BMC Medical Education BMC

BMC Med Educ. 2025 Feb 26;25:313. doi: 10.1186/s12909-025-06896-3

# Student evaluations of teaching do not reflect student learning: an observational study

R O Gilbert[1,✉], D R Gilbert[2]

Author information Article notes Copyright and License information

PMCID: PMC11863847 PMID: 40012037

## Abstract

### Background

Student Evaluations of Teaching (SET), of individual instructors and of courses, are routinely utilized by university administrators for consequential decisions regarding individual faculty members, courses, and curricula. Despite their ubiquity, much evidence exists that they are biased, amongst other factors by expected and received student grades. To our knowledge, this issue has not been examined in veterinary education until recently. Furthermore, it remains unclear whether observed combinations of higher grades and more favorable student evaluations using common survey instruments reflect enhanced learning. Our study evaluates the relationship between (A) student evaluations of courses in a veterinary curriculum, (B) grades earned in those courses, and (C) an independent measure of learning in those subjects.

### Methods

The Veterinary Educational Assessment (VEA) is an independent, external examination in basic sciences subjects prepared by the National Board of Medical Examiners and administered by the International Council for Veterinary Assessment and is taken by Ross University School of Veterinary Medicine (RUSVM) students in their fifth semester of study. It offers an external means of measuring student learning in specific subjects and relating them to course evaluations. RUSVM has three terms each year with three separate intakes of students. Course evaluations and student grades were recorded for courses from fall 2018 to summer 2022, spanning 12 cohorts of students, and 160 individual courses. Courses were aligned to the relevant section of the VEA taken by each cohort. Spearman correlation coefficients were calculated.

### Results

Mean course evaluations were significantly positively correlated to median grade in the course (*rho* = 0.35, $P < 0.0001$) and the proportion of students earning A-grades (*rho* = 0.38, $P < 0.0001$). The relationship between course evaluation and relevant VEA score was negative (*rho* = -0.18, $P = 0.02$), indicating that students judged courses favorably when higher grades were expected without necessarily learning more from those courses.

### Conclusions

We confirmed the well-known relationship between SET and student grades but, for the first time in veterinary medicine, describe a small but negative and statistically significant relationship between SET and an independent measure of learning. SET should be interpreted with caution; their use for evaluation of teachers or courses may have unintended consequences including reduced expectations for student achievement.

**Keywords:** Student evaluation of teaching, Veterinary education, Veterinary educational assessment

## Background

Student evaluations of teaching (SET) are now a standard tool in post-secondary education to evaluate the teaching effectiveness of faculty. They inform decisions regarding wage increases, promotion and tenure, and inform curriculum development. A survey of four-year, liberal arts colleges in the United States found that 94% “always used” SET in assessing faculty teaching [1]. SET tend to be conducted at the end of a course and ask students to evaluate their instructor(s) as well as the course overall. Students are asked to indicate their level of agreement with statements like “The instructor was available to students outside of class", "Overall, I learned a lot from this instructor", or "The course was well planned", and responses are typically recorded via a Likert scale, with options ranging from "Strongly Disagree" to "Strongly Agree" (or similar). In addition, students may be asked more general questions, such as "Overall, how would you rate this course/instructor?" Again, responses are collected via a Likert scale.

Implicit in the perceived value of SET is the assumption that students' reported perceptions of teaching quality are an accurate measure of actual teaching quality: if SET are to be used as a basis for individual promotion or merit or for decisions regarding courses, they should be a reliable measure of teaching effectiveness. There is a significant body of evidence indicating that this assumption is not justified. Research on SET has suggested the presence of biases according to, amongst other factors, gender, physical attraction, race, ethnicity, age, culture, likability, prior interest in the subject matter, and the perception of leniency in making policy exemptions [2–10]. In addition, and most relevant to our present purposes, studies repeatedly demonstrate a positive correlation between high grades (or expected grades) in courses and SET. This correlation has been interpreted by some as validating SET: higher marks are indicative of increased learning and, therefore, good teaching [9, 11]. On this interpretation, good teaching leads to student learning, which, in turn, leads to increased course grades and higher SET. Following Johnson, one might call this explanation the *teacher-effectiveness theory* [12]. On the other hand, the correlation has been attributed to a biasing effect "in the sense that the effect of grading on teaching evaluations represents a factor not related to either effective teaching or student learning." [12, 13] Though different mechanisms have been posited to explain the precise nature of the bias, the common feature among *grade-bias theories* is that the expectation of higher grades—distinct from high-quality teaching—drives more favorable SET [3, 13–15]. Several recent multi-section validity studies and meta-analyses seem to indicate that the teacher-effectiveness theory is not a good explanation of the observed positive correlation, lending support to the grade-bias theories [3, 4, 14, 16]. Buttressing this argument, there is evidence that instructors who feel vulnerable to unfavorable hiring decisions (e.g. adjunct faculty) face the most intense pressure to achieve high student evaluations—and, indeed, adjunct instructors do tend to give higher grades than full time faculty [17].

To our knowledge, there are few reports on these issues in the context of veterinary education [18–21]. With regard to medical education, the known biases of SET and the difficulties involved in relating student evaluations to actual learning have previously been noted, as has the basic dichotomy of (medical) students rating courses on the basis of their ability to pass examinations whereas their teachers seek more broadly to prepare them to become doctors in the context of a comprehensive curriculum [19]. At least one attempt to relate the extant literature on SET to veterinary education seems to endorse the effectiveness theory [20]. However, the veterinary educational literature has also been interpreted to support the grade-bias theory [15]. It seems that no specific data exist to interrogate this question (grade bias vs. teaching effectiveness) specifically in the context of veterinary education, although there is evidence that receiving or expecting a higher grade drives better course evaluation in a veterinary context [21]. The aim of this paper is to evaluate the relationship between SET for courses in a veterinary curriculum, grades earned in those courses, and independent measures of learning in those subjects. In addition to being focused on veterinary education in particular, the use (and availability) of an independent, external measure of learning distinguishes the current study from others in the literature.

## Methods

This study was reviewed by the IRB of Ross University School of Veterinary Medicine (RUSVM) and deemed to be exempt from review and from individual student consent in accordance with U.S. Federal Regulations 45 CFR 46.102. reference 24 – 03.

RUSVM has three academic terms each year. Intake of new students occurs tri-annually as well, coordinated with the beginning of the academic terms. During the period of this study, entering classes numbered 180 to 200 students. Attrition over the first four semesters was approximately 15%. All students took all courses in the same sequence. Courses in this study employed an arbitrary pass point of 70%. A-grades were allocated for students with 90% or higher.

In the context of this study, all courses were offered in the first four semesters of study. All courses were required (there was no opportunity to select courses) and all were taken in the prescribed sequence. The specific survey tool used or SET had been in use for some time at RUSVM and was uniform for all courses, but was discontinued in 2022 following a study by a task force that pointed out its biases and inconsistencies [18]. Its origin is unknown, and we make no claims about its value or design quality. SET during this period were conducted online at the end of each course. Students had electronic access to the survey instrument from three weeks before the end of term and continuing until two weeks after the end of the term. Therefore, some students would have completed the survey before knowing their final grades and others would have known the final outcome of each course. No final examination was worth more than 40% of the course grade; students completing the surveys prior to the examination therefore had substantial knowledge of their performance in the course. Completion of the survey was not required and responses were anonymous.

The categories/questions according to which students were asked to evaluate courses were:

1. The course was well planned, organized and followed a coherent pattern.

2. Learning activities used in this course [lectures, small groups, demonstrations, laboratories, case discussions, quizzes, etc.] were effective.
3. The course was intellectually challenging and stimulating.
4. Examination questions for this course challenged me to critically think about concepts I was presented.
5. The examinations were administered according to the Student Handbook.
6. The course conformed to the schedule published in the syllabus.
7. The amount of time I invested in this course was appropriate for the allotted credits.
8. Rate the course overall.

Each question was answered using a 5-point Likert scale.

We explored the relationships between component questions as well as those between the questions and results on the Veterinary Educational Assessment (VEA) in order to probe the question of teacher-effectiveness versus grade bias.

The VEA is an independent, standardized, external examination in basic sciences subjects composed by the National Board of Medical Examiners and administered by the International Council for Veterinary Assessment and is taken by RUSVM students in their fifth semester of study. Students take a 240-item, web-based, multiple-choice examination with sections on anatomy, physiology, pharmacology, microbiology, and pathology. The interval between specific courses and the VEA varied from one to four semesters. Students were aware of the scheduling of the VEA, which is always administered in the first week of the fifth semester. They were urged to take it seriously although it was a low stakes exam for them and did not affect their academic standing. Although its principal use was to benchmark curricular efficacy, results are significantly correlated with performance in the North American Veterinary Licensing Examination [22].

Courses were aligned with components of the VEA as depicted in Table 1 below:

Table 1.

Alignment of courses taught and VEA sections

| VEA Section | Course Subject titles |
|---|---|
| Anatomy | Gross Anatomy I, Gross Anatomy II, Microscopic Anatomy & Embryology |
| Physiology | Physiology I, Physiology II |
| Microbiology | Principles of Infectious Diseases, Immunology, Parasitology, Bacteriology, Virology, Veterinary Public Health |
| Pathology | Pathology I, Pathology II, Clinical Pathology |
| Pharmacology | Pharmacology |

Open in a new tab

We recorded SET, course grades, and the percentage of students with grades in the A-range or F-range for 160 individual courses between the fall 2018 and summer 2022 terms: the first four terms of study for 12 cohorts of students. Each course was aligned with the relevant section of the VEA taken by each cohort. To further explore the relationship between high course grades and SET, we calculated modified Hofstee cut points for each course [23].

Mean course evaluations and final course grades were not normally distributed. Therefore, Spearman correlation coefficients were calculated to explore the relationships between the median grade for the course, the mean SET, the percentage of A-grades, the percentage of F-grades, and VEA score. We also explored the relationship between the individual questions in the SET.

We had available only mean scores for overall course evaluations and for components of the evaluation. Original raw data are unfortunately not available. We acknowledge that it is statistically inappropriate to use means values for Likert scale scores [24], but contend that it is or was common practice to use SET means in making administrative decisions [16, 24]. Had we been able to, we would have used interpolated medians rather than means for the SET responses [24].

## Results

Mean course evaluations were significantly positively related to median grades in the courses (rho = 0.35, $P < 0.0001$). In addition, they were significantly positively correlated with the proportion of students earning final grades in the A-range (rho = 0.38, $P < 0.0001$). On the other hand, the relationship between course evaluations and the corresponding sectional VEA scores was negative (rho = -0.18, $P = 0.025$), as was the correlation between the VEA scores and the percentage of grades in the A-range (rho = -0.31, $P = 0.0001$). These data are summarized in Table 2.

Table 2.

Relationship of response to “rate the course overall” to course median grade and VEA scores. (Cells contain Spearman’s correlation coefficient (rho), *P*-value.) for all correlations in this table, $n = 160$

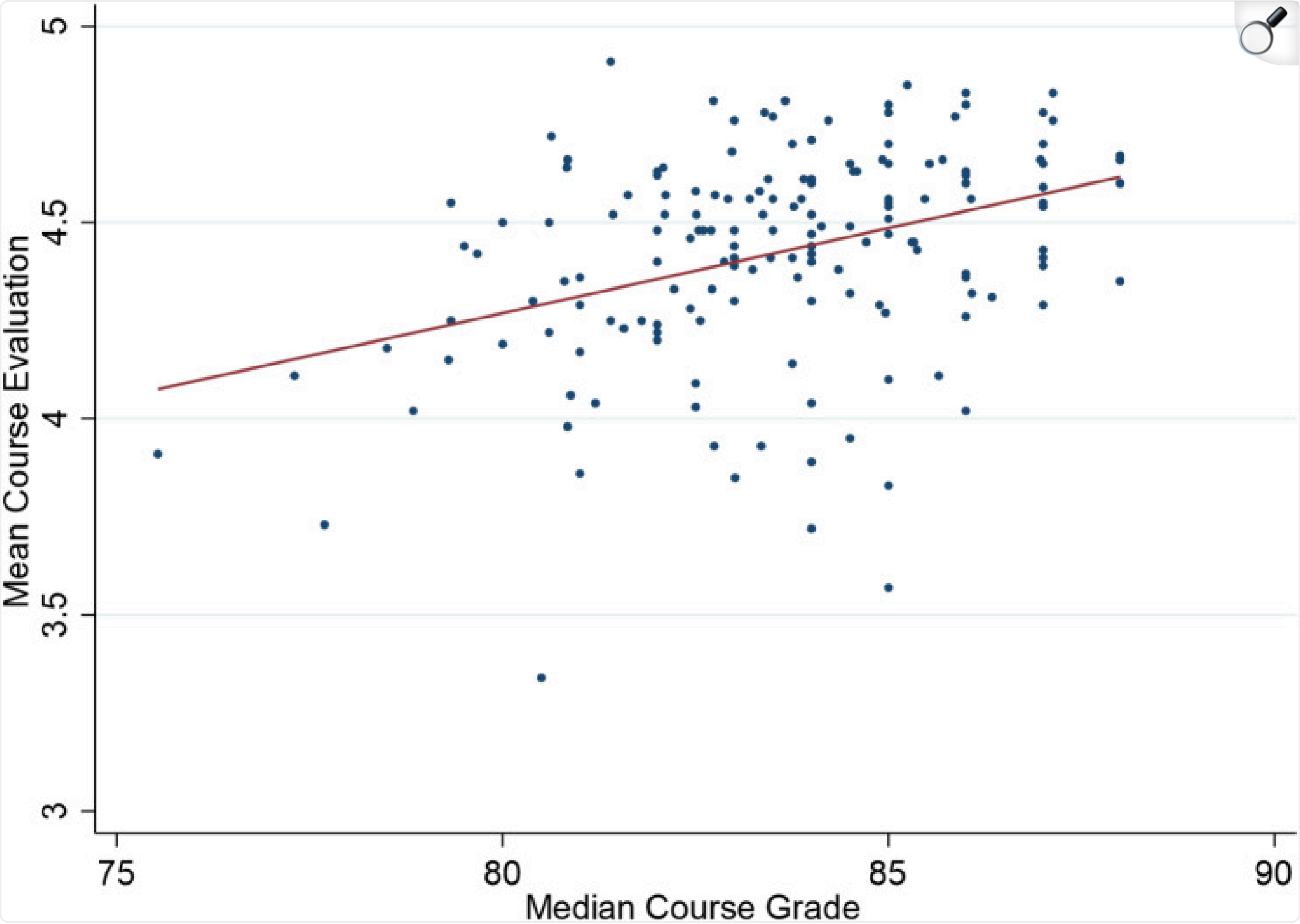
| | **Overall evaluation** | **Median Grade** | **Percent A-grades** | **Percent F-grades** | **Modified Hofstee cut point** | **VEA score** |
|---|---|---|---|---|---|---|
| Overall evaluation | 1 | | | | | |
| Median Grade | 0.35<br>$P < 0.0001$ | 1 | | | | |
| Percent A-grades | 0.38<br>$P < 0.0001$ | 0.82<br>$P < 0.0001$ | 1 | | | |
| Percent F-grades | -0.14<br>$P = 0.078$ | -0.51<br>$P < 0.0001$ | -0.18<br>$P = 0.023$ | 1 | | |
| Modified Hofstee cut point | 0.22<br>$P = 0.006$ | 0.78<br>$P < 0.0001$ | 0.41<br>$P < 0.0001$ | -0.77<br>$P < 0.0001$ | 1 | |
| VEA Score | -0.18<br>$P = 0.025$ | -0.19<br>$P = 0.017$ | -0.31<br>$P = 0.0001$ | Not significant | Not significant | 1 |

Open in a new tab

A modified Hofstee method was used to determine an “objective” passing score as an alternative to the arbitrary 70% pass point used in reality during the period of this study. Not surprisingly, the calculated cut score is highly correlated to the median grades for each course (*rho* = 0.78, $P < 0.0001$), and to the percentage of students awarded grades in the A-range (*rho* = 0.41, $P < 0.0001$) because A grades were similarly allocated arbitrarily to students awarded more than 90%. The calculated cut point was even more strongly (and negatively) related to the proportion of students failing each course, because of the clustering of grades around the cut point (*rho* = -0.77, $P < 0.0001$). The percentage of students passing courses with arbitrary cut points was correlated to the median course evaluation (*rho* = 0.18, $P = 0.017$) but had the modified Hofstee cut point been used, this relationship became non-significant (*rho* = 0.06, $P = 0.41$). Similarly, the percentage of students awarded A-grades using the arbitrary 90% standard was significantly correlated to mean course evaluation (*rho* = 0.22, $P = 0.005$) but if grades had been corrected according to the modified Hofstee, this relationship also became non significant (*rho* = 0.09, $P = 0.27$).

The positive relationship between median course grades and mean course evaluations are shown in Fig. 1. Figure 2 shows the scatter plot and trend line for the negative relationship between mean course evaluations and VEA scores.

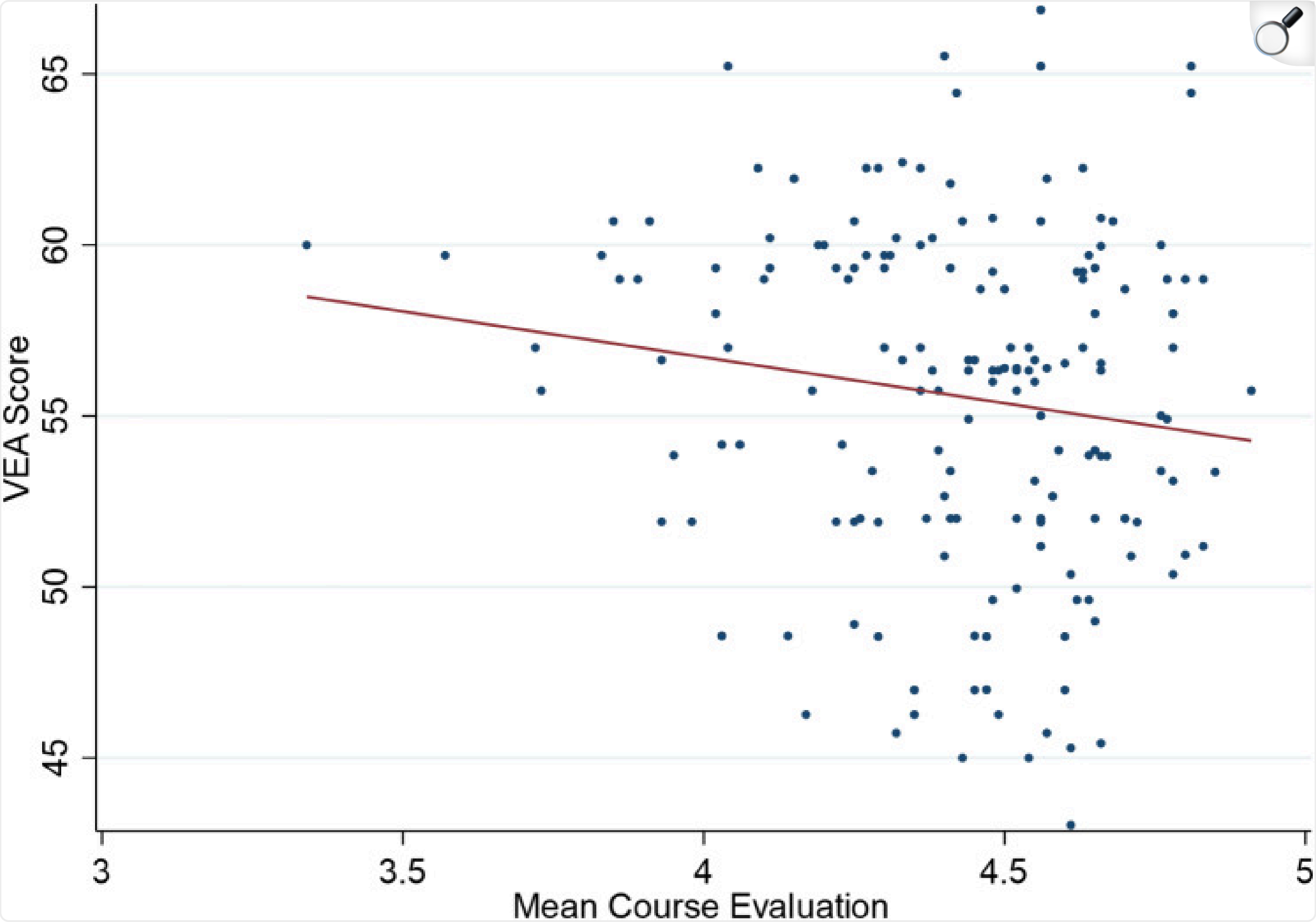
Fig. 1.



Open in a new tab

Scatter plot and regression line showing relationship between median course grade and mean course evaluation ($P = 0.0001$)

Fig. 2.



Open in a new tab

Scatter plot and regression line showing relationship between mean course evaluation and corresponding VEA scores ($P = 0.025$)

Notably, responses to all component questions of the student evaluation were significantly correlated to each other. Spearman correlation coefficients ranged from 0.63 to 0.95, and for all $P < 0.0001$. This might suggest a mindless completion of the evaluation once a student had formed an overall opinion of the course ("halo / horn effect") [25]. Responses to all individual questions were either unrelated to VEA scores ($P > 0.05$) or negatively correlated. All were significantly and positively correlated with course median grade and with the proportion of the class earning grades in the A-range. Even responses to the question "of whether or not courses adhered to the published syllabus" also varied with the median grade—even though essentially no courses deviated at all from the syllabus and, if they had, it would have constituted a basis for grade appeals and formal complaints.

## Discussion

Our study confirms the well-established grade bias to which SET is subject. For the first time in a veterinary educational context, the negative relationship between SET and actual learning is illustrated.

A novel aspect of the study is that we were able to take advantage of the VEA as an external and independent measure of learning to explore the relationship between SET and learning. While the negative relationships between SET and VEA scores and between VEA scores and course grades are not strong, they are statistically significant. A weak, statistically significant correlation implies that the relationship was unlikely to have been observed by chance alone, but that other factors are involved in the outcome, which is not surprising. The existence of significant negative correlations in this case, even if numerically weak, undermines arguments that favorable SET (and high course grades) generally reflect superior teaching and, by implication, learning. Hence, these significant negative correlations provide reason to prefer grade-bias theories over teacher-effectiveness theories as explanations of the relationship between course grades and evaluations. In turn, this necessitates extreme caution in the use of SET for informing important, career or curriculum-defining decisions such as faculty remuneration, promotion or tenure, course revision, or grading policies. The additional perspective provided by examining the theoretical difference that using modified Hofstee cut points would have provided casts additional light on the negative relationship between course evaluations and VEA scores, and course grades and VEA scores, because it suggests that, in general, courses with very high grades, had indefensibly high grades. Stated differently, for these courses, several students scored higher grades than they deserved relative to a calculated pass point, or passed a course they arguably should not have. Our results may reflect a poorly designed course evaluation survey instrument. However, the instrument is not unlike others in common use and few, if any, reports describe instruments free of bias [26]. Given the reported weakness of SET as measures of teaching, it may be desirable to use objective measures of learning as a proxy for determining effectiveness of teaching, such as pre- and post-course testing [26]. Pre- and post-course testing suffers from the weakness of being susceptible to instructors teaching specifically to this assessment, making an external, objective instrument such as the VEA even more desirable when such a measure is available.

The use of the VEA is appealing for many reasons. Apart from the fact that it is external and independent of any veterinary school, it is a carefully curated examination, prepared by the National Board of Medical Examiners, and administered by the International Council for Veterinary Assessment, specifically to benchmark institutional programs and individual student progress toward qualification as a veterinarian licensed to practice in North America. Because the VEA is a low-stakes examination for students and is used primarily to benchmark institutional programs, as explained above, there is little incentive to do well, which may distort our study. On the other hand, students are made aware of the benefits to them of taking the VEA seriously as a measure of their own preparedness for the stage of the curriculum. They are reminded that there is a good correlation between the VEA and student scores in the North American Veterinary Licensing Examination (NAVLE) [24], which is of obvious importance to them, and exhorted to take the VEA seriously. Our own data (unpublished) confirm that VEA scores are well correlated (*rho* = 0.68, $P < 0.0001$) to individual student performance in the NAVLE. These observations suggest that, overall, VEA scores are indicative of student mastery of material, even though this is not a high stakes examination required for our students to progress in the program.

Importantly, with regard to the design of this study, instructors of the included courses had control over the content, pace, methods of assessment, and course design more generally. Hence, our study is not obviously subject to a criticism that has been made of multi-section validity studies: that wresting control of courses from instructors and placing it in the hands of researchers conducting studies undermines the quality of the courses, obfuscates the relevance of any observed correspondence between evaluations and course grades, and reduces the scope of any conclusions [12].

The consistent correlation between component questions of the SET, and their relationship to grades earned in courses, supports the contention that students seem prone to complete surveys without attention to detail, once having decided their overall opinion of the course (the halo/horns effect) [27]. An extreme instance of lack of diligence in completion of course evaluations is a study by Uijtdehaage and O'Neal in which 66% of students completed assessments for a fictitious instructor and 49% did so even when a photograph of the phantom teacher was included [28].

## Limitations

There are a number of limitations to this study that require consideration. First the lack of a one-to-one relationship between subjects taught and VEA sections is a weakness of this study, but we believe that the availability of an external, objective measure that cannot be manipulated by the instructors or the institution provides a unique opportunity to study student evaluation of teaching and actual learning, offsetting this acknowledged deficit.

Secondly, in order to interpret the negative correlation between SET/grades and performance on the external examination as not supporting teacher-effectiveness theories, one needs to assume that: (*i*) student learning is an acceptable measure of teaching effectiveness (or a significant component thereof); and (*ii*) measuring performance on the external examination is a reasonable proxy for assessing learning. Proponents of teacher-effectiveness theories might well grant (*i*) but argue, against (*ii*), that course grades are a better indicator of learning than a standardized examination. However, recent studies exploring and modeling the mechanisms relating SET to grade inflation cast some doubt against the validity of such an argument [29]. Our study cannot settle this debate.

Third, our study allowed SET completion during a period that spanned the final examinations and course results, so some students completed the evaluations without knowing their final scores. However, for all of the courses, the weight of the final examination was 40% or less, with the rest of the points allocated for assessments during the term. The result is that students had a good basis for estimating their final course grades even before the final examination. We therefore believe our conclusion that student evaluation of courses was biased by actual or expected grades. This is supported by the work of Bailey et al., who reported that completion of SET surveys after final examinations lowered the response rate but did not influence actual ratings [21]. Bailey et al. also confirmed that students' self-reported expected grades were positively correlated with all SET items in their survey, observations that lessen concern over the timing of the evaluations in our study and support the observation of consistent grade bias.

Lastly, as noted above, using interpolated medians rather than means for the SET responses would have been statistically more appropriate. Unfortunately these were not available to us because when the specific survey instrument was discontinued, the computer platform housing it was also decommissioned; that led to loss of the original raw data - only summaries of SET responses remained for our investigation.

## Conclusion

Student evaluations of teaching should be interpreted very cautiously; their use for evaluation of teachers or courses may be counter-productive and penalize some of the most effective teachers and provide unintended incentives for teachers to lower expectations of student mastery of material. In particular, there is evidence that inflated grades may be associated with more favorable SET scores, but that student mastery of subjects could be less than the high grades (and associated SETs) may imply.

## Acknowledgements

Not applicable.

## Abbreviations

**IRB**
: Institutional review board

**RUSVM**
: Ross university school of veterinary medicine

**NAVLE**
: North American veterinary licensing examination

**SET**
: Student evaluation of teaching

**VEA**
: Veterinary educational assessment

## Author contributions

Study conception and design: ROG and DRG; data acquisition: ROG; analysis of results: ROG; draft manuscript preparation and manuscript revision: ROG and DRG. Both authors reviewed the results and approved the final version of this manuscript. Both authors agree to be accountable for all aspects of the work.

## Funding

None.

## Data availability

Data are available from the corresponding author upon reasonable request.

## Declarations

### Ethics approval and consent to participate

This study was reviewed by the IRB of Ross University School of Veterinary Medicine and deemed to be exempt from review and from individual student consent in terms of U.S. Federal Regulations 45 CFR 46.102.

### Consent for publication

Not applicable.

## Competing interests

The authors declare no competing interests.

## Footnotes

### Publisher’s note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

1. Miller J, Seldin P. Changing practices in faculty evaluations: can better evaluation make a difference? American Association of University Professors; 2014. https://www.aaup.org/article/changing-practices-faculty-evaluation . [cited 2023 November 22].

2. Boring A. Gender biases in student evaluations of teaching. J Public Econ. 2017;145:27–41. [Google Scholar ]

3. Boring A, Ottoboni K, Stark P. Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. ScienceOpen research. 2016.

4. Fan Y, Shepherd LJ, Slavich E, Waters D, Stone M, Abel R, et al. Gender and cultural bias in student evaluations: why representation matters. PLoS ONE. 2019;14(2):e0209749. [DOI ] [PMC free article] [PubMed] [Google Scholar ]

5. Hamermesh DS, Parker A. Beauty in the classroom: instructors pulchritude and putative pedagogical productivity. Econ Educ Rev. 2005;24(4):369–76. [Google Scholar ]

6. Wolbring T, Riordan P. How beauty works. Theoretical mechanisms and two empirical applications on students’ evaluation of teaching. Soc Sci Res. 2016;57:253–72. [DOI ] [PubMed] [Google Scholar ]

7. Smith BP. Student ratings of teaching effectiveness: an analysis of end-of-course faculty evaluations. Coll Student J. 2007;41(4):788. [Google Scholar ]

8. Joye S, Wilson JH. Professor age and gender affect student perceptions and grades. J Scholarsh Teach Learn. 2015;15(4):126–38. [Google Scholar ]

9. Feistauer D, Richter T. Validity of students' evaluations of teaching: biasing effects of likability and prior subject interest. Stud Educational Evaluation. 2018;59:168–78. [Google Scholar ]

10. Boswell SS. Academic entitlement and Ratemyprofessors.com evaluations bias student teaching evaluations: implications for faculty evaluation and policy-lenient professors' occupational health. Heliyon. 2024;10(8):e29473. [DOI ] [PMC free article] [PubMed] [Google Scholar ]

11. Cohen PA. Student ratings of instruction and student achievement: A Meta-analysis of multisection validity studies. Rev Educ Res. 1981;51(3):281–309. [Google Scholar ]

12. Johnson VE, Springer LEIC, Ebook C. SpringerLink. Grade inflation: a crisis in college education. New York: Springer; 2003. [Google Scholar ]

13. Clayson DE, Frost TF, Sheffet MJ. Grades and the student evaluation of instruction: A test of the reciprocity effect. Acad Manage Learn Educ. 2006;5(1):52–65. [Google Scholar ]

14. Clayson DE. Student evaluations of teaching: are they related to what students learn?? A Meta-Analysis and review of the literature. J Mark Educ. 2009;31(1):16–30. [Google Scholar ]

15. Anderson RE, Choi KS, Hair JF. Cognitive consistency theory and student evaluation of teacher effectiveness. J Experimental Educ. 1975;44(2):64–70. [Google Scholar ]

16. Uttl B, White CA, Gonzalez DW. Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. Stud Educational Evaluation. 2017;54:22–42. [Google Scholar ]

17. Sonner BS. A is for adjunct: examining grade inflation in higher education. J Educ Bus. 2000;76(1):5–8. [Google Scholar ]

18. Rolph KE, King A, Larde H, Kim ST, Ragland N, Herve Claude LP et al. Critique of approaches for evaluating teaching and proposal for a new institutional policy. J Vet Med Educ. 2023;50(5):499–507, e20220072. 10.3138/jvme-2022-0072. [DOI ] [PubMed]

19. Looi JCL, Anderson KJ. Between SET and ASP: balancing the scales of student evaluation of teaching (SET) and teachers' assessments of student performance (ASP) for medical school

education in psychiatry. Australasian Psychiatry: Bull Royal Australian New Z Coll Psychiatrists. 2018;26(6):659–61. [DOI] [PubMed] [Google Scholar]

20. Beran TN, Donnon T, Hecker K. A review of student evaluation of teaching: applications to veterinary medical education. J Vet Med Educ. 2012;39(1):71–8. [DOI] [PubMed] [Google Scholar]

21. Bailey MR, Lane IF, Biddix JP. Veterinary student evaluations of teaching: scores and response rate when administered before or after final exams. J Vet Med Educ. 2023;51(6):785–94, e20230128. 10.3138/jvme-2023-0128. [DOI] [PubMed]

22. Danielson JA, Wu TF, Molgaard LK, Preast VA. Relationships among common measures of student performance and scores on the North American veterinary licensing examination. J Am Vet Med Assoc. 2011;238(4):454–61. [DOI] [PubMed] [Google Scholar]

23. Burr SA, Whittle J, Fairclough LC, Coombes L, Todd I. Modifying hofstee standard setting for assessments that vary in difficulty, and to determine boundaries for different levels of achievement. BMC Med Educ. 2016;16:34. [DOI] [PMC free article] [PubMed] [Google Scholar]

24. Zumrawi AA, Macfadyen LP. Proposed metrics for summarizing student evaluation of teaching data from balanced likert scale surveys. Cogent Educ. 2023. 10.1080/2331186X.2023.2254665.

25. Noor N, Beram S, Yuet FKC, Gengatharon K, Rasidi MSM. Bias, halo effect and Horn effect: a systematic leterature review. Int J Acad Res Bus Social Sci. 2023;13(3):1116–40. [Google Scholar]

26. Mendoza Diaz NV, Sotomayor T. Effective teaching in computational thinking: A bias-free alternative to the exclusive use of students' evaluations of teaching (SETs). Heliyon. 2023;9(8):e18997. [DOI] [PMC free article] [PubMed] [Google Scholar]

27. Shevlin M, Banyard P, Davies M, Griffiths M. The validity of student evaluation of teaching in higher education: love me, love my lectures? Assess Evaluation High Educ. 2000;25(4):397–405. [Google Scholar]

28. Uijtdehaage S, O'Neal C. A curious case of the Phantom professor: mindless teaching evaluations by medical students. Med Educ. 2015;49(9):928–32. [DOI] [PubMed] [Google Scholar]

29. Stroebe W. Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. Basic Appl Soc Psychol. 2020;42(4):276–94. [Google Scholar ]

## Associated Data

*This section collects any data citations, data availability statements, or supplementary materials included in this article.*

### Data Availability Statement

Data are available from the corresponding author upon reasonable request.

Articles from BMC Medical Education are provided here courtesy of **BMC**