Can Faculty Change Initial Impressions on Student Evaluations of Teaching?

J. Ross Pruitt

Michael R. Dicks

Daniel S. Tilley*

* Graduate Research Assistant, Professor, and Professor, Department of Agricultural Economics, Oklahoma State University.

Contact: Michael R. Dicks, Oklahoma State University, Dept. of Agricultural Economics, 314 Agricultural Hall. Stillwater, OK 74078, Phone: (405)744-6163, Fax: (405)744-8210, e-mail: michael.dicks@okstate.edu

Abstract

Expectations of a course and instructor are formed prior to engagement in the course which affects learning. By understanding the factors that are involved in setting these expectations, instruction and student learning may be improved. This paper seeks to determine if student expectations set the basis for course and instructor appraisal and what factors affect these expectations or changes from expectations. Results indicate that while instructor appraisal does not change over the course of the semester, variables related to instructor appraisal do change. How worthwhile students view the course can also positively or negatively impact overall course appraisal.

Introduction

"To summarize, current evidence suggests that although grades and workload predict how students will evaluate their courses, this might not be because students are rewarding instructors for awarding high grades or for assigning low amounts of work. Quite the contrary, ratings might be an accurate reflection of how well a course is taught. Good teaching, in this view, leads to better learning, and this in turn leads to both good grades and high course ratings" (Remedios and Lieberman 2008).

The practice of students evaluating courses and instructors has a long and controversial history. Much of the research has focused on testing the validity of student evaluations of teaching (SET) as a result of instructor's concern over the use of the SET in faculty evaluations (Whitworth, Price, and Randall, 2002). We take a different approach in our analysis of the SET. If student expectations of a course and instructor are developed prior to engagement in the course and these expectations affect learning, then identifying and understanding the factors involved in setting these expectations can improve instruction and student learning.

This paper addresses two specific objectives: 1) to determine if student expectations set the basis for the overall course and instructor appraisal; and 2) to determine what factors affect these expectation or changes from expectations. This paper is not designed to provide an alternative to SET, but to provide a better understanding of the relative importance of factors that influence students' ranking of instructors and courses.

Literature Review

Students express their views on instructors and courses through SET. A key argument against the use of SET as a measure of teaching effectiveness is that students do not have adequate *a*

1

priori knowledge to critique instruction until after having been in the workforce for several years (Theall and Franklin, 2001). This argument is dismissed by Costin, Greenough, and Menges (1971) who point out that student ratings are stable across several years where fellow faculty members' evaluations of teaching are not. As Theall and Franklin (2001) point out, students are there and have experienced the full course experience. The research provided in Fleming, Bazen, and Wetzstein (2005) is particularly informative by indicating the impact of externalities (e.g. class time, size, and schedule) on SET scores. One of the research findings suggests that upperclassmen may resent taking introductory agricultural courses. Merritt (2008) and Widmeyer and Loy (1998) also found that externalities such as appearance and descriptions of the instructor have a direct impact on SET scores.

Wetzstein, Broder, and Wilson (1984) analyzed what students thought of their instructors and courses in introductory macroeconomics and intermediate microeconomics. This evaluation was conducted on the first day of class (prior to discussion of the course content) as well as the end of the term. Using a Bayesian method, the authors corrected for the reputation of the professor in the microeconomics course and demonstrated that the graduate student may have performed better in the macroeconomics course than the professor. This result was not immediately apparent given the unadjusted results which suggested both instructors performed at the same level.

Other studies have administered evaluation questionnaires in the early stages of the semester as well as the more traditional end of semester evaluation. Kohlan (1973) administered evaluations in selected classes at the end of the second class hour and again during the last week of the semester. Results of the study found that evaluations conducted early in the semester are stable across the semester. Kohlan (1973) suggests this may be due to the fact that little new

information regarding the ability of the instructor is presented after the first few classes and underlines the importance of positive early impressions.

Whitworth, Price, and Randall (2002) find that instructor's gender did affect the quality ratings received and that course evaluations cannot be compared across course category. The implication of the latter hypothesis is that comparisons cannot be made across different business disciplines and level of the course (e.g. graduate versus undergraduate or lower division courses and upper division courses). They concluded that administrators should refrain from comparing one instructor to another. McKeachie (1997) further validates this conclusion with the finding that student evaluations in lower level courses have lower validity than do evaluations in upper level courses. Germain and Scandura (2005) also call into question the construct validity of SET and discuss its relevance in greater detail. Furthermore, McKeachie (1979) concludes that evaluations have been linked to students' course grades. If this does occur, not only is the criterion contaminated as suggested by McKeachie (1979), but calls into question the timing of the evaluations. Simply put, there may not be an ideal time for student evaluations to occur. Development of a tool to account for grade inflation present in SET scores would increase validity, but would be difficult to apply in practice, given that evaluations can not necessarily be compared across courses, instructors, or years.

A limited number of studies have focused on timing of the actual evaluation instrument. However, these studies (Frey, 1976; Witt and Burdalski, 2003) have supported the effectiveness of the SET at the end of the semester. Frey (1976) divided students of introductory calculus classes into two subsets with one group evaluating the instructor prior to end of the term and the other half during the first week of the subsequent term, with the conclusion being that the results were "not reliably different." That finding helps to explain why SET scores are consistent across

time, i.e. students who are asked about the course a few years later feel the same as they did at the conclusion of the course (Costin, Greenough, and Menges 1971). Witt and Burdalski (2003) administered SET during the eleventh week of a fourteen week term with a follow-up evaluation on the last day. Results included significant differences in the evaluation of the instructor's ability to allow students to express their ideas and ask questions, communication skills and knowledge of the instructor, and the clarity of course objectives. Students in this study self-reported that opinions were no worse at the end of the semester although in the actual evaluation responses there were negative changes. Such findings do raise the question of whether students knowingly or unknowingly anchor their responses when given the same survey at different points in the semester.

Bejar and Doyle (1976) conducted an evaluation at the beginning and end of a summer semester. Initial evaluations were conducted on the first day of the course prior to the students seeing the instructors, with none of the 76 participating students knowing who the instructor was. Use of factor analysis showed that students were able to separate their expectations from the evaluations. However, the structures of the measured expectations and final evaluations were similar. Bejar and Doyle (1976) state this relationship might be the result of the learning process from previous instructors which is similar to McKeachie's (1997) statement regarding the lower validity of evaluations in lower division courses due to lack of a broad educational experience. Additionally, Bejar and Doyle (1976) note that the fact this research was conducted in the summer term might lead to more or less homogenous results compared to a regular semester.

Remedios and Lieberman (2008) also document expectations of students and compare results from before and after the semester among approximately six hundred students who enrolled in psychology courses at a Scottish university. Students were asked to complete the

4

questionnaire prior to registering for classes with the follow-up questionnaire being given during enrollment for the following term. Findings included grades, study hours, and perceived difficulty did have a marginally small impact on ratings. Courses where students felt involved including being stimulating, interesting, and useful largely determined course ratings (Remedios and Lieberman 2008). Remedios and Lieberman (2008) also found students are sensitive to different qualities of courses and are not likely to rate all aspects highly if one area of the course is rated highly.

The literature is resplendent with analyses that conclude that expectations affect the SET, but few identify or measure the relative impact of factors that affect expectations or changes between expectations and final SET. This study first determines if expectations and final evaluation differs for individual characteristics (e.g. instructor presentation of material) and overall instructor and course scores. We then measure the importance of several factors (e.g. change in grade) in changing scores of the characteristics.

Conceptual Model

At Oklahoma State University (OSU), students evaluate their instructor as well as the course on several different factors in each category which are shown in table 1. Merritt (2008) has suggested that students have formed expectations (opinions) of both course and instructor prior to the first day of class or within the first meetings. These expectations may be formed from other students, websites, professors, and/or advisors in addition to the student's own prior interaction with the instructor.

Actual experiences in the classroom may or may not alter the student's expectations of the course and/or instructor. No change between expectation and actual experience of the

course or instructor would indicate that information obtained from various sources of the instructor/course was consistent with the actual experiences, that is the expected utility from the course $E(U_{initial})$ is equal to the actual utility received $E(U_{final})$.

We hypothesize that students experience in the classroom and with the instructor are consistent with their expectations. More succinctly,

(1) $\Delta E(U) = E(U_{final}) - E(U_{initial}) = 0$

The expected utility is observed through ratings of the overall instructor appraisal (score) as well as if the student viewed the course as a good course (or course score).

Students have three options regarding their opinions on instructor and course scores across the semester: increase, decrease, or no change. This may oversimplify the issue as students who initially rate courses/instructors the lowest (highest) cannot rate courses/instructor any lower (higher). However, a principle issue for this research is to determine what factors lead to a student changing their evaluations of courses and instructors. Variability (or lack thereof) in instructor score is hypothesized to be a function of instructor related variables ("preparation," "effort," "presentation," "knowledge," "explain," and "attitude") and course score as a function of course related variables ("workload," "assignments," "tests," "adequately involved," and "worthwhile"). The probability that a student will choose option j (increase appraisal, decrease appraisal, or no change in appraisal) can be calculated by;

(2)   Prob (option j) $= \dfrac{e^{V_j}}{\sum e^{V}}$

where $V_j = XB$ is a vector of appropriate independent variables as listed in the preceding paragraph for the model (course or instructor score) and their associated parameter estimates.

Questions regarding the instructor were rated on a zero to four-point scale while questions regarding the course and additional questions, not typically asked on the standard OSU

evaluation, were asked on a zero to five-point scale. The latter can be compared to the standard OSU evaluation that asks students to rate courses on a zero to four-point scale. However, the standard evaluation puts "not applicable" in the middle of the ratings whereas our evaluation had "not applicable" at the end of all answer choices and undecided in the middle. Students who responded to questions with "not applicable" at the beginning of the semester had those responses treated as undecided. We assumed that a "not applicable" response implied that the student was not able to form an expectation. However, responses of "not applicable" at the end of the semester were considered to be irrelevant to the course and were discarded. The change in the rating for an instructor is calculated as the difference between the initial and final response as shown in equation (1). Only the variables mentioned in the preceding paragraph were included to determine the factors that affect the changes in instructor or course appraisal that were recorded in an attempt to identify the factors that most directly impact changes in instructor or course appraisal over the course, if changes do in fact occur.

The significant instructor and course related variables are hypothesized to be affected by other collected variables. These other variables can be viewed as the externalities that are examined in Fleming, Bazen, and Wetzstein (2005) and include the degree to which students view the instructor as fair, whether or not the classroom affects impressions of the course and instructor, whether the students likes to ask/answer questions during class time, and time of day the course meets. The impact of changes in grade expectations and attendance also are factors that are examined. Questions regarding how long the student has attended OSU (including the current semester) as well as how many the student accumulated at an institution other than OSU-Stillwater were also included on the final evaluation procedure as transfer students may form and change expectations regarding courses and instructors differently than students who have

completed all coursework at OSU. The number of hours transferred to OSU was asked in ten

hours blocks (0-10, 11-20, etc.) with the final category being more than 60 hours.

It is hypothesized that variables related to the student will not change over the course of

the semester (i.e. a student will continue to be a student in the college of agriculture, the course

will continue to be required, etc.). Additional information from the semester will impact views

of the student and may alter their responses to questions provided on the evaluation instrument.

Variables such as whether or not the instructor presents material in a manner appealing to

students (presentation), or students' views of whether or not the course is worthwhile may

significantly affect the overall instructor or course appraisal

Concern over whether a student would knowingly try to anchor question responses (i.e.

answer a question consistently across the semester), led to the development of an additional,

"control" questionnaire. Students who completed this separate questionnaire were asked to

evaluate all instructors and courses they were enrolled in other than the class they were currently

in. Comparison of the control and non-control students will be used to determine the existence

of anchoring.

Data

Twenty two courses in the College of Agricultural Sciences and Natural Resources (CASNR)

participated in this research. Participating classes were from the departments of animal science,

agricultural economics, agricultural communication, education, and leadership, plant and soil

science, horticulture, and natural resource ecology and management. Of the twenty-two courses,

one course was being offered for the first time by a full professor and two courses being taught

by an instructor for the first time. There were two freshmen, four sophomore, nine junior, and

seven senior level courses. Seventeen instructors participated with nine of those being full professors, two associate professors, five assistant professors, and one graduate student instructor.

The proctor was introduced by the instructor of the course prior to the evaluations being distributed. Two proctors were used at the beginning of the semester due to some courses evaluating instructors at the same time, while only one proctor was used at the end of the semester to ensure continuity among verbal instructions. The proctors were not involved with the participating courses. Evaluations were completed within the first two weeks of the fall 2007 semester with the time of the evaluation being determined by the instructor to allow for the least amount of intrusion to the instructor. The final round of evaluations was conducted from November 15th through December 7th. Students, on average, completed the questionnaire in fifteen minutes at both points in the semester.

Students were assigned an individual identification code that would identify their responses at the beginning and end of the semester. The code number was a five digit alphanumeric code based on information only known by the student. The first digit was the first letter of the high school where the student graduated. Digits two and three were the student's birth month (January was 01, February as 02, etc.) with the final two digits being the last two digits of the student identification number. Problems arose at the end of the semester as students did not remember the last two digits as Oklahoma State assigns multiple identification numbers for students. The proctor encouraged students at the end of the semester to list multiple code numbers if the student was unable to remember the last two digits used at the beginning of the semester. This allowed several questionnaires to be successfully matched with evaluations at the beginning of the semester.

An informational cover sheet was included that listed the title of the research, a student's rights as a research volunteer, how to determine their individual identification code, and that the research would occur twice in the semester. The proctor did not announce that students would be given an additional opportunity to evaluate the instructor and course towards the end of the semester until the initial evaluation was completed. This was done in an attempt to have the student feel this was an actual evaluation. Students were encouraged to keep the cover sheet in case they had questions later as the sheet had contacts of persons who could answer those questions. The informational cover sheet also indicated the confidentiality of all responses that included the fact that instructors would not see the results until after grades had been submitted.

A total of 867 responses were collected initially in addition to the 155 control questionnaires. A total of 897 evaluations were completed at the end of the semester. Of these, 423 evaluations were successfully matched by identification code numbers with an additional 67 responses successfully matched to a control questionnaire from earlier in the semester. Differences among students completing the evaluation were most obvious in the percentage of males completing the evaluation instruments. Approximately half of all unmatched initial evaluations were completed by males, while this percentage fell to forty percent in evaluations that were successfully matched.

Results

Comparison of means among the different subsets of completed evaluations is included in tables 2 through 4 (by all courses, upper division courses, and lower division courses). The overall instructor appraisal did not change. However, other instructor related variables were significantly different over the course of the semester contrary to the existing literature on SET

(denoted by superscripts in table 2 through 4). More importantly, the results change when the overall instructor appraisal is segregated between courses at the upper and lower division level. The overall appraisal of instructors in upper division courses decreased while it increased in lower division courses

Cross tabulations were also calculated for course and instructor variables and shown in tables 5 through 6 (initial ratings are in the rows with final ratings in the columns). These tables show the distribution of changes by direction and magnitude of the change rather than the mean change. Students who answered course related questions with "not applicable" in the beginning of the semester had those answers grouped as undecided while "not applicable" responses at the end of the semester were treated as non-responses. Chi-square tests were also conducted with results showing that the distribution of scores had significantly changed over the course of the semester for instructor and course related variables. Of the 423 evaluations collected in all courses, 56 students decreased their overall instructor appraisal rating while 231 did not change, and 136 increased their opinions of instructor appraisal. This can be compared to overall course appraisal which had 31 students decrease their ratings, 154 exhibiting no change, and 234 students increasing opinions of overall course appraisal among 419 observations. However, because the "not applicable" response was scored as a zero in the initial data for the overall course rating this would provide an increase in overall course rating at the end of the semester.

Changes in means for course variables were generally the same regardless of whether all, upper division, or lower division courses were analyzed. This finding should not be surprising due to a lack of information available to students when the initial evaluation was conducted. Students may also realize the difference in abilities between friends who give them information

11

on courses and instructors prior to enrollment. Changes in student's views of course variables (accuracy of syllabus, entertaining instructor, distractions in classroom, etc.) were also significantly different across the semester. Differences do arise by the level of the course which is often masked when participating courses are aggregated. Expected attendance as reported at the beginning of the semester seems to be understated while students either overestimated their abilities, underestimated course difficulty, or some combination of these two factors led to decreased expected grades at the end of the semester.

A pooled means test for matched student evaluation responses shows that students' mean evaluations of course and instructor related variables do change during a semester. Students in upper division courses typically overrated overall instructor appraisal while students in lower division courses underestimated overall instructor appraisal. Student views including their ability to maintain focus in the classroom, distractions that occur in the classroom, and whether the instructor is able to actively involve students positively changed across the semester regardless of division.

As previously mentioned, a control questionnaire was developed to test for the presence of anchoring by students to their initial responses. The null hypothesis was that responses would not be significantly different (in terms of the mean) between students who evaluated courses and instructors twice and students who only evaluated the course and instructor once, i.e. H_0: $\mu_1 = \mu_2$, where μ_1 is the mean of the final evaluation among students whose evaluations were matched and μ_2 is the mean from matched control questionnaires. A difference in means test was conducted for all course and instructor variables. Variables that were significantly different in the control group relative to end of semester matched evaluations are also included in tables 2 through 4. Control questionnaires were distributed in classes with at least one hundred students,

care should be taken with these results since only 67 responses (11 in upper division courses) were successfully matched across the semester compared to the 423 evaluations successfully matched across the semester. A further note regarding these findings is that students may not have fully understood the instructions leading to the failure to reject the null hypothesis in many cases. In many cases, anchoring does not seem to be present among these two groups. However, students who completed the control question typically had been at OSU fewer semesters and transferred fewer hours from another institution to OSU. Students completing the questionnaire in the lower division courses did not rate variables significantly different than their counterparts who completed two evaluations. Results from the control questionnaire are biased to larger one and two thousand division courses which prohibit meaningful conclusions being drawn in upper division courses.

Multinomial Logit Modeling

Given that students generally do change opinions on instructor and course related variables after the first two weeks of classes, the reasons for the change were evaluated. This was done by a multinomial logit that was estimated in PROC CATMOD in SAS 9.1. An ordered probit model would have been an appropriate model to use as well, but given the distribution of the available matched evaluation data a multinomial logit was employed. The data could change in discrete units from ±4 given that the rating system employed allowed students to rate instructors/courses from zero to four scale. Changes in variables related to instructors approximated a normal distribution with the majority of responses between -1 and 2. This compares to the course related variables having the majority of responses between -1 and 4. The nature of the data allowed for estimation of models based on students who did not change, decreased, or increased

13

their appraisal of the instructor. Estimation of the multinomial logit in this instance was less cumbersome than the ordered probit where some levels of change were not observed.

Likelihood ratio tests were conducted to determine the appropriateness of the pooled model versus separate models for the upper and lower division courses. The test statistic for the instructor appraisal model was 14.07 as there were seven degrees of freedom while the course appraisal model was 12.59 due to only six degrees of freedom. Both pooled models were rejected and the models for the upper and lower division courses were favored. These results are shown in tables 7 and 10. Due to the marginal effects of a multinomial logit not being equal to the parameter estimates, these were calculated and are shown in tables 9 and 12 for the instructor and course models, respectively. Marginal effects for a multinomial logit are calculated as

$$(4) \qquad \frac{\partial P_j}{\partial x_i} = P_j [\beta_j - \overline{\beta}]$$

where P_j is the probability of the jth alternative and β_j is the parameter estimate of the jth alternative and $\overline{\beta}$ is the sum of parameter estimates times the probability for all alternatives (Greene 2003). The probability of the jth alternative is calculated as $P_j = \frac{e^{\beta_j ' x_i}}{1 + \sum_{k=1}^{2} e^{\beta_k ' x_i}}$ is adapted from Greene (2003). As there are three alternatives in the model (increase, decrease, no change) the multinomial logit model produces parameter estimates for two models whose coefficients are relative to the omitted model, which is no change in this case. The numeral one which appears in the probability equation is a result of the parameters that are not estimated in PROC CATMOD. Additionally, the probability for no change in instructor/course appraisal would have a numeral one in the numerator instead of the product of the parameter estimates for the jth alternative. Marginal effects are calculated at the mean for all independent variables. The marginal effects shown in these tables are expected percentage changes given a one unit increase

in an independent variable, *ceteris paribus*. The maximum amount of change for any student was ±4. Therefore, a student in an upper division course who increased their presentation score by 4 is 80% less likely not to have decreased their instructor appraisal score. Marginal effects for all courses are reported even though the models for upper and lower division courses are preferred.

"Presentation" was a key factor in explaining change in instructor scores. "Attitude" was also important for upper division courses leading to instructor who increased their rating in this respect were more likely to have a positive rating. Instructor's ability to "explain" material was only significant in explaining a decrease in instructor appraisal relative to no change in instructor appraisal regardless of course division.

Results vary by course level in explaining the change in course scores. Key factors that led to positive changes in course score for lower division courses were the "workload" of the course, "adequate involvement" of students, and how "worthwhile" they feel the class is. This may be contrasted with students in upper division courses who were concerned about "tests" and how "worthwhile" the class was viewed in explaining a positive change in course score. This finding on the "worth" of the class is consistent with Remedios and Lieberman (2008). These factors were the same in explaining a decrease in course appraisal among students in upper division courses, but no factors were significant in explaining changes among students in lower division courses.

Additional models were estimated to determine the factors that led to changes (positive, no change, or decrease) in each of the instructor and course related[1]. Marginal effects were also estimated for these variables affecting a positive/negative change in the dependent variable. There were no variables that significant in each model estimated as well as differences existed

---

[1] Available from the authors upon request.

between course divisions. The number of hours transferred to OSU was significant in several of the models estimated in explaining changes in the various dependent variables. However, marginal effects calculated for this variable were typically small (less than 5%) suggesting that increasing the number of hours transferred to OSU (in 10 hour blocks) would have a negligible impact on changing the probability of a positive/negative increase in the dependent variable. Marginal effects on the whole across all models explaining change in the instructor/course related variables were typically less than ten percent.

Conclusion

Students base course and instructor expectations on a multitude of sources including friends, professors, web resources and actual experiences with the instructor. These expectations are important in determining the final assessment of the course and instructor but are challenged and may change over the course of the semester. Instructors that understand the factors that form these expectations may learn to influence these factors to improve the student experience and increase overall course and instructor ratings.

Results from the comparison of matched evaluations suggest that on the whole, instructor appraisal does not change over the semester. However, these finding masks the difference between students in upper division and lower division courses. Upper division students tended to decrease their evaluations while lower division students tended to increase their evaluations.

Students do change their expectations of instructor related variables (their "preparation", "presentation" of material, ability to "explain" material, etc.) over the course of the semester. This is contrary to Merritt's (2008) assertion that evaluations measure snap judgments that occurred at the beginning of the semester. Additional evidence in our research that supports our

findings is that while their was no significant difference between control and non-control students for overall instructor rating, differences did occur between the groups for two factors of instruction, presentation of material and adequate involvement of students. This finding however needs further research as only classes with initial enrollments of at least one hundred students were selected to be part of the control group. This fact may be the reason that students rated presentation of material and adequate involvement of students differently. At the very least, instructors of large, often introductory courses are seemingly not negatively impacted by the size of the course, especially on instructor appraisal.

Consistent across the semester (whether first impressions or comparison results across the semester), was the need for instructors to "present" material in an effective and engaging way. The magnitude of this coefficient at both points in the semester dwarfed coefficients of other instructor related variables. This may be a key way to improve overall ratings of instructor appraisal or performance. Instructors who want to maintain high levels of instructor appraisal throughout the semester should focus on presenting material effectively and taking the time to carefully explain material. Similarly, the ability of instructors to motivate students to see the "worth" in the course is an important factor that determines overall course appraisal. Only one instance occurred where its parameter estimate was not the largest ("workload" in lower division courses explaining a positive change relative to no change in overall course appraisal).

The research contained in this paper identifies several factors that may lead to changes in overall instructor/course rating. For lower division courses the factors "presentation" and "explain" significantly lowered instructor expectation while only "presentation" significantly increased instructor expectation. This can be contrasted with upper division courses where "presentation" and "attitude" significantly increased expectations of the instructor while

"presentation" and "explain" decreased expectations of the instructor. Results from course rating models show consistently across course divisions that the worth of the class is extremely important in either decreasing or increasing course expectations. The exception to this is in negative expectation changes in lower division courses, but this is likely due to a lack of observations where students decreased course expectations.

However, the paper falls short in identifying the characteristics considered by the students in evaluating these factors. For instance, the factor "presentation" had a significant effect on changes in student expectations but what we can not identify, or measure the effect of, is the characteristics of "presentation" that students use in rating this factor. Identifying and quantifying the importance of these characteristics may increase student expectations, learning, and overall course and instructor score. Where this research falls short is determining what students mean when they read the statement "presentation of material", "effort devoted to teaching", and "this course is worthwhile to me". While some of the questions asked on the SET may be viewed as straightforward, students may interpret the question in a different way leading to answers that are not as straightforward as they appear to be.

It is entirely possible that results presented herein are not indicative of evaluations campus wide at Oklahoma State University. The fact that having students evaluate instructor twice may have lead to students consider responses at the end of the semester which may not be indicative of typical evaluations conducted at only the end of the semester. Concerns over this fact can be dismissed given the results Wetzstein, Broder, and Wilson (1984) present showing that while instructor reputation may drive initial impressions what is done by the instructor does impact final evaluations given that a graduate student outperformed a known professor once

accounting for reputation. Yet the underlying result of this study is that you can change

students' minds on instructor related variables over the course of the semester.

References

Bejar, I.I., and K.O. Doyle. 1976. "The Effect of Prior Expectations on the Structure of Student Ratings of Instruction." 13:151-54.

Costin, F., W.T. Greenough, and R.J. Menges. 1971. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." *Review of Educational Research*. 41:511-35.

Fleming, R.A., E.F. Bazen, and M.E. Wetzstein. 2005. "Measuring the Impact of Externalities on College of Agriculture Teaching Evaluations." *Journal of Agricultural and Applied Economics*. 37:635-645.

Frey, P.W. 1976. "Validity of Student Instructional Ratings: Does Timing Matter?" *Journal of Higher Education*. 47:327-36.

Germain, M-L., and T.A. Scandura. 2005. "Grade Inflation and Student Individual Differences as Systematic Bias in Faculty Evaluations." *Journal of Instructional Psychology*. 32:58-67.

Greene, W.H. 2003. *Econometric Analysis*. Upper Saddle River NJ: Prentice Hall.

Kohlan, R.G. 1973. "A Comparison of Faculty Evaluations Early and Late in the Course." *Journal of Higher Education*. 44:587-595.

McKeachie, W.J. 1979. "Student Ratings of Faculty: A Reprise." *Academe*. 65:384-397.

McKeachie, W.J. 1997. "Student Ratings: The Validity of Use." *American Psychologist*. 52:1218-1225.

Merritt, D. J. 2008. "Bias, the Brain, and Student Evaluations of Teaching." *St. John's Law Review*. 82:235-87.

Remedios, R., and D.A. Lieberman. 2008. "I Liked Your Course Because You Taught Me Well: The Influence of Grades, Workload, Expectations, and Goals on Students' Evaluations of Teaching." *British Educational Research Journal*. 34:91-115.

Theall, M., and J. Franklin. 2001. "Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?" *New Directions for Institutional Research*. 109:45-56.

Whitworth, J.E., B.A. Price, and C.H. Randall. 2002. "Factors that Affect College of Business Student Opinion of Teaching and Learning." *Journal of Education for Business*. 77:282-289.

Widmeyer, W.N., and J.W. Loy. 1988. "When You're Hot, You're Hot! Warm-Cold Effects in First Impressions of Persons and Teaching Effectiveness." *Journal of Educational Psychology*. 80:118-21.

Witt, J., and D. Burdalski. 2003. "Regarding the Timing of Student Course/Instructor Evaluations." *Journal of the Academy of Business Education*. Proceedings Issue. Available at http://www.abe.villanova.edu/proc2003/witt.pdf.

Table 1. Questions asked on Actual Oklahoma State University's SET form.

Variable Name	Question
Prep[a]	Preparation and effort
TeachEffort[a]	Effort devoted to teaching
Present[a]	Presentation of material
Knowledge[a]	Knowledge of subject
Explain[a]	Ability to explain subject matter
Attitude[a]	Positive attitude toward students
InstrOverall[a]	Overall INSTRUCTOR appraisal
Workload[b]	The workload is appropriate for the hours of credit
Assignments[b]	Assignments are relevant and useful
Tests[b]	Testing and evaluation procedures are good
Involve[b]	Students are adequately involved
Worthwhile[b]	This course is worthwhile to me
CourseOverall[b]	Overall, this is a GOOD course
Syllabus[b]	The syllabus is an active reflection of the course experience
ActiveInvolve[b]	The instructor is able to actively involve me in class
Entertain[b]	The instructor is entertaining
Ask[b]	I don't like to ask questions during class time
Answer[b]	I don't like to answer questions during class time
Fair[b]	The instructor treats students fairly
CalledOn[b]	I don't like to be called on during class time
Focus[b]	I am able to maintain focus in class
Visualaids[b]	Learning in this class is aided by charts, graphs, and presentations
Stories[b]	Learning in this class is aided by stories, games, and real world applications
Classroom[b]	The classroom negatively impacts my perception of the course and instructor
Distract[b]	Distractions from other students negatively impact my perception of the course
ExpAttendance[c]	I expect to miss the following number of classes
ExpGrade	I expect my grade to be
TransferredHours[d]	I transferred the following number of hours to OSU
OSU Semesters	I have been at OSU this many semesters including the current semester

[a] Options were Very High, High, Average, Low, or Very Low.
[b] Options were Definitely Yes, Yes, Undecided, No, Definitely No, or Not Applicable.
[c] Options were 0 to 2 classes, 3 to 4 classes, 5 to 7 classes, or more than 7 classes.
[d] Options were 0 hours, 1 to 10 hours, 11 to 20 hours, 21 to 30 hours, 31 to 40 hours, 41 to 50 hours, 51 to 60 hours, or more than 60 hours

Table 2. Comparison of Means for All Collected Questionnaire Types for All Courses.

Variable	Matched initial Evaluations		Matched final Evaluations		Matched (Final) Controls	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Dependent Variables						
InstrOverall	3.504	0.708	3.563	0.685	3.687	0.556
CourseOverall	1.871	1.693	3.216[a]	1.065	3.269	0.750
Student Characteristics						
Transferred Hours	---	---	2.396	2.607	1.418[b]	1.932
OSU Semesters	---	---	4.031	4.610	2.075[b]	1.627
Instructor Characteristics						
Prep	3.322	0.735	3.532[a]	0.677	3.687[b]	0.528
TeachEffort	3.423	0.694	3.603[a]	0.641	3.716	0.486
Present	3.390	0.693	3.284[a]	0.873	3.507[b]	0.637
Knowledge	3.165	0.792	3.667[a]	0.638	3.761	0.495
Explain	3.550	0.651	3.418[a]	0.825	3.582[b]	0.581
Attitude	3.322	0.735	3.645[a]	0.647	3.537	0.876
Course Characteristics						
Workload	1.660	1.677	3.173[a]	0.958	3.25	0.66
Assignments	1.650	1.684	3.168[a]	1.024	3.30	0.76
Tests	1.043	1.548	3.088[a]	1.087	3.14	0.76
Involve	1.969	1.657	3.242[a]	0.916	2.91[b]	1.06
Worthwhile	2.069	1.668	3.105[a]	1.137	3.04	1.17

Student Views

Syllabus	1.771	1.687	3.126[a]	1.107	3.22	0.97
ActiveInvolve	1.995	1.662	2.988[a]	1.151	2.68[b]	1.30
Entertain	2.525	1.582	3.211[a]	1.088	3.34	1.02
Ask	1.756	1.392	2.055[a]	1.278	2.12	1.29
Answer	1.704	1.342	2.184[a]	1.318	2.14	1.48
Fair	2.327	1.611	3.308[a]	0.899	3.24	0.91
CalledOn	1.842	1.386	2.209[a]	1.354	2.20	1.49
Focus	2.246	1.456	2.888[a]	1.136	2.73	1.23
Visualaids	1.962	1.613	3.105[a]	0.939	3.12	0.86
Stories	2.036	1.649	2.981[a]	1.131	3.11	0.92
Classroom	1.286	1.042	1.664[a]	0.963	1.56	0.92
Distract	1.411	1.183	1.787[a]	1.092	1.83	1.06
ExpAttendance	0.196	0.474	0.484[a]	0.750	0.44	0.71
ExpGrade	3.771	0.508	3.438[a]	0.691	3.28	0.73

[b] Mean is significantly different from corresponding mean in matched initial evaluation column at the 5% level
[c] Mean is significantly different from corresponding mean in final matched evaluation column at the 5% level

Table 3. Comparison of Means for All Collected Questionnaire Types in Upper Division Courses.

Variable	Matched initial Evaluations		Matched final Evaluations		Matched (Final) Controls	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Dependent Variables						
InstrOverall	3.654	0.552	3.538[a]	0.742	3.909[b]	0.302
CourseOverall	2.159	1.701	3.210[a]	1.154	3.545	0.522
Student Characteristics						
Transferred Hours	---	---	2.752	2.781	3.091	2.773
OSU Semesters	---	---	5.150	4.402	3.800[b]	1.317
Instructor Characteristics						
Prep	3.410	0.713	3.509	0.737	3.818[b]	0.405
TeachEffort	3.491	0.650	3.568	0.704	3.818	0.405
Present	3.521	0.623	3.261[a]	0.952	3.818[b]	0.405
Knowledge	3.256	0.777	3.598[a]	0.730	3.818	0.405
Explain	3.592	0.581	3.376[a]	0.915	3.727[b]	0.467
Attitude	3.410	0.713	3.654[a]	0.658	3.818	0.405
Course Characteristics						
Workload	2.090	1.659	3.185[a]	0.998	3.455	0.688
Assignments	2.090	1.674	3.202[a]	1.037	3.182	1.168
Tests	1.356	1.668	3.096[a]	1.163	3.100	0.738
Involve	2.487	1.520	3.352[a]	0.869	3.000	1.183
Worthwhile	2.433	1.612	3.150[a]	1.174	3.091	1.221

Student Views

Syllabus	2.157	1.646	3.065[a]	1.173	3.545	0.522
ActiveInvolve	2.551	1.485	3.129[a]	1.147	3.091	1.136
Entertain	3.069	1.247	3.215	1.151	3.727	0.467
Ask	1.944	1.340	2.100	1.204	2.400	1.174
Answer	1.919	1.299	2.173[a]	1.279	2.400	1.174
Fair	2.824	1.367	3.378[a]	0.827	3.727	0.467
CalledOn	2.056	1.317	2.219	1.310	2.400	1.174
Focus	2.601	1.280	2.948[a]	1.127	3.273	0.647
Visualaids	2.405	1.444	3.083[a]	1.039	3.500	0.527
Stories	2.575	1.461	3.132[a]	1.037	3.545	0.522
Classroom	1.397	1.023	1.636[a]	1.050	1.800	1.317
Distract	1.598	1.120	1.737	1.126	1.800	1.135
ExpAttendance	0.197	0.449	0.447[a]	0.697	0.700	0.949
ExpGrade	3.812	0.413	3.539[a]	0.643	3.636	0.674

[b] Mean is significantly different from corresponding mean in matched initial evaluation column at the 5% level
[c] Mean is significantly different from corresponding mean in final matched evaluation column at the 5% level

Table 4. Comparison of Means for All Collected Questionnaire Types in Lower Division Courses.

Variable	Matched initial Evaluations		Matched final Evaluations		Matched (Final) Controls	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Dependent Variables						
InstrOverall	3.317	0.828	3.593[a]	0.609	3.643	0.586
CourseOverall	1.513	1.618	3.222[a]	0.947	3.214	0.780
Student Characteristics						
Transferred Hours	---	---	1.946	2.298	1.089[b]	1.552
OSU Semesters	---	---	2.651	4.495	1.804[b]	1.482
Instructor Characteristics						
Prep	3.212	0.749	3.561[a]	0.595	3.661	0.549
TeachEffort	3.339	0.738	3.646[a]	0.552	3.696	0.502
Present	3.228	0.741	3.312	0.767	3.446	0.658
Knowledge	3.053	0.797	3.751[a]	0.491	3.750	0.513
Explain	3.497	0.727	3.471	0.696	3.554	0.601
Attitude	3.212	0.749	3.635[a]	0.635	3.482	0.934
Course Characteristics						
Workload	1.127	1.545	3.159[a]	0.909	3.214	0.653
Assignments	1.106	1.533	3.127[a]	1.008	3.327	0.668
Tests	0.656	1.290	3.079[a]	0.989	3.143	0.773
Involve	1.328	1.597	3.106[a]	0.956	2.893	1.039
Worthwhile	1.619	1.628	3.048[a]	1.091	3.036	1.175

Student Views

Syllabus	1.314	1.623	3.201[a]	1.017	3.161	1.023
ActiveInvolve	1.307	1.615	2.814[a]	1.134	2.600	1.328
Entertain	1.851	1.693	3.206[a]	1.008	3.268	1.087
Ask	1.521	1.423	2.000[a]	1.364	2.073	1.317
Answer	1.439	1.350	2.198[a]	1.367	2.091	1.531
Fair	1.714	1.680	3.222[a]	0.975	3.143	0.943
CalledOn	1.577	1.426	2.196[a]	1.410	2.164	1.549
Focus	1.810	1.542	2.815[a]	1.145	2.625	1.287
Visualaids	1.418	1.647	3.133[a]	0.800	3.055	0.891
Stories	1.367	1.628	2.797[a]	1.214	3.019	0.961
Classroom	1.148	1.051	1.698[a]	0.847	1.519	0.841
Distract	1.180	1.220	1.849[a]	1.047	1.830	1.051
ExpAttendance	0.196	0.504	0.529[a]	0.809	0.389	0.656
ExpGrade	3.720	0.602	3.312[a]	0.728	3.214	0.731

[a] Mean is significantly different from corresponding mean in matched initial evaluation column at the 5% level
[b] Mean is significantly different from corresponding mean in final matched evaluation column at the 5% level

Table 5. Cross tabulation of overall instructor appraisal across the semester in all courses.

	End of Semester					
	Very Low	Low	Average	High	Very High	Total
Very Low	0	0	0	0	2	2
Low	0	0	0	3	0	3
Average	0	2	3	7	20	32
High	0	2	10	55	62	129
Very High	1	2	10	52	192	257
Total	1	6	23	117	276	423

Table 6. Cross tabulation of overall course appraisal across the semester in all courses.

	End of Semester					
	Undecided	Definitely No	No	Yes	Definitely Yes	Total
Not applicable/ Undecided	17	4	9	81	73	184
Definitely No	0	0	0	0	0	0
No	0	0	1	0	0	1
Yes	10	0	3	75	67	155
Definitely Yes	1	0	1	16	61	79
Total	28	4	14	172	201	419

Table 7. Multinomial Logit Estimates of a Positive Change in Instructor Score Relative to a No Change in Instructor Score.

Variable	All Courses		Upper Division Courses		Lower Division Courses	
	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
Intercept	-2.391***	0.257	-3.256***	0.470	-1.857***	0.346
Prep	0.112	0.307	0.697	0.532	-0.278	0.387
TeachEffort	0.660**	0.324	0.812	0.507	0.364	0.457
Present	1.545***	0.280	0.861**	0.426	1.944***	0.411
Knowledge	0.574**	0.259	0.290	0.396	0.775	0.361
Explain	0.326	0.239	0.455	0.356	0.315	0.323
Attitude	0.293	0.234	0.885**	0.417	0.080	0.300
N	422		233		189	
-2LL	500.009		229.970		241.247	

Note: Three asterisks denote significance at the 1% level, two asterisks at the 5% level, and one asterisk at the 10% level.
A log likelihood ratio test favored separate estimated models for the upper and lower division courses over the pooled (all courses) model.

Table 8. Multinomial Logit Estimates of a Negative Change in Instructor Score Relative to No Change in Instructor Score.

Variable	All Courses		Upper Division Courses		Lower Division Courses	
	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
Intercept	-2.453***	0.257	-2.733***	0.364	-2.374***	0.415
Prep	-0.272	0.281	-0.298	0.431	-0.408	0.397
TeachEffort	-0.234	0.309	-0.014	0.421	-0.416	0.533
Present	-1.542***	0.250	-2.154***	0.380	-1.039***	0.380
Knowledge	0.158	0.231	-0.036	0.318	0.424	0.357
Explain	-0.888***	0.224	-1.047***	0.301	-0.790**	0.368
Attitude	0.001	0.215	-0.543	0.332	0.314	0.290
N	422		233		189	
-2LL	500.009		229.970		241.247	

Note: Three asterisks denote significance at the 1% level, two asterisks at the 5% level, and one asterisk at the 10% level.
A log likelihood ratio test favored separate estimated models for the upper and lower division courses over the pooled (all courses) model.

Table 9. Marginal Effects of a Change in Instructor Score for All Course Types.

Variable	All Courses			Upper Division Courses			Lower Division Courses		
	No Change	Decrease	Increase	No Change	Decrease	Increase	No Change	Decrease	Increase
Prep	0.011	-0.023	0.013	0.004	-0.030	0.026	0.065	-0.023	-0.042
TeachEffort	-0.038	-0.025	0.063	-0.025	-0.004	0.030	-0.038	-0.034	0.071
Present	-0.016	-0.140	0.156	0.161	-0.201	0.040	-0.260	-0.102	0.362
Knowledge	-0.059	0.008	0.051	-0.006	-0.004	0.011	-0.146	0.015	0.130
Explain	0.038	-0.076	0.038	0.077	-0.098	0.021	-0.011	-0.058	0.069
Attitude	-0.024	-0.003	0.027	0.019	-0.053	0.034	-0.029	0.020	0.009

Table 10. Multinomial Logit Estimates of a Positive Change in Overall Course Score Relative to No Change in Overall Course Score.

Variable	All Courses		Upper Division Courses		Lower Division Courses	
	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
Intercept	-1.298***	0.226	-1.317***	0.286	-1.273***	0.395
Workload	0.215**	0.089	0.006	0.128	0.576***	0.145
Assignments	0.128	0.100	0.116	0.144	0.207	0.156
Tests	0.192**	0.089	0.252**	0.112	-0.111	0.170
Involve	0.186**	0.092	0.148	0.134	0.297**	0.143
Worthwhile	0.705***	0.118	1.103***	0.220	0.445***	0.151
N	416		228		188	
-2LL	506.197		288.278		191.595	

Note: Three asterisks denote significance at the 1% level, two asterisks at the 5% level, and one asterisk at the 10% level.
A log likelihood ratio test favored separate estimated models for the upper and lower division courses over the pooled (all courses) model.

Table 11. Multinomial Logit Estimates of a Negative Change in Overall Course Score Relative to No Change in Overall Course Score.

Variable	All Courses		Upper Division Courses		Lower Division Courses	
	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
Intercept	-1.361***	0.228	-1.476***	0.296	-1.592***	0.456
Workload	-0.101	0.155	-0.194	0.222	-0.065	0.236
Assignments	-0.009	0.159	0.009	0.222	-0.125	0.234
Tests	-0.371***	0.132	-0.616***	0.202	-0.049	0.212
Involve	0.100	0.163	0.216	0.209	-0.115	0.307
Worthwhile	-0.471***	0.166	-0.473**	0.230	-0.331	0.230
N	416		228		188	
-2LL	506.197		288.278		191.595	

Note: Three asterisks denote significance at the 1% level, two asterisks at the 5% level, and one asterisk at the 10% level.
A log likelihood ratio test favored separate estimated models for the upper and lower division courses over the pooled (all courses) model.

Table 12. Marginal Effects of Multinomial Logit Estimates for Change in Course Appraisal.

Variable	All Courses			Upper Division Courses			Lower Division Courses		
	No Change	Decrease	Increase	No Change	Decrease	Increase	No Change	Decrease	Increase
Workload	-0.046	-0.006	0.052	0.001	-0.005	0.004	-0.099	-0.007	0.106
Assignments	-0.028	-0.002	0.030	-0.028	-0.001	0.029	-0.035	-0.004	0.039
Tests	-0.039	-0.012	0.051	-0.053	-0.018	0.071	0.019	0.001	-0.020
Involve	-0.041	0.000	0.042	-0.037	0.003	0.034	-0.051	-0.005	0.056
Worthwhile	-0.150	-0.022	0.172	-0.255	-0.025	0.281	-0.076	-0.010	0.085