Regarding the Timing of Student Course/Instructor Evaluations

Jerome Witt, Ph.D.  Philadelphia University
Doreen Burdalski, M.B.A.  Philadelphia University

Abstract

The literature on student evaluations is vast. This research adds to that stream an empirical study comparing student evaluations on the last day of the semester to those completed three weeks earlier. The “official” evaluations were conducted during the 11th week of a 14-week semester. The original questions were replicated for the instrument administered on the last day of class. Results were compared question-by-question. Although students self-reported that their opinions were the same or better at the end of the semester, the actual last-day evaluations were lower for all ten questions in the instrument.

I. INTRODUCTION

In the School of Business Administration at a small mid-Atlantic university, student evaluations of courses and instructors are customarily conducted during the 11th week of the 14-week semester. This practice has been questioned by some faculty who feel that such an early measurement lacks validity. It is argued by some that semester-long projects (as well as other integrative teaching techniques) are a significant part of the learning experience and that by conducting evaluations three weeks before the end of the semester the benefits of those experiences are being systematically discounted.

The purpose of this study was to re-measure the student evaluations on the last day of the semester and compare those evaluations to those recorded during the eleventh week.

II. BACKGROUND

The body of research conducted on student evaluations of instructors and courses is too vast to be comprehensively reviewed in this paper. Cashin (1995) estimated over 1500 references on the subject as of that publication date. There are three extremely useful papers that summarize the stream of literature (Aleamori & Hexner, 1980; Cashin, 1988 and 1995) that dates back to the 1920’s (Remmers & Brandenburg, 1927; Brandenburg & Remmers, 1927; Remmers, 1928) and continues to the present (Stimpert, & Antonuccio, 2003).

Cohen (1981) and Feldman (1989b) found positive correlations between student learning and instructor ratings. The higher students performed on an external exam, the higher the instructor rating. Positive relationships were also found between instructor ratings by students and instructor self ratings (Feldman, 1989a; Marsh & Dunkin 1992) and alumni ratings (Overall & Marsh, 1980; Braskamp & Ory, 1994).

Research that specifically studies the timing of evaluations is limited. Cashin (1995) summarized that the timing of the evaluations was not significant. Frey (1976) compared the final exam performances of several sections of a course with the student ratings of the instructors, collecting the evaluations at various times during the semester. Half of the students
were contacted during the last week of classes, the other half during the first week of the following term. Frey concluded “the analyses of variance indicated that (1) the final exam performances of the several calculus sections were reliably different, (2) the instructional ratings reliably differentiated among the seven sections, and (3) the time of data collection (before the exam vs. after the exam) did not significantly affect the ratings.” Canaday, Mendelson and Hardin (1978) reported a study “designed to determine the effect of timing by administering a rating to equivalent random groups at three times.” Students were given evaluations right before the final examination, right after the final examination and after they received their grades. No significant differences were found. Feldman (1979) concluded that evaluations administered at any time during the second half of the term seemed to yield similar ratings.

III. HYPOTHESES

Consistent with the purpose of this study, there are ten individual hypotheses, one for each of the 10 questions on the evaluation instrument (see below.)

Ho: There is no difference between students’ evaluations during the eleventh week of the semester and on the last day of classes.
Ha: There is a difference between students’ evaluations during the eleventh week of the semester and on the last day of classes.

To test the hypotheses an alpha level of .10 was used. Although an alpha level of .05 is more commonly used for statistical testing, a level of .10 is within acceptable range (McDaniel and Gates) and is commonly used for managerial decision making.

IV. METHODOLOGY

Evaluation Instruments

The instrument used by the University for course evaluations is comprised of ten items for which the student is asked to rate the course on a 5-point scale: Well above average/Above average/Average/Below average/Well below average. The 10 questions are:

1. How would you rate the instructor’s willingness to help students?
2. How would you rate the instructor in terms of making students feel free to ask questions, disagree, express their ideas?
3. How would you rate the instructor’s communication skills?
4. How would you rate the instructor’s knowledge of the subject?
5. How would you rate the teaching methods (lectures, discussions, presentations) used in this course?
6. How clear were the objectives of this course?
7. How would you rate the evaluation and grading methods (tests, papers, oral presentations) in this course?
8. How would you rate the textbooks used in this course?
9. Overall, how would you rate the effectiveness of this instructor?
10. Overall, how would you rate your learning experience in this course?
For collection of data specifically for this study, the ten 5-point scale items were reproduced. Added were an introductory explanation of the study and three study-related questions. The first asked whether the student completed the earlier, official evaluation; the next two asked those who answered “yes” whether they thought their evaluation of a) the instructor and b) the overall learning experience was probably lower, higher or about the same as when they completed the “official” evaluation.

Data Collection

Data for the “official” evaluations were collected during the 11th week of two 14-week semesters (fall 2002 and spring 2003). The additional data used specifically for this study were collected at the end of the same two semesters. Data is from 14 sections of 9 different courses taught by 5 different professors. There were 306 responses, all usable. The data from the official evaluations represented 314 responses.

V. ANALYSIS AND DISCUSSION

Although the responses to the ten questions are technically categorical, they were treated as interval data, a common practice in marketing research. Aaker, Kumar and Day (1995) state: “… this assumption is controversial but is adopted because it permits more powerful methods of analysis to be used.”

To test the hypotheses ten separate t-tests were performed using EXCEL. The tests employed 2-tail, 2-sample equal variance parameters. Results are shown in Table 1.

Table 1. Summary of t-tests on all questions.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (E)</td>
<td>314</td>
<td>313</td>
<td>312</td>
<td>311</td>
<td>309</td>
<td>304</td>
<td>309</td>
<td>295</td>
<td>310</td>
<td>307</td>
</tr>
<tr>
<td>n (L)</td>
<td>306</td>
<td>305</td>
<td>306</td>
<td>303</td>
<td>305</td>
<td>302</td>
<td>305</td>
<td>289</td>
<td>305</td>
<td>306</td>
</tr>
<tr>
<td>Mean (E)</td>
<td>4.40</td>
<td>4.54</td>
<td>4.37</td>
<td>4.60</td>
<td>4.17</td>
<td>4.14</td>
<td>4.11</td>
<td>3.41</td>
<td>4.31</td>
<td>4.14</td>
</tr>
<tr>
<td>Mean (L)</td>
<td>4.31</td>
<td>4.42</td>
<td>4.25</td>
<td>4.50</td>
<td>4.07</td>
<td>4.02</td>
<td>4.01</td>
<td>3.38</td>
<td>4.26</td>
<td>4.11</td>
</tr>
<tr>
<td>Diff. (L-E)</td>
<td>-0.09</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.10</td>
<td>-0.10</td>
<td>-0.12</td>
<td>-0.10</td>
<td>-0.03</td>
<td>-0.05</td>
<td>-0.03</td>
</tr>
<tr>
<td>t-test, p =</td>
<td>0.13</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.15</td>
<td>0.09</td>
<td>0.16</td>
<td>0.70</td>
<td>0.43</td>
<td>0.72</td>
</tr>
<tr>
<td>Significance</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
<td>sig.</td>
</tr>
</tbody>
</table>

(E) = early in semester (the “official” evaluations)  
(L) = late in semester (data collected for this study on the last day of class)

Although not all differences are statistically significant it is interesting to note that the means for all questions are lower (less favorable) when measured late in the semester. This might suggest that there is some across-the-board negative bias at play, perhaps end-of-semester “burnout”. Another possibility might be an apprehension related to grades. Cohen (1981) found that students with higher grades tend to give higher evaluations so it could be that fear of a low grade might engender lower evaluations.
Students who stated that they completed the early evaluation were asked to rate whether their evaluations of the a) instructor and b) overall learning experience were lower, higher or about the same on the last day of class. As indicated in Table 2, the overwhelming majority (97% for instructor and 98% for learning experience) said that their evaluation was the same or higher on the last day. This contradicts the less favorable evaluations discussed above.

Table 2: Self-reported comparisons with earlier evaluation

<table>
<thead>
<tr>
<th>Evaluation is:</th>
<th>Instructor</th>
<th>Overall learning experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>probably lower at end</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>probably higher at end</td>
<td>10%</td>
<td>16%</td>
</tr>
<tr>
<td>probably the same</td>
<td>87%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Four of the ten course/instructor evaluative questions showed significant differences between the early and late measurements. They are:

Q2. How would you rate the instructor in terms of making students feel free to ask questions, disagree, express their ideas?
Q3. How would you rate the instructor’s communication skills?
Q4. How would you rate the instructor’s knowledge of the subject?
Q6. How clear were the objectives of this course?

VI. SUMMARY AND CONCLUSIONS

This data, viewed as a whole, clearly indicates that there are extraneous variables at play in student evaluations. Most conspicuous is that, when asked directly, students said their last-day evaluations were the same as or higher than during the 11th week, but the evaluations were, in fact, consistently lower.

The strong agreement between early and late measurement in questions 8, 9 and 10 suggest a high level of reliability by virtue of test-retest measurement. Nevertheless, the mean responses are consistently lower for these three as well as all other questions, so it is safe to conclude that student perceptions tend to be diminished by the last day of the semester.

The meaning is unclear when it comes to the specific questions that showed statistical significance. More questions are raised than answered in this regard. Results of the t-tests indicate that by the last day of the semester students have a less favorable perception of instructors’ willingness to answer questions or accept students’ ideas (question 2.) If considered in tandem with similar perceptions regarding the instructors’ communication skills (question 3), knowledge of the subject (question 4) and clarity of course objectives (question 6) we see a wide-ranging negative bias developing sometime during the last three weeks of the semester. It is very unlikely that instructors communication skills or knowledge would in fact decline during this period so it is clear that there are factors at work “beneath the surface” that tend to sour the students’ perceptions. One can almost imagine the students saying “Things are going wrong and it’s the instructor’s fault!”

Clearly, we should look for some additional insights by “massaging the data”, that is to say studying the data by course and by instructor. It would be useful to examine the data to determine whether this negative end-of-semester bias is instructor or course specific but, unfortunately, a caveat (see Limitations) with the participating faculty precludes such an investigation.
Perhaps the more meaningful findings are those that indicate no significant differences, especially questions 9 and 10, where the students rated the overall effectiveness of the instructor and the learning experience. Looking at the p-values for these two questions (plus question 8 which evaluates the textbooks) we see quite strong agreements between the samples.

**Limitations and Future Research**

This study was significantly restricted by a caveat with the instructors of the sections in which data were collected. In order to assure confidentiality the researchers assured that the data would be analyzed and reported in aggregate only, that there would be no analysis or reporting by section, course or instructor. This limitation prevented us from a closer examination of the puzzling differences in specific questions that might have led to more comprehensive understanding of the underlying influences.

Future research is suggested to confirm the generalizability of these findings across a wider range of instructors, disciplines and institutions. In addition, a closer inspection of the questions that yielded statistical significance is called for. Will these differences be found across all instructors and sections or do they occur only in certain settings? Permission of the participating faculty from this study would be required to use this data study for further analysis. Alternatively, new data can be collected without these restrictions.

In future studies additional questions should be included to probe more deeply students’ self-reported changes in perceptions. A series of Likert questions could be devised to measure the changes more indirectly than in this study.

**VII. REFERENCES**


