Using Student Evaluations of Teaching to Support Faculty: Use Proportions Instead of Means to Analyze SETs

B. D. McCullough
Drexel University

Raluca Teodorescu
Montgomery College

We analyze student evaluations of teaching (SETs) for 17 courses covering 760 students taught by 11 faculty. We discuss theoretical and practical implications of analyzing SETs using means and proportions. We find that both methods provide similar results for some questions and different results for others. When the results are different, we find that proportions lead to a more accurate and complete analysis that is essential when using SETs for faculty training and support. Eight out of 11 faculty members asked that the department evaluation reports should be based on proportions rather than means, and none preferred means.

INTRODUCTION

Student Evaluations of Teaching (SETs) are common instruments administered at the end of the semester and involve questionnaires given to students, in which students respond by marking one of several choices, e.g., Strongly Agree, Agree, Disagree or Strongly Disagree. Occasionally, these questionnaires also include one or two free response questions, but most of the questions are Likert-like format. Although there is controversy surrounding many aspects of SETs (Stark and Freishtat, 2014), they continue to be one of the most popular methods of assessment among academic administrators. Additionally, they are tools that inform faculty who wish to improve their teaching. Within the context of physics education (Henderson, et al., 2014) interviewed 72 physics instructors and found out that 90% of them reported that their institution uses SETs as primary way of assessing physics teaching effectiveness. The authors found that SETs are the most common way of evaluating physics instructors effectiveness among other methods that include peer observations, teaching portfolios, and students performance. Additionally, they also found that there is little overlap in the ways faculty evaluate their own teaching, how their teaching is evaluated by their own institutions, and what research suggests are best practices in assessment and evaluation of teaching. In this paper, we make a case that a careful analysis of the metrics used by physics and astronomy SETs is needed from both administrative and instructional perspective. Specifically, we focus on the theoretical and practical aspects of SETs analysis using means and proportions.

We analyze the results of the evaluations of 17 physics and astronomy courses taught by 11 faculty in two consecutive semesters. Among the 11 faculty, five were part-time and six were full-time faculty.

Almost all courses (all except two) have been taught in active learning format. 760 students completed the evaluations and the percent of the students who completed the evaluations in each class ranged from about 50% to about 90%. The wording and the format of the questions used have been discussed and decided by the faculty prior to the administration. These discussions sought to clarify aspects like questions relevance for the content and for the teaching approach, questions clarity, the alignment between the department evaluation and the university requirements for such instruments, as well as the administration procedures. In the end, faculty decided that 22 questions (19 Likert and 3 free response) are necessary and sufficient for collecting student feedback. In this paper, we focused on the 19 questions that can be grouped in the following categories:

a) general questions Q1, Q2, Q3
b) questions about the course Q4, Q5, Q6, Q7, Q8
c) questions about the amount learned Q9, Q10, Q11, Q12, Q13, Q14
d) questions about the instructor Q15, Q16, Q17, Q18, Q19

MEANS VS. PROPORTIONS

In this section we compare the results of SETs analyzed via means and proportions, and we present theoretical and practical implications for assessing instructor performance in the courses. Additionally, we discuss the challenges and the limitations that an administrator and a faculty member can face when trying to interpret SETs results for guiding faculty teaching improvement. Our research is motivated by the considerations highlighted in the Introduction and by previous findings related to quantitative analyses of SETs. McCullough and Radson (2011) (hereafter "M&R") compared means and proportions for a data set comprising 737 students in 49 classes from various disciplines taught by 22 instructors. They employed univariate and bivariate analysis and found that proportions are more appropriate for analyzing SETs than means of ordinal variables. The authors also found that when almost all the responses are uniformly high or uniformly low, means and proportions reach the same conclusions. It is in the middle that they can differ and they argue that proportions give more accurate insights than means. They suggest that more research is necessary and some must be field-specific because students responses depend, in part, on the subject being taught. The present paper extends this research.

Theoretical Considerations

The field of Measurement Theory is devoted to the concept that our analyses of data should be correct in a meaningful way (Hand, 2010). One of the field's standard results, universally reproduced in elementary statistics book, is that there are four types of data:

- *nominal* – observations can be placed in categories, *e.g.*, North, South, East, West
- *ordinal* – observations can be placed into ordered categories where one category is, in some sense, bigger than another, *e.g.*, Olympic medals: Gold, Silver, and Bronze
- *interval* – numeric measurements are made without reference to a true zero, *e.g.* temperature in degrees Fahrenheit or Celsius
- *ratio* – numeric measurements are made with reference to a true zero, *e.g.*, weight or pressure.

One cannot compute an average for nominal or ordinal data, because the preliminary concepts of addition and division are not defined for such data. Assigning numbers to the categories does not imply that the categories themselves will admit the division operator. The matter is summed up nicely by Hand (1996, p.462): The arithmetic mean is not defined for ordinal scales. Consequently, from a theoretical perspective the use of means to analyze students evaluations results is questionable because evaluations-related data is ordinal. This idea is further detailed by M&R [3]. Due to space considerations, we analyze the results of only two key questions often used by administrators:

Q3: How difficult is the subject matter? A = Very Easy, B = Easy, C = So-so, D = Hard, E = Very Hard

Q19: Overall rating of the instructor. A = The least, E = The most

For converting to a Likert Scale we used A = 1 (the least), B = 2, C = 3, D = 4, E = 5 (the most)

FIGURE 1
SCATTERPLOT FOR QUESTION 3 – p_{345} IS THE PROPORTION OF STUDENTS WHO RESPONDED 3,4 or 5

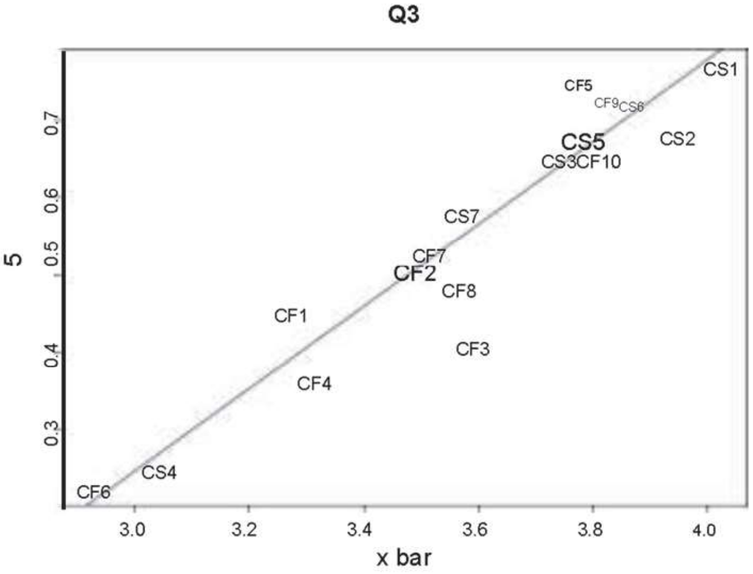
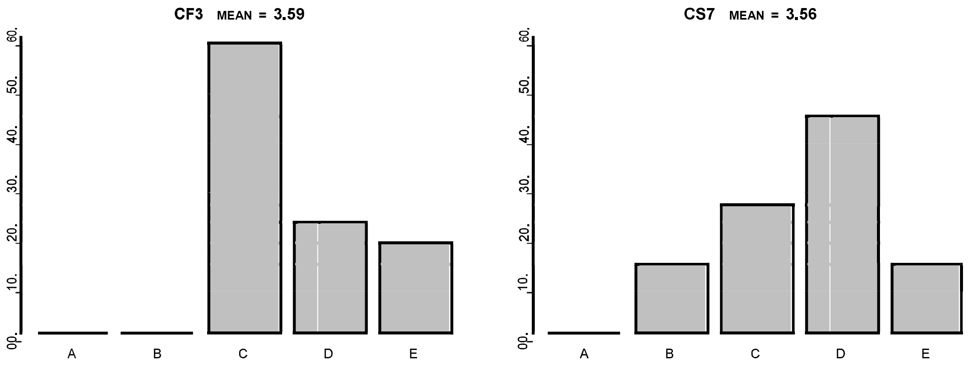


Figure 1 shows a largely linear relation between the mean and the proportion for Q3, though there are significant anomalies. An administrator who analyzes only the means (shown on the abscissa) might think that classes CF3, CF2, CF7, CF8 and CS7 were all about the same level of difficulty. Yet the proportions (shown on the ordinate) indicate that the students perceive the sections levels of difficulty differently. In CS7, 15% of the students think the subject matter is easy (rated it a 1 or a 2); 10% of the students in CF2, CF7 and CF8 think the subject matter is easy; and none of the students in CF3 think the subject matter is easy. This situation is revealed starkly by the barplots of these results, shown in Figure 2. For space considerations we show only CF3 and CS7. As a practical matter, there is a large difference between thinking that students perceive CF3 and CS7 to be roughly the same, as means would indicate, and the truth, which is revealed in the barplots.

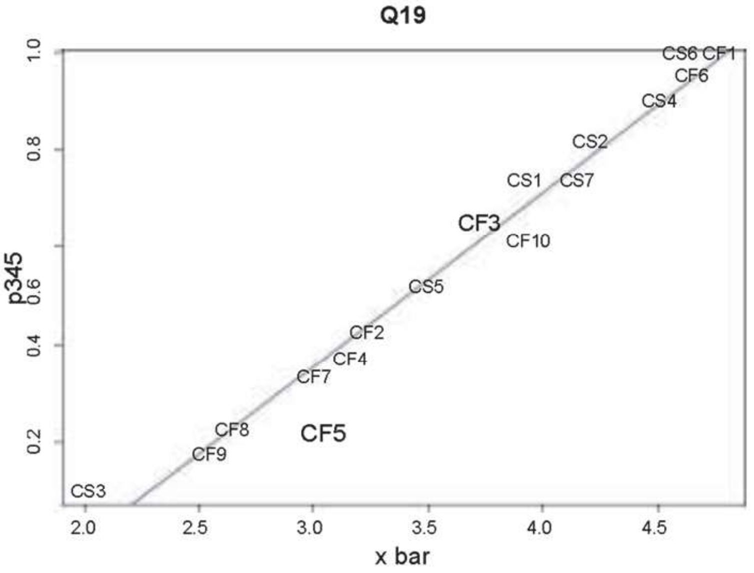
As another example of what can be obscured by means but revealed by proportions, consider Question 19. As shown in Figure 3, there is close agreement between means and proportions with two notable exceptions: CS3 and CF5. Looking at the means, it might appear that CF7 and CF5 are similar, since they both have a mean of 3.0. Looking at the proportions, they are clearly different. The raw data, presented in Table I, bear this out. Observe that in Table 1, $\bar{x}$ for CF7 = (1 × 4 + 2 × 9 + 3 × 16 + 4 × 13 + 5 × 2)/44 = 132/44 = 3.

FIGURE 2
BARPLOTS FOR QUESTION 3, SECTIONS CF3 and CS7



As is obvious, while both sections have similar percentages for categories B and E, CF7 has 30% for category D while CF5 has only 18%. Without a doubt, CF5 is much more like CF8 and CF9 (which have means of 2.5) than CF7.

FIGURE 3
SCATTERPLOT FOR QUESTION 19: p_{345} IS THE PROPORTION OF STUDENTS WHO RESPONDED 3,4 or 5



For the record, M&R presented similar plots that show a decidedly nonlinear relation between the mean and the proportion, so it is not always the case that the relation is largely linear, as in the present data. They show that means and proportions can lead to dramatically different conclusions concerning instructor performance, and when there is a discrepancy, examination of the underlying data favors the proportion. This is because the mean is particularly susceptible to the influence of outliers, whereas the proportion does not suffer this defect. We note that of the remaining 17 questions, the results of means and proportions agreed linearly (i.e., the plot of one against the other was very nearly a straight line) for 11 questions (as in Figure 3, with the exceptions of CS3 and CF5), and showed disagreement for 6 (as in Figure 1)

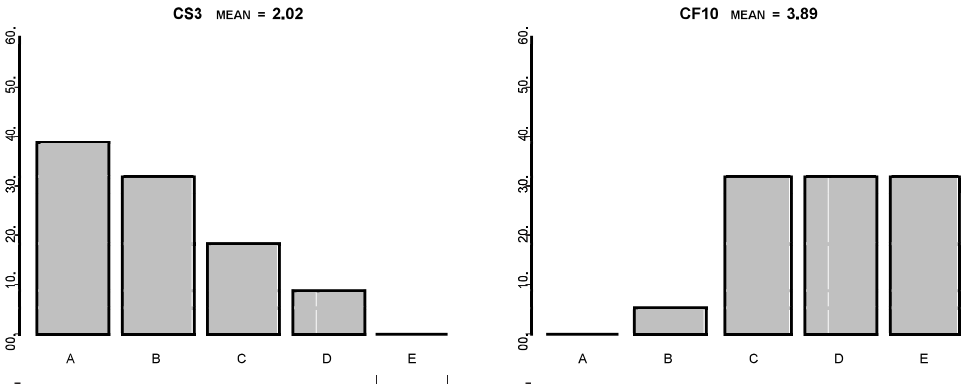
TABLE 1
THE COUNTS, MEANS AND PROPORTIONS FOR COURSES CF7 AND CF5 FROM FIGURE 3

	CF7		CF5	
$\bar{x}$	3.00		3.04	
p_{45}	0.35		0.22	
	count	percent	count	percent
A. 1	4	09	0	0
B. 2	9	20	6	22
C. 3	16	36	15	56
D. 4	13	30	5	18
E. 5	2	5	1	4
sum	44	100	27	100

Practical Considerations

Let us consider the courses CF10 and CS3 and imagine that the instructors solicited advice for improving their teaching. Suppose that the consultant uses data only from Q19 (Figure 3). This is certainly not recommended and perhaps not true in practice. However, it is an informative example that can illustrate the limitations that means create in practice, in contrast with a much richer information that proportions can offer, when SETs are used to support faculty. Both courses are introductory calculus-based courses for science and engineering majors. Both courses have been taught by experienced instructors but they used different teaching strategies and had different experiences with the particular courses. CF10 has been taught by a male instructor for the first time and he used worksheets inspired by the University of Washington Tutorials [6], while CS3 has been taught by a female instructor many times previously and she used a strategy similar to SCALE UP [7]. Figure 4 shows the barplots for both courses, while Tables II, III, and IV show corresponding means and proportions as well as possible practical advices offered to instructors interested to improve their teaching strategies.

FIGURE 4
BARPLOTS FOR QUESTION 19, SECTIONS CS3 and CF10



For course CF10, 19% of the students completed the evaluations, while 61% of the students completed the evaluations for course CS3. From Figure 4, one can observe that for CF10 most of the students provided positive feedback about the instructor. In contrast, for CS3 about half provided negative feedback.

TABLE 2
POSSIBLE PRACTICAL ADVICE BASED ON MEAN ANALYSIS

Course	Instructional Method	Mean
CF10	worksheets	3.38
CS3	SCALE-UP	1.97

A faculty mentor who needs to provide guidance for teaching improvement and uses means (see Table 2) could suggest:

(a) for the CF10 instructor. This semester is a good start. For future, try to enhance what you have done.
(b) for the CS3 instructor. Given that you have taught this course 3 times, you should consider making significant changes.

However, if the mentor uses proportions (see Table 3), he or she can provide much richer advice as illustrated below:

(a) for the CF10 instructor. The methods you used to motivate students to provide feedback are not working (based on Table 3). It seems that your students appreciated your teaching and this is encouraging, especially because this is your first time teaching the course. However, given the low proportion of survey participants, it is hard to conclude that your instructional method satisfied the entire class. Let's come up with some methods of encouraging students to complete the evaluations and see how you can implement them in the future. If your p12 will have the same value when a much higher proportion of students complete the evaluations than this may suggest that you are an impressive instructor.
(b) for the CS3 instructor. The methods you used to motivate students to provide feedback are working, but you can probably improve them (based on Table 3). In the few semesters you taught this course, you mastered ways of encouraging students to provide feedback. However, about a half of the students who completed the evaluations did not seem to appreciate your way of teaching. Let's look at some individual aspects of your teaching and analyze them separately (based on Table IV).

TABLE 3
POSSIBLE PRACTICAL ADVICE BASED ON p_{12} ANALYSIS

Course	Instructional Method	p_{12}
CF10	worksheets	0.05
CS3	SCALE-UP	0.71

Consequently, we notice that in practice, the use of means, in addition to not being supported theoretically, does not offer information as rich as proportions. Thus, the choice of proportions versus means is justified not only theoretically, but also for practical purposes.

Given the fact that in most institutions SETs are offered in online environments and their results are automatically calculated, we envision that changes in computing algorithms can be implemented relatively straightforward and with minimal effort. Thus, such measures might not be difficult to implement in practice.

We caution the readers that the analysis featured in this section is not meant to draw any conclusions about the effect on student evaluations that the University of Washington Tutorials vs. the SCALE UP pedagogy have. An analysis performed on many more courses is suitable for such a study. We also caution the readers that this analysis is not meant to show what practical advice has to be given to faculty who wish to improve their teaching. Providing feedback based on one evaluation question is certainly limited and possibly wrong. The examples showed are used exclusively for illustrating the difference between the two methods.

CONCLUSIONS

In this paper, we compared the results of students' evaluations of teaching. We show that the use of proportions for analyzing teaching evaluations is statistically more rigorous than the use of means and offers more advantages in practice. In the case of physics and astronomy courses we analyzed there is substantial agreement between means and proportions, especially at the extremes of both measures. However, in the intermediate range the measures are most likely to disagree, and this is problematic: often faculty who seek and need support for teaching improvement are those with evaluation results in the middle range. Perhaps more important is the fact that, when using SETs to guide and support faculty, mentors can offer misleading guidance if they do not rely on proportions. Additionally, Q19 discussed here is one of the most-used questions by administrators and basing the guidance on incorrect information can have negative impacts on faculty. Moreover, faculty seems to favor proportions. The results of this study were shown to the eleven faculties in question, and they were asked whether they would like to continue with means or switch to proportions. Eight answered proportions, one answered, both, one answered means, and the last responded, "I need more information."

In the next phase of this study, we plan to focus on how better to use proportions and on understanding the types of support that should be given to faculty whose evaluations indicate low proportions of satisfied students.

ACKNOWLEDGEMENTS

We thank Dr. Michael Lujan, without whom this work would have not been possible. We also thank Profs. Xiangyun Qiu and Oleg Kargaltsev, Dr. Carol O'Donnell, and Mr. Peter Hahn for their valuable feedback. This work was supported by the George Washington University Office of Academic Planning and Assessment and the George Washington University Department of Physics. We also thank Dean Muhammad Kehnemouyi for supporting the publication of this article.

REFERENCES

Beichner, R. J., Saul, J. M., Abbott, D. S., Morse, J. J., Deardorff, D. L., Allain, R. J., Bonham, S. W., Dancy, M. H. and Risley, J. S. (2007). The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project, in E. F. Redish (Ed.), *Research-Based Reform of Introductory Physics*.

Hand, D. J. (2010). *Measurement Theory and Practice: The World through Quantification*, New York: Wiley.

Hand D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society - Series A*, 159 (3), 445-492.

Henderson, C., Turpen, C., Dancy, M., & Chapman, T. (2014). Assessment of teaching effectiveness: Lack of alignment between instructors, institutions, and resaerch recommendations, *Physical Review Special Topics – Physics Education Research*, 10, 01010601-01010620.

McCullough, B. D. & Radson, D. (2011). Analyzing student evaluations of teaching: comparing means and proportions, *Evaluation and Research in Education* 24, (3), 183-202.

McDermott, L., Shaffer P. & University of Washington Physics Education Group (2002). *Tutorials in Introductory Physics*, NJ: Prentice Hall.

Stark, P. B. & Freishtat, R. (2014). An Evaluation of Course Evaluations, ScienceOpenResearch DOI: 10.14293/S2199-1006.SOR-EDU.AOFRQA.v1