2020-12-01

Relationship between Course Evaluations and Course Grades in Six Allied Health Programs

Emily Weidman-Evans
LSU Health Shreveport, School of Allied Health Professions, LA, USA

Sandra Hayes
LSU Health Shreveport, School of Allied Health Professions, LA, USA

Teresa Bigler
LSU Health Shreveport, School of Allied Health Professions, LA, USA

Follow this and additional works at: https://hpe.researchcommons.org/journal

Relationship between Course Evaluations and Course Grades in Six Allied Health Programs

Emily Weidman-Evans*, Sandra Hayes, Teresa Bigler

LSU Health Shreveport, School of Allied Health Professions, 1501 Kings Highway, Shreveport, LA 71130-3932, United States

Abstract

Purpose: Most faculty members are evaluated by their administration, in part, based upon student evaluation results. This study sought to determine the relationship between course evaluation scores and course grades, in order to ensure academic rigor and optimize faculty evaluation practices.

Methods: Overall course evaluation results were compared to average course grades for six allied health programs, over 3 years (n = 256 observations).

Results: There was no overall relationship between evaluation scores and course grades; however, individual differences between programs were seen.

Conclusion: Grades achieved in a course did not affect the overall course evaluations. However, since there was a correlation between grades earned and course evaluations for some programs in the School of Allied Health, the importance of utilizing multiple assessment tools to determine the effectiveness of a course must be stressed.

© 2020 King Saud bin Abdulaziz University for Health Sciences. Production and Hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Course evaluations; Course assessment; Faculty evaluation; Effectiveness; Quality assurance; Allied health

1. Introduction

Most faculty members are evaluated by their administration, in part, based upon student evaluation results. In fact, they are claimed to be the most widely used tool in the overall evaluation of teaching in higher education.[1] Proponents state that they are a valuable tool,[2] reliable and quantitative measures of teaching effectiveness, while they are giving students a voice.[1] Critics, however, state that they are "meaningless quantification", and give students "the power of their pencils to get even". This topic has become so hotly debated that some institutions are actually prohibiting their use in promotion and tenure decisions.[3,4]

There is a perception of a relationship between faculty evaluations and student grades in higher education. If grades are a valid marker of student learning, this is appropriate. However, if students are evaluating courses and instructors based not on what they learned but on only their desired grade, this relationship can lead (intentionally or not) to grade inflation and "teaching down", due to the importance placed on the

---

* Corresponding author.

E-mail addresses: eevans@lsuhsc.edu (E. Weidman-Evans), shayes1@lsuhsc.edu (S. Hayes), tbigle@lsuhsc.edu (T. Bigler).

evaluations by some institutions, as posited by Stroebe.[5] This possible "negative" relationship is supported by research conducted by Boring, Ottoboni, and Stark who investigated the relationship between evaluations, student grade expectations and final grades. No significant correlation was found between student evaluation results and final grades but there was a significant positive correlation between evaluations and expected grades.[6]

In spite of the legitimate concern that potentially affects both academic rigor and teaching outcomes, there has been very little research completed regarding student evaluations of teaching in healthcare education. One study by Bowen and colleagues showed that medical students' evaluations of their general surgery clerkship correlated positively with student grades.[7]

In a pilot study (IRB approved: STUDY00000915), the faculty in the School of Allied Health Professions (SAHP) Physicians Assistant (PA) Program analyzed the relationship between course evaluation results and final grades for the purpose of quality assurance within the PA program and to potentially shape the SAHP faculty evaluation practices. Forty-three course offerings ("observations") were included in the analysis. Results of a single linear regression analysis found that there was a significant relationship between course grades and evaluation scores in this program (p = 0.01; 95% CI 0.11, 0.94); however, the r^2 was low-medium at 0.37. Due to this discordance between the presence of a statistically significant relationship and wellness-of-fit of the data, an additional analysis was performed based upon nominal data. This showed that courses with a mean grade of $\geq$ 90% were 11 times more likely to have a median evaluation score of 5 (the highest possible).

Based upon these findings and the paucity of information in the current health education literature, the research was expanded to include each of the other five SAHP programs at this institution. These include cardiopulmonary sciences (CPS), clinical laboratory sciences (CLS), occupational therapy (OT), physical therapy (PT), and speech-language pathology (SLP). Of the six programs, two of these offer a bachelor's degree (CPS, CLS); three offer a master's degree (OT, PA, SLP), and one offers a clinical doctorate (PT). The number of students in each cohort ranges from five (CPS) to 39 (PA).

The purpose of this study was primarily to determine the relationship between course evaluation scores and course grades, while the secondary purpose was to determine the relationship between course evaluation scores and course credit hours (a surrogate marker for course difficulty, as shown by Szafran and colleagues).[8] An additional goal was to contribute to the health education literature in this arena to ensure academic rigor by potentially minimizing grade inflation and to optimize faculty evaluation practices.

2. Materials & methods

2.1. Data collection

This was a retrospective observational study. At the end of every semester, prior to final grades being submitted, students anonymously complete their course and instructor evaluations electronically. All programs' course evaluations utilize a 5-point Likert-type scale, with 5 being the most favorable or highest score and 1 being the least favorable. Evaluations are completed electronically, and all programs have at least a 90% completion rate. Upon IRB approval of this project, each Program Director submitted data to the primary investigator for approximately three cohorts (years) of courses. The data submitted for each course from the previous three to four years included:

- Mean course evaluation score: The "course evaluation score" that was used for analysis was the one that related to either the first or last "general" question on the evaluation tool (e.g., "Rate the course overall.")
- Average course grade: The course grade was calculated on a 4.0 scale, based upon the number of students who made As (4.0), Bs, (3.0), Cs (2.0), or Ds (1.0) in a course. (Fs are not allowed in any of the courses in the School of Allied Health for progression.)
- Number of credit hours: Credit hours were recorded based upon course catalog entries.

For consistency, the investigators compiled all useable data from the first 3−4 semesters of didactic work, regardless of program length and input this into an Excel spreadsheet for comparison. "Useable data" was defined as having all three necessary data points—course grade, evaluation score, and credit hours. Each course offering was considered an "observation". (For example, the course "PA Professions" was offered in 2016, 2017, and 2018. The data collected would result in three "observations.") This translated to each program having 40−50 observations per program except OT, which did not have enough observations to meet the criteria of 40−50 observations with all three data points.

Table 1
Summary of data.

Program	Evaluation score (mean (SD))	Grade (mean (SD))	Credit hours (mean (SD))
Composite (n = 265)	4.52 (0.44)	3.49 (0.39)	2.89 (1.33)
CLS (n = 45)	4.63 (0.31)	3.17 (0.42)	3.13 (0.95)
CPS (n = 47)	4.56 (0.36)	3.32 (1.00)	2.36 (0.41)
OT (n = 31)	4.76 (0.31)	3.84 (0.23)	2.71 (1.47)
PA (n = 43)	4.08 (0.44)	3.48 (0.31)	3.98 (1.92)
PT (n = 44)	4.60 (0.47)	3.82 (0.15)	2.64 (0.97)
SLP (n = 46)	4.53 (0.36)	3.45 (0.31)	2.52 (0.81)

2.2. Data analysis

An a priori sample size determination was conducted using. It was determined that in order to achieve 80% power with a level of significance of 0.05 and to detect a small effect size (corresponding to a multiple square correlation of 0.10), at least 90 observations were required.[9] Multiple linear regression was performed using the Microsoft Excel (2010) data analysis add-on to determine correlation between course evaluation scores (dependent variable), course grade and/or course credit hours (independent variables, respectively). This was completed for the composite data, and for individual programs. Lack of collinearity was established between the two independent variables, as well, by performing simple linear regression (VIF = 1.08). Data were plotted as assessed for nonlinear correlations (exponential, logarithmic, polynomial, and power).

3. Results

Table 1 summarizes the data for each of the individual programs analyzed in this study. Table 2 details the multiple linear regressions performed to determine the predictive relationships between the dependent variable (course evaluations) and the two independent variables (grades and credit hours). Analysis of the composite data showed neither course grade nor credit hours was a predictor of course evaluation scores (F (1254) = 21.30, p = 0.08, with an r^2 of 0.02). The r^2 values for comparisons within individual programs were also low, indicating little to no predictive relationship within any of them, even though the relationship was statistically significant for some.

4. Discussion

The overall lack of relationship between evaluation scores with either grades or credit hours supports the use of course evaluations as a measure of course effectiveness. It appears that students are, in general, evaluating courses based upon other measures, with the assumption that these measures are appropriately related to course quality.

There were marked differences between the individual programs' results in this study. Several potential differences between programs were further investigated, to determine if they might explain the differences in results, but there were few patterns to be discerned. It is interesting to note that all programs EXCEPT PA (which had the highest r^2 value, and a statistically significant predictive relationship between course evaluations and course grades) have two or fewer instructors in each course. The PA program,

Table 2
Results of multiple linear regression to show relationships between evaluation scores and independent variables.

Program	r^2	Grades			Credit hours		
		Regression coefficient	95% CI	p-value	Regression coefficient	95% CI	p-value
COMPO-SITE	0.02	0.12	(−0.01, 0.27)	0.08	−0.02	(−0.06, 0.02)	0.32
CPS	0.08	0.22	(0.004, 0.44)	0.046[a]	−0.02	(−0.11, 0.08)	0.72
CLS	0.10	0.08	(−0.30, 0.46)	0.68	0.14	(0.004, 0.27)	0.04[a]
OT	0.10	−0.45	(−1.04, 0.14)	0.13	−0.067	(−0.16, 0.02)	0.15
PT	0.02	0.95	(−0.02, 1.92)	0.06	0.07	(−0.08, 0.22)	0.33
PA	0.26	0.90	(0.42, 1.38)	<0.001[a]	0.20	(0.02, 0.18)	0.01[a]
SLP	0.01	0.07	(−0.33, 0.47)	0.73	0.03	(−0.12, 0.18)	0.66

[a] Statistically significant based upon a α = 0.05.

however, typically has four or more per course, with some having over 15 in a semester.

It is acknowledged that the degree granted and number of students may greatly impact the teaching and assessment methods of a program. That said, degree granted was not related to the discrepancy between programs, as one of the programs that showed a statistically significant result (CPS) is a bachelor's program, while the other (PA) is a master's. The number of students enrolled in each cohort was also not related, as CPS has the fewest number per cohort, and PA the greatest.

Lastly, it was determined that all programs handled course evaluations essentially the same way (i.e., giving students time in class to do them, and scheduling them ahead of time), with the exception of the PA program having students complete all evaluations for a semester at one time, as opposed to separate times allotted for different courses.

While there have been studies in non-healthcare fields that show the usefulness of student evaluations of courses and teaching,[2] there is just as much research that highlights the impact of student biases and how it affects their evaluations.[6,10] In addition, while grades are the most commonly used measure to reflect student learning, they do not always predict student performance in subsequent related courses.[11,12] Therefore, if the applicability of knowledge and skills gained from a course to subsequent, related courses is a surrogate measure of teaching effectiveness, students are not able to truly evaluate this facet at the termination of said course. All of these factors contribute to evaluation scores and should be considered when weighing the results.

4.1. Limitations

There were several limitations to this study. The authors acknowledge that the study would have been strengthened if individual students' grades and the evaluations they submitted could be compared; however, all evaluations in this institution are anonymous, which makes such a comparison impossible.

It is believed by all program directors that course evaluations are likely strongly related to the course coordinator or primary instructor(s) and their rapport with students. This institution has separate evaluations for courses and instructors, so this faculty-specific information was not included in the analysis. However,

the small and close-knit nature of this SAHP, as well as the expectation of confidentiality regarding individual instructor evaluations, makes an analysis of course vs. instructor evaluations infeasible using the current study design and data collection/analysis techniques. If another institution utilized combination evaluations that included both course and instructor, these results may not be as generalizable.

Every program uses a different evaluation tool, so the "general" question was worded slightly differently. Some programs had an outright "Rate the course in general" question, while others had to choose the question that best met this need, such as, "How well did this course meet the stated objectives?" The investigators acknowledge that these two statements are not evaluating the same concept and, thus, could have affected overall results. However, they also felt that this method was more valid than calculating an "overall" score for evaluations based upon all of the questions, since each program has different expectations and needs, and, thus, number and type of evaluation questions.

The method of calculating course grade was necessary due to differences in record keeping between the programs. However, differences in numbers of students in each program (ranging from five to 39) could skew the GPA-like calculations used for this part of the analysis.

While OT was an outlier in the number of useable observations that were available, discussion with the Program Director indicated that the courses that were missing were, in fact, at random, which makes listwise deletion of the data appropriate, and will not result in bias in the analysis.[13]

5. Conclusion

Student course evaluations remain important tools for course assessment. Based on the results from this study, most faculty and administrators can be assured that course evaluations are not dependent upon either grades achieved in that course or its difficulty. However, since there was a significant relationship detected for some programs in the SAHP, the importance of utilizing multiple assessment tools to determine the effectiveness of a course must be stressed. Faculty and administrators should take these variations and influencing factors into consideration when reviewing course evaluations.

References

1. Kulik JA. *Student ratings: validity, utility, and controversy. New directions for institutional research. No 109, Spring*. Jossey Bass; 2001:9−25. https://doi.org/10.1002/ir.1.
2. Spooren P, Brockx B, Mortelmans D. On the validity of student evaluation of teaching: the state of the art. *Rev Educ Res*. 2013;83(4):598−642. https://doi.org/10.3102/0034654313496870.
3. Flaherty C. Arbitrating the use of student evaluations of teaching. *High Educ*; August 31, 2018. https://www.insidehighered.com/quicktakes/2018/08/31/arbitrating-use-student-evaluations-teaching. Accessed July 26, 2019.
4. Flaherty C. Teaching eval shake-up. *High Educ*; May 22, 2018. https://www.insidehighered.com/news/2018/05/22/most-institutions-say-they-value-teaching-how-they-assess-it-tells-different-story. Accessed July 26, 2019.
5. Stroebe W. Why good teaching evaluations may reward bad teaching: on grade inflation and other unintended consequences of student evaluations. *Perspect Psychol Sci*. 2016;11(6):800−816. https://doi.org/10.1177/1745691616650284.
6. Boring A, Ottoboni K, Stark PB. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Sci Res*. 2016. https://doi.org/10.14293/S2199-1006.1. SOR-EDU.AETBZC.v1.
7. Bowen RE, Grant WJ, Schenarts KD. The sum is greater than its parts: clinical evaluations and grade inflation in the surgery clerkship. *Am J Surg*. 2015;209(4):760−764. https://doi.org/10.1016/j.amjsurg.2014.10.023.
8. Szafran RF. The effect of academic load on success for new college students: is lighter better? *Res High Educ*. 2001;42(1):27−50. Available at: https://link.springer.com/content/pdf/10.1023/A:1018712527023.pdf. Accessed July 26, 2019.
9. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175−191.
10. MacNell L, Driscoll A, Hunt AN. What's in a name? Exposing gender bias in student ratings of teaching. *Innovat High Educ*. 2015;40(4):291−303. https://doi.org/10.1007/s10755-014-9313-4.
11. Carrell SE, West JE. Does professor quality matter? Evidence from random assignment of students to professors. *J Polit Econ*. 2010;118(3):409−432. https://doi.org/10.1086/653808[3]. ISSN 0022−3808.
12. Braga M, Paccagnella M, Pellizzari M. Evaluating students' evaluations of professors. *Econ Educ Rev*. 2014;41:71−88. https://doi.org/10.1016/j.econedurev.2014.04.002.
13. Kang H. The prevention and handling of missing data. *Korean J Anesthesiol*. May 2013;64(5):402−406. https://doi.org/10.4097/kjae.2013.64.5.402.