Too Early for Physics? Effect of Class Meeting Time on Student Evaluations of Teaching in Introductory Physics

R. G. Tobin, Department of Physics & Astronomy, Tufts University, Medford, MA

This paper reports observations that show a significant effect of class meeting time on student evaluations of teaching for an introductory college physics class. Students in a lecture section with an early-morning meeting time gave the class and instructors consistently lower ratings than those in an otherwise nearly identical section that met an hour later, even for aspects of the course that were seemingly identical across the two sections.

Student evaluations of teaching (SET) are almost as extensively debated as they are widely used.[1-5] Research has shown that evaluation scores can be significantly affected by such extraneous factors as the gender,[6-8] race,[9] age,[4,8] verbal and nonverbal mannerisms,[8,9] and physical attractiveness[10] of the instructor, students' grade expectations,[7] the physical environment of the classroom,[11] and even the weather on the day the evaluation is completed.[12] Studies differ on whether SET have *any* meaningful correlation with objective measures of student learning, and, if so, whether the correlation is positive or negative.[1,12,13] Nevertheless there is no sign that SET are going away, so it is important that educators and administrators be aware of, and objectively document, as many of the factors that affect them as possible.

Description

The course studied here was the first semester of calculus-based introductory physics at a selective private research-intensive university in the United States. 141 students completed the course, predominantly freshman engineering students. For lectures the class was divided into two sections. Each section met for 50 minutes three times per week. One section (the "early section") met Monday and Wednesday 8:30–9:20 a.m. and Thursday 9:30–10:20 a.m. The other section (the "later section") met an hour later: Monday and Tuesday 9:30–10:20 a.m. and Thursday 10:30–11:20 a.m. Both sections were taught in the same room by the same instructional team of an experienced professor and an advanced graduate student. That is, in a given week a section might have had two classes led by the professor and one by the graduate student, or vice versa, but over the course of the semester each section had 60% of the classes led by the professor and 40% by the graduate student. Each instructor attended the other's lectures, and the two instructors closely coordinated their presentations, using the same lecture materials (slides, examples, in-class questions, demonstrations). Thus the students in the two sections received nearly identical instruction from the same instructors. The lecture format was based on Peer Instruction,[14,15] with much time devoted to conceptual questions (using electronic student response devices) and both small group and all-class discussion.

In the early section 57 students completed the course and 84 in the later section. Students selected their lecture sections during course registration and had no obvious reason other than class time to select one or the other. Records from the electronic response system show that students rarely attended the "wrong" section, perhaps in part because they received participation credit only in their assigned sections. Table I compares the composition of the two sections. The only identifiable and significant difference is that the early section had a lower proportion of engineering students. Comparison of SET scores for engineering and liberal arts students in each section revealed no significant difference between the two groups, and in fact on almost all items liberal arts students gave slightly higher scores than engineering students, so this difference in student population cannot likely account for the lower overall scores in the early section.

Table I. Composition of the two sections.

	Early section	Later section
Number of students completing course	57	84
Percent first-year	90%	90%
Percent engineering	54%	71%
Percent female	31%	26%
Evaluation response rate	89%	87%

Apart from the lectures, students in the two sections of the class were treated identically. They were mixed in recitation and laboratory sections, did the same homework, attended (or not) the same office hours, and took the same exams, which were graded anonymously and without regard to lecture section. All graded work was returned to students in their recitation sections or picked up in the department office.

At the end of the course students completed a university-mandated standard course evaluation. The evaluation was anonymous, online, and completed during the final week of classes and reading period, before the final examination. The instructor was given access to the results only after final grades were submitted. As shown in Table I, the participation rates were 89% (51/57) for the early section and 87% (73/84) for the later section, motivated in part by the promise of a small amount of extra credit to all students in whichever section had the higher participation, or to both if they exceeded 90%. (In fact all students in both sections received the extra credit.) The evaluation included both multiple-choice and open-response questions. The multiple-choice questions offered five options from “Very Poor” to “Excellent,” which for purposes of analysis were associated with numerical values from 1 to 5, respectively.

Results

Figure 1 compares the distribution of SET ratings in the two sections, for the five multiple-choice questions for which the difference between the sections was most significant. The mean evaluation scores for those questions are compared graphically in Fig. 2, and summarized in Table II.

Across all questions, including seven others that are not listed in Table II, the students in the early section gave the class and the instructor lower mean scores than those in the later section. For the overall evaluation of the course and of the instructor, which are the items given the greatest weight in faculty evaluations at this institution, the early section’s mean ratings were more than 0.3 lower than the later section’s (effect size 0.4). Examination of the distribution of individual ratings (Fig. 1) shows that the difference in means arose from an overall shift of the distribution, not from a small number of outliers.

Two measures were used to assess the statistical significance of the differences between the sections, and the results are shown in Table II. The p_t values were calculated using a two-sided t-test with unequal variances. This and similar parametric measures are commonly used in studies of SET. There are legitimate objections to this statistical approach, however, including the assignment of interval values to ordinal data,[16] the non-normal distribution of the ratings, and the likelihood that individual students’ ratings are not entirely independent. The p_{MW} values are determined from a Mann-Whitney non-parametric test that is better suited to ordinal data that is not normally distributed. (p represents the probability that a difference between the two samples at least as large as the

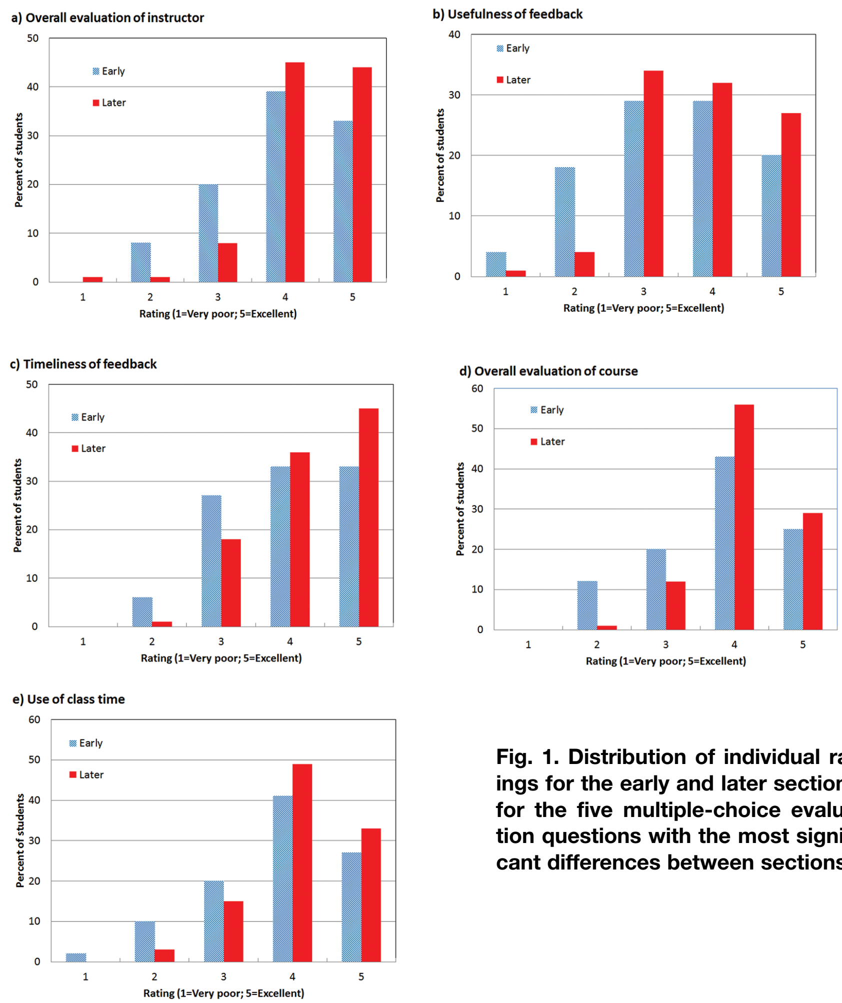

Fig. 1. Distribution of individual ratings for the early and later sections, for the five multiple-choice evaluation questions with the most significant differences between sections.

Table II. Summary of selected multiple-choice items from the student evaluation for the two sections. The scale on all items was from 1 to 5, with 5 being the most favorable. In the comparison, the difference is (Late – Early). The column labeled p_t gives the p-values for a two-sided t-test with unequal variances, while the column labeled p_{MW} gives p-values for a Mann-Whitney nonparametric significance test. The items listed are those for which at least one of the p-values was lower than 0.1, and * denotes items for which a real difference between the two sections can be ruled out owing to the structure of the course. There were seven other multiple-choice items; on all items the scores for the later section were higher than those for the early section, but the differences for the individual items were not statistically significant.

	Early (N = 51)		Later (N = 73)		Comparison			
Item	Mean	Std. Dev.	Mean	Std. Dev.	Diff.	Effect size	p_t	p_{MW}
Overall evaluation of instructor	4.0	0.9	4.3	0.8	0.3	0.4	0.05	0.08
*Usefulness of feedback on assignments, exams, and other work	3.4	1.1	3.8	0.9	0.4	0.4	0.05	0.08
*Timeliness of feedback on assignments, exams and other work	3.9	0.9	4.3	0.8	0.3	0.4	0.06	0.09
Overall evaluation of course	3.8	1.0	4.1	0.7	0.3	0.4	0.04	0.11
Use of class time to promote learning	3.8	1.0	4.1	0.8	0.3	0.3	0.08	0.16

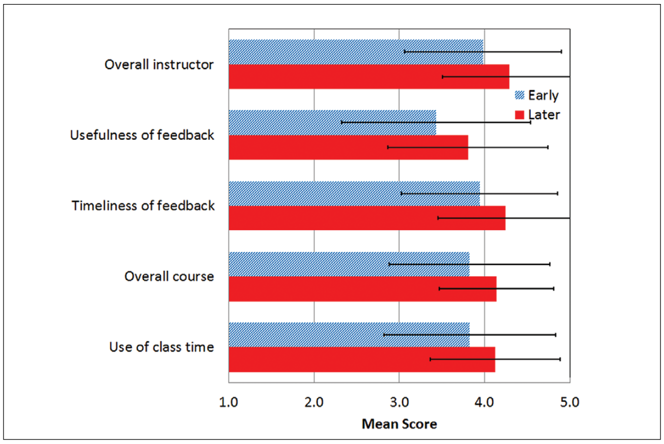

Fig. 2. Comparison of mean evaluation scores for the five items included in Fig. 1 and Table II. Error bars represent one standard deviation.

observed difference would have appeared if the two samples were drawn from the same distribution—that is, if there were no real difference between the two sections.) For the overall evaluations of the course and instructor, for example, the p_t (p_{MW}) -values were 0.04 (0.11) and 0.05 (0.08), respectively, indicating a high probability that the difference between sections is a real effect.

Discussion

Is it possible that the students in the later section really did receive a superior educational experience, even though the instructors, materials, and pedagogical approach were identical? The instructors might have been more energetic or enthusiastic. Also, for two of the three classes each week the early section's lecture preceded the later section's (the later section had its second class of the week a day before the early section) so the instructors had an opportunity to learn and improve.

Table III. Comparison of course performance for the two sections. There are no statistically significant differences between the two sections. *p*-values are calculated using a two-sided *t*-test with unequal variances.

	Early (N = 57)		Later (N = 84)		Comparison		
	Mean	Std. Dev.	Mean	Std. Dev.	Diff.	Effect size	p
Average midterm exam grade (%)	74.0	10.1	73.0	10.6	-1.0	-0.1	0.57
Final exam grade (%)	68.7	12.4	71.5	11.4	2.8	0.2	0.18
Overall course grade (%)	81.7	7.3	82.9	6.9	1.2	0.2	0.33

There is little evidence in the students' course performance, however, for the hypothesis that students in the early section received significantly inferior instruction. Table III compares the average grades for the two sections both on the final exam (identical for both sections and graded anonymously) and for the overall course (including homework, labs, exams, and participation). Fig. 3 shows the grade distributions for the two sections. While students in the later section did score slightly higher (effect size 0.2), the differences are not statistically significant (p = 0.18 and 0.33, respectively), and come largely from a small number of students in the early section who received unusually low scores. Yet Fig. 1 shows that the difference in evaluation scores was distributed across the class, and not due to a small number of highly dissatisfied students.

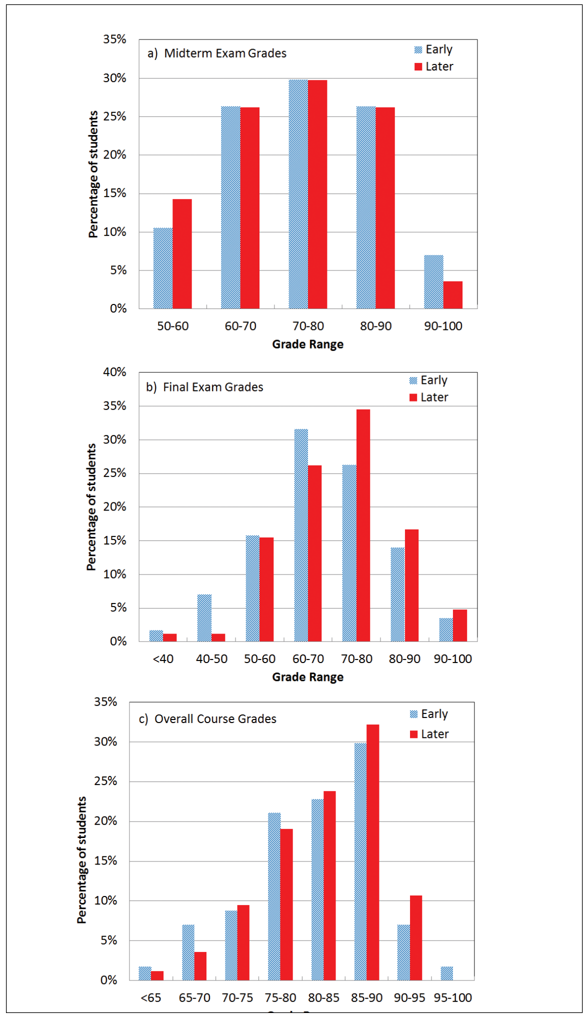

Fig. 3. Comparison of grade distributions for the two sections for (a) midterm exams (average of two exams), (b) final exam, and (c) overall course grade.

Table III and Fig. 3 also compare the midterm exam scores for the two sections. This quantity is a reasonable proxy for students' perception of their course performance at the time when they completed the course evaluation, which has been shown to be correlated with course evaluations.[7] There was no significant difference in midterm grades between the two sections, and in fact the earlier section scored slightly higher (effect size 0.1), so the difference in course evaluations cannot be attributed to a difference in students' grades at the time of the evaluation.

Further, two of the questions with the largest and most significant differences between the two sections were those about the timeliness and usefulness of written feedback provided on assignments and exams (effect size 0.4, p_t = 0.06 and 0.05, p_{MW} = 0.09 and 0.08, respectively). Because homework and exams from the two sections were intermingled and graded by the same people, and were returned in recitation section, it is not plausible that there were real differences on these dimensions.

One area in which the classroom experience may have differed between the sections, despite every effort to provide identical instruction, is in the liveliness and quality of student engagement in in-class discussions. Indeed, both instructors felt that the early section was noticeably less enthusiastic, and that it was more difficult to prompt discussion in that section. The data in Table III and Fig. 3 show that any such difference did not affect course performance, but it could conceivably account for the differences in students' ratings of the effective use of class time, and could have led to generally lower overall satisfaction in the early section that affected responses on other items. Even if that is the case, however, the most likely reason for the difference in liveliness seems to be the difference in class time. It is possible that the larger proportion of liberal arts students in the early section led to lower enthusiasm, but the instructors' perception was of an overall level of lethargy, rather than a subset of the class that was disengaged. Further, some of the students most active in discussion were in liberal arts, and as noted above, liberal arts students actually gave slightly higher SET scores than engineering students. It is possible however, that a less discussion-centered class would not be subject to as large an effect of class time on SET results.

Therefore, while it is conceivable that some other difference between the sections accounts for the difference in evaluation scores, the most plausible interpretation is that students whose physics class met at 8:30 and 9:30 a.m. were less happy with their experience than those whose class met at 9:30 and 10:30 a.m., simply because of the early hour, and their relative discontent was expressed in their evaluation of all aspects of the class. In any event, the students appeared to be unaware of the source of their relative dissatisfaction; in the open-response questions not a single student commented on the class meeting time. Critical comments in both sections centered on aspects that were the same in both sections, such as the difficulty of the homework, dissatisfaction with the labs, and frustration with the amount of lecture time spent in discussion.

That an early class time can result in lower evaluation scores may not be surprising to any instructor who has tried to interest a roomful of sleepy teenagers in the intricacies of Newton's laws, or to any of those early-morning physics students. Still, as long as SET continue to play an important role in the institutional evaluation of instruction, instructors and those who evaluate them need to be aware of the extraneous factors that can affect the numbers, and to have evidence of those effects. It is unusual to have a relatively well-controlled instructional context that isolates the effect of class time. Moreover, the observed differences in course and instructor ratings are significant not only statistically but also professionally: At the institution in question a difference of 0.3 out of a range of 4 is taken seriously in the evaluation of instructors' effectiveness and can have a real effect on a person's career. The possibility that a one-hour difference in class meeting time can result in a shift of this magnitude is an important factor for instructors and those evaluating them to take into consideration.

Acknowledgments

The author thanks his co-instructor, Jeremy Wachter, and the anonymous reviewers of the manuscript.

References

1. D. E. Clayson, "Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature," *J. Marketing Educ.* **31**, 16–30 (April 2009).
2. P. Spooren, B. Brockx, and D. Mortelmans, "On the validity of student evaluation of teaching: The state of the art," *Rev. Educ. Res.* **83**, 598–642 (Dec. 2013).
3. S. L. Benton and W. E. Cashin, "Student Ratings of Instruction in College and University Courses," in *Higher Education: Handbook of Theory and Research,* Vol. 29, edited by M. B. Paulsen (Springer, Netherlands, 2013), pp. 279–326.
4. F. Zabaleta, "The use and misuse of student evaluations of teaching," *Teach. High. Educ*. **12**, 55–76 (Feb. 2007).
5. H. W. Marsh, "Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness," in *The Scholarship of Teaching and Learning in Higher Education, an Evidence-Based Perspective*, edited by R. P. Perry and J. C. Smart, (Springer, Netherlands, 2007), pp. 319–383.
6. L. MacNell, A. Driscoll, and A. N. Hunt, "What's in a name: Exposing gender bias in student ratings of teaching," *Innov. High. Educ.* **40**, 291–303 (Aug. 2015).
7. A. Boring, K. Ottoboni, and P. B. Stark, "Student evaluations of teaching (mostly) do not measure teaching effectiveness," *ScienceOpen Research* (2016). DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1
8. J. Arbuckle and B. D. Williams, "Students perceptions of expressiveness: Age and gender effects on teacher evaluations," *Sex Roles* **49**, 507–516 (Nov. 2003).
9. D. J. Merritt, "Bias, the brain, and student evaluations of teaching," *St. John's Law Rev.* **81**, 235–288 (2008).
10. D. S. Hamermesh and A. Parker, "Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity," *Econ. Educ. Rev*. **24**, 369–376 (2005).
11. M. C. Hill and K. K. Epps, "The impact of physical classroom environment on student satisfaction and student evaluation of teaching in the university environment," *Acad. Educ. Lead. J.* **14**, 65–79 (2010).
12. M. Braga, M. Paccagnella, and M. Pellizzari, "Evaluating students' evaluations of professors," *Econ. Educ. Rev.* **41**, 71–88 (2014).
13. T. Beleche, D. Fairris, and M. Marks, "Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test," *Econ. Educ. Rev.* **31**, 709–719 (2012).
14. E. Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).
15. C. H. Crouch and E. Mazur, "Peer Instruction: Ten years of experience and results," *Am. J. Phys.* **69**, 970–977 (Sept. 2001).
16. T. R. Knapp, "Treating ordinal scales as interval scales: An attempt to resolve the controversy," *Nurs. Res.* **39**, 121–123 (March/April 1990).

R. G. Tobin, Department of Physics & Astronomy, Tufts University, Medford, MA 02155; Roger.Tobin@tufts.edu; http://rtobin.phy.tufts.edu/