., Massey, D., Nasrawi, D., Tower, M. & Lee, M. (2023). Student motivations, perceptions and opinions of participating in student evaluation of teaching surveys: a scoping review, *Assessment & Evaluation in Higher Education*, DOI: 10.1080/02602938.2023.2199486

Wagenaar, T. (1995). Student Evaluation of Teaching: Some cautions and suggestions, Teaching Sociology, 23(1), 64-68. https://doi.org/10.2307/1319382