

Good Practice: Suggestions for Development and Analysis of Course Evaluations

by Plamen Iossifov,
Doctoral Candidate, Economics and
Graduate Research Assistant,
Center for Teaching Effectiveness,
University of Delaware

and

Gabriele Bauer, Assistant Director
Center for Teaching Effectiveness,
University of Delaware

March, 2005

Table of Contents

1. Introduction.....3

2. Good Practice: Suggestions for Development of Course Evaluations.....3

 2.1 Course Evaluations Content.....3

 2.2 Questions Content.....4

 2.3 Answer Sets9

3. Practical Suggestions for Analysis of Course Evaluations12

 3.1 Processing Responses to Open-Ended Questions12

 3.2 Processing Responses to Closed-Ended Questions12

References.....14

Appendix 1. Web Shots of UD’s On-line Course Evaluation System.....15

1. INTRODUCTION

Course evaluations serve two main purposes. They provide instructors with students' feedback on certain areas of instruction, such as teaching methods, course organization, and assessment methods. The student feedback can be used for continuous improvement of teaching effectiveness. Course evaluations are also used by administrators in monitoring instructors' performance and making personnel decisions (Arreola & Aleamoni, 1990; TEC, 2005).

Course evaluations, developed in line with good practice, have been shown to be reliable and valid instruments for measuring teaching effectiveness (Arreola, 1995; IDEA, 2005). Research studies have shown that, contrary to the commonly held beliefs among faculty, being popular with students and giving easy grades do not translate in high overall ratings. There is strong evidence that students are capable of discerning effective from ineffective teaching (Arreola, 1995). At the same time, certain student characteristics do influence course evaluations (e.g., students' motivation in taking the class, the level of the course). Therefore, analysis of course evaluations should control for factors that affect student ratings (Arreola, 1995).

Course evaluations constitute only one source of data about teaching. If one wishes to evaluate all elements of university teaching, the course evaluations must be used in combination with multiple data sources, such as products of student work, peer evaluations, and instructor self-evaluation (Cashin, 1988).

2. GOOD PRACTICE: SUGGESTIONS FOR DEVELOPMENT OF COURSE EVALUATIONS

2.1 Course Evaluations Content

Most questions in course evaluations ask students to rate different aspects of their experience with a course. As a result, most answer sets that appear in course evaluations are rating scales. Typically course evaluation include 4-point or 5-point ratings scales; the scales need to be balanced (refer to section 2.3). The main types of rating scales, with suggestions for preferred specification, are:¹

- Frequency scale—Almost always, Frequently, Sometimes, Occasionally, Hardly ever
- Agreement scale—Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree
- Quality scale—Very good, Good, Neither good nor poor, Poor, Very poor

Aspects of students' experience with a course that are typically addressed in end-of-term evaluations and measured with rating scales include (Braskamp & Ory, 1994):

- Course organization and planning.
- Communication skills.

¹ Section 2.3 provides in-depth guidance on how to construct user-defined rating scales.

- Instructor-student interaction, rapport, student involvement.
- Course difficulty, student workload.
- Grading, exams – assessment and feedback.
- Student self-rated learning (e.g., amount learned in this course, effort spent in learning the material).

Course evaluations also include open-ended questions to elicit written student comments about the course and instructor, such as suggestions for improving the course, and response to particular projects.

2.2 Questions Content

Course evaluations help identify faculty strengths in their teaching and areas for enhancement; they also provide input for personnel decisions at the department level. Different types of questions are better suited for each of these tasks. Low-inference questions are more effective in helping faculty identify strengths and areas for change. High-inference questions are more useful for carrying out oversight responsibilities (Arreola & Aleamoni, 1990).

Low-inference questions seek to collect information on objectively measurable aspects of teaching and are typically paired with a frequency scale (Arreola & Aleamoni, 1990). For example, frequency of soliciting student questions, writing important concepts on the board, and giving real-world examples (Theall & Franklin, 1991).

High-inference questions are typically paired with quality or agreement scales and ask for students' subjective judgment on the degree of achievement of course outcomes (TEC, 2005). High inference questions, such as "How would you rate this professor against others in your department/college?" or "Overall, would you recommend this professor to your classmates?" do not provide students with clear criteria for evaluation. It is recommended that such questions be used very carefully and infrequently.

In the remainder of this section, we offer suggestions for wording of course evaluation questions. Examples have been taken from UD's on-line course evaluation question banks²:

- A question's wording should bring the issue to the level of the individual student, instead of asking the student to make a judgment on how the issue was perceived by the entire student population. *Example 1.*
- Questions should address issues within students' competency that are relevant to students (Scriven, 1995). *Example 2.*
- Questions should focus strictly on the course being evaluated and should not ask for comparisons with other courses. At the same time, it is admissible for a question to define a point of comparison, in which case the latter should be made as unambiguous as possible. *Example 3.*

² Problematic questions are identified by a grey-shaded header row.

- Each question should address only one aspect of students' experience with a course.
Example 5. Complex questions can be broken down into a number of simpler questions.
Example 4. Never attempt to simplify a question by moving part of what is being asked in the answer entries. *Example 4.*
- Questions with objective answers, known to faculty, should not be included in course evaluations. For instance, in *Example 4*, question 1 should not be included, and only the relevant subset of questions 2, 3, 4 should be used.

Example 1

Problematic: (1) Question requires student to make a judgment on how the issue was perceived by the entire student population.

The instructor presented information at an appropriate pace.

- (a) Strongly Agree
- (b) Agree
- (c) Neither Agree nor Disagree
- (d) Disagree
- (e) Strongly Disagree

Suggested:

The instructor presented information that I could follow.

- (a) Strongly Agree
- (b) Agree
- (c) Neither Agree nor Disagree
- (d) Disagree
- (e) Strongly Disagree

Source: Theall & Franklin (1991, p. 87)

Example 2

Problematic: (1) Complex question not relevant to students.
(2) Inappropriate rating scale.

To what extent do you believe you have met the following course objectives: Critique theories and models from biological, social, behavioral and epidemiological science that address the bio-psycho-socio-cultural needs of individuals, families and aggregates/target populations in the community setting.

- (a) Always
- (b) Frequently
- (c) Sometimes
- (d) Rarely
- (e) Never

Suggested:

Clarity of course objectives.

- (a) Very Good
- (b) Good
- (c) Neither Good nor Poor
- (d) Poor
- (e) Very Poor

Considering all aspects of the course, I rate the course.

- (a) Very Good
- (b) Good
- (c) Neither Good nor Poor
- (d) Poor
- (e) Very Poor

Example 3

Problematic: (1) Ambiguous reference group for comparison.
(2) Not relevant from student's point of view.

I have learned as much in this course as in other courses at U.D.

- (a) Strongly Agree
- (b) Agree
- (c) Neither
- (d) Disagree
- (e) Strongly Disagree

Suggested:

The instructor encouraged questions and class participation sufficiently for a class this size.

- (a) Strongly Agree
- (b) Agree
- (c) Neither Agree nor Disagree
- (d) Disagree
- (e) Strongly Disagree

Given the class size, did the instructor return written work/exams in a reasonable time?

- (a) Always
- (b) Frequently
- (c) Sometimes
- (d) Rarely
- (e) Never

Example 4

Problematic: (1) Subject matter of question shifted to the answer set.
(2) *Non-response answers* included alongside *regular answers*.³

Did you find that meeting the prerequisites for this class helped you understand the material?

- (a) Yes
- (b) No
- (c) No prerequisites stated
- (d) None stated, but there should be
- (e) I did not meet the prerequisites

Suggested:

1. Were there prerequisites for taking this class?

- (a) Yes
- (b) No

2. If you've answered "No" to question 1, do you think that there should be prerequisites for taking this class?

- (a) Yes
- (b) No

3. If you've answered "Yes" to question 1, have you met the prerequisites for this class?

- (a) Yes
- (b) No

4. If you've answered "Yes" to question 3, did you find that meeting the prerequisites for this class helped you understand the material?

- (a) Yes
- (b) No

³ The answer set to each question contains two types of answers: (1) *regular answers* which upon selection provide part or all of the information sought from the respondent by that particular question; and (2) *non-response answers* which upon selection indicate the reason why the respondent did not answer the question (e.g. Not Applicable, Refused to Answer).

Example 5

Problematic: (1) Complex question that forces students to average their ratings of two unrelated features of course materials.
(2) Unbalanced rating scale.

The accuracy and timeliness of the materials received were:

- (a) Excellent
- (b) Very Good
- (c) Good
- (d) Satisfactory
- (e) Unsatisfactory

Suggested:

1. The accuracy of the materials received was.
 - (a) Very Good
 - (b) Good
 - (c) Neither Good nor Poor
 - (d) Poor
 - (e) Very Poor

2. The timeliness of the materials received was.
 - (a) Very Good
 - (b) Good
 - (c) Neither Good nor Poor
 - (d) Poor
 - (e) Very Poor

2.3 Answer Sets

Most questions in course evaluations ask students to rate different aspects of their experience with a course. As a result, most answer sets that appear in course evaluations are rating scales. UD's On-line Course Evaluation System encourages users to choose from a set of approved rating scales (Appendix 1, Figure 2). Users also have the option of creating their custom rating scales, using the "Multiple Choice" option of the form "Create a question" (Appendix 1, Figure 1). Custom answer sets should conform with the following good practice:

- Rating scales should be balanced (Arreola & Aleamoni, 2000):
 - The same number of answer set entries should be provided for expressing positive and negative perceptions. *Example 7.*
 - In case a neutral category is included, care should be taken to ensure lack of bias (positive or negative). *Example 7.*

- The difference in intensity of positive or negative attitudes conveyed by all pairs of neighboring rating categories must be comparable. *Example 7.*
- Answer set entries should not include non-response answers. *Example 6.* If the relevance of the question needs to be determined, use a chain of master and dependent questions. *Example 4.*⁴
- Different entries in the answer set should be mutually exclusive. *Example 8.*
- The answer set to a question should make sense in light of what is being asked. *Example 8.*

Example 6

Problematic: (1) Answer set entries include non-response answer.

In this course, I improved the quality of my writing.

- (a) Strongly agree
- (b) Agree
- (c) Disagree
- (d) Strongly disagree
- (e) No opinion/does not apply

Suggested:

In this course, I improved the quality of my writing.

- (a) Strongly agree
- (b) Agree
- (c) Neither Agree nor Disagree
- (d) Disagree
- (e) Strongly disagree

⁴ *Independent question* is a question, for which the decision on the part of the respondent of whether to give a valid response or a non-response, does not depend on her response to a prior question. The choice between a valid response or a non-response to a *dependent question*, on the other hand, is fully determined by the response given to a prior question, called a *master question*. The entry in the pre-defined answer set of the master question, which upon selection triggers a non-response to the dependent question is called a *trigger answer*.

Example 7

Problematic:(1) Unbalanced rating scale:

- Only one negative category.
- Neutral category has positive bias.
- Intensity of attitude conveyed by the neighboring categories “Excellent and Very Good” and “Satisfactory and Unsatisfactory” is arguably not of comparable magnitude.

Rate the effectiveness of the required readings in achieving the course objectives.

- (a) Excellent
- (b) Very Good
- (c) Good
- (d) Satisfactory
- (e) Unsatisfactory

Suggested:

Rate the effectiveness of the required readings in achieving the course objectives.

- (a) Very Good
- (b) Good
- (c) Neither Good nor Poor
- (d) Poor
- (e) Very Poor

Example 8

Problematic: (1) The answer set is irrelevant to what is being asked in the body of the question.

Indicate the level of learning achieved in communicating effectively.

- (a) Synthesis (can design, formulate)
- (b) Analysis (can analyze, explain why)
- (c) Application (can recognize, apply)
- (d) Comprehension (can describe, explain)
- (e) Knowledge (can recall, repeat)

Problematic:(1) Different entries in the answer set are not mutually exclusive

Indicate your anticipated grade in this course.

- (a) A or B
- (b) B or C
- (c) C or D
- (d) D or F

3. PRACTICAL SUGGESTIONS FOR ANALYSIS OF COURSE EVALUATIONS

Developers of course evaluations should keep in mind that different types of questions require different analytical techniques to process collected responses. Responses to *closed-ended questions* can be analyzed using a variety of statistical techniques, whereas *open-ended questions* do not readily lend themselves to statistical analysis. The use of *mixed-type questions* should be avoided, in favor of using a set of *closed and open-ended questions*.⁵

3.1 Processing Responses to Open-Ended Questions

The processing of open-ended questions starts with the generation of a query containing all responses given to each open-ended question. The query can be used by the course instructor and department chair to glean insights on different aspects of students' experience. Analysis of qualitative data is central. For example, content analysis of student responses allows for identification of main themes. The main themes can be presented in a table alongside frequencies of appearance in student responses.

Qualitative data reflect students' perceptions and need to be balanced against other data, such as instructor perceptions, course goals, and course curriculum. Department chairs and instructors should be cautious of overinterpreting students' written comments (Theall and Franklin, 1991). Typically, only a small percentage of students answer open-ended questions and these tend to be the students, who are most satisfied and most dissatisfied with the instructor. Therefore, in contrast with answers to closed-ended questions, students' written comments tend to be not representative of class attitudes. A general rule of thumb is that answers to open-ended questions should mainly be used by faculty for course enhancement purposes. When a large percentage of the class has opted to respond, the information may be used as one input in taking personnel actions.

3.2 Processing Responses to Closed-Ended Questions

The responses to a closed-ended question can be mapped into values of a *categorical, ordinal or interval variable*, depending on the characteristics of the answer set entries (UCLA, 2005a). This is achieved by assigning a numeric score to each answer set entry and translating respondents' selections from checkmarks to numeric values.⁶ *In UD's On-line Course Evaluation System,*

⁵ *Closed-ended questions* have pre-defined set of answers, from which respondents select the ones applicable to them. *Open-ended questions* do not have a pre-defined set of answers. Instead, they allow respondents to provide responses in free-text form. *Mixed-type questions* allow respondents to provide responses in free-text form in addition to selecting from a pre-defined set of answers.

⁶ In the case of *categorical variables*, the ordering of different answer set entries from highest to lowest, implied by their numeric scores, does not carry over to the choices they represent. For example, if the question seeks to establish student's race, and the answers from which the student chooses are converted into integers, the resulting ordering of different races is nonsensical. In the case of *ordinal variables*, both the ordering and the spacing of numeric scores carries over to the choices they represent. For example, if the answer set to a question represents a rating scale (e.g., Excellent, Very Good, Good, Satisfactory, and Unsatisfactory) with respective numeric scores (e.g., 5, 4, 3, 2, and 1), responses to that question can be mapped to values of an ordinal variable, if one is willing to assume that the difference in intensity of positive or negative attitudes conveyed by all pairs of neighboring rating categories is of comparable magnitude. In the case of *interval variables*, the ordering but not the spacing of numeric

(continued)

rating categories of approved rating scales (Appendix 1, Figure 2) are assigned numeric scores from 1 to 5, in such a way that the most positive category receives the highest score, and the most negative—the lowest. When designing custom rating scales, using the “Multiple Choice” option of the form “Create a question” in UD’s On-line Course Evaluation System (Appendix 1, Figure 1), users should always place the most positive category in the text-box labeled “A”, and the most negative one—in the text-box labeled “E”.

Whereas categorical variables are straightforward to identify, the choice between ordinal and interval variables is often subjective. *As a general rule of thumb, responses to questions with answer sets representing balanced rating scales are best represented by ordinal variables, whereas responses to questions with answer sets representing unbalanced rating scales are best represented by interval variables.* Statistical techniques appropriate for categorical variables are also appropriate for both ordinal and interval variables, whereas the opposite is not true.

As noted in the preceding chapter, the use of unbalanced rating scales in course evaluations should be avoided. Thus, course evaluation questions should mainly use categorical and ordinal variables. Statistical techniques commonly used in the analysis of categorical and ordinal variables are reviewed below. Statistical tests appropriate for interval variables are discussed in UCLA (2005b).

Tabulation of responses

Appropriate for analysis of both categorical and ordinal variables. For each entry of the answer set associated with a given question, the output from the procedure shows the percentage of students, who have judged the question to be applicable and have answered it, who have chosen that particular answer set entry. In the case of ordinal variables, an easy way to further summarize students’ responses is to calculate for each question the difference between the percentage of students, who have selected the two positive rating categories, and the percentage of students, who have selected the two negative rating categories.

Averaging of rating scores

Appropriate for analysis of ordinal variables only. For each question, the output from the procedure shows the average of the numeric scores corresponding to students’ selections from the answer set associated with the question. In the UD’s On-line Course Evaluation System, rating categories are assigned numeric scores in such a way that the most positive category receives the highest score, and the most negative—the lowest. Therefore, higher average of numeric scores indicates more positive student perceptions, on average.

scores carries over to the choices they represent (UCLA, 2005a). In the above example, the difference in intensity of attitude conveyed by the neighboring categories “Excellent and Very Good” and “Satisfactory and Unsatisfactory” is arguably not of comparable magnitude, so responses to that question form the values of an interval variable.

REFERENCES⁷

- Arreola, R. & Aleamoni, L., 2000, *Assessing Student Learning Outcomes: A CEDA Workshop Resource Document*, (Memphis: Center for Educational Development and Assessment).
- , 1990, “Practical Decisions in Developing and Operating a Faculty Evaluation System,” *New Directions for Teaching and Learning*, 43.
- Arreola, R., 1995, “Student Rating Form Selection and Development Kit Part I: Some Common Misconceptions and Beliefs,” in *Developing a Comprehensive Faculty Evaluation System*, (Bolton, MA: Anker Publishing Company).
- Boston, C., 2002, “The Concept of Formative Assessment,” *Practical Assessment, Research & Evaluation*, 8(9), <http://pareonline.net/getvn.asp?v=8&n=9>.
- Braskamp, F. & Ory, F., 1994, *Assessing Faculty Work*, (San Francisco: Jossey-Bass).
- Cashin W., 1988, *Student Ratings of Teaching: A Summary of the Research*, IDEA Paper No. 20, Manhattan, KS: The IDEA Center at Kansas State University.
- Scriven, M., 1995, “Student Ratings Offer Useful Input to Teacher Evaluations,” *Practical Assessment, Research and Evaluation*, 4(7), <http://PAREonline.net/getvn.asp?v=4&n=7>.
- TEC, 2005, List of Evaluation Glossary Terms, <http://ec.wmich.edu/Glossary/glossaryList.htm>, (Western Michigan University: The Evaluation Center).
- Theall, M. & Franklin, J., 1991, “Using Student Ratings for Teaching Improvement,” *New Directions for Teaching and Learning*, 48.
- UCLA Academic Technology Services, 2005a, “What is the difference between categorical, ordinal and interval variables?,” www.ats.ucla.edu/stat/mult_pkg/whatstat/nominal_ordinal_interval.htm.
- UCLA Academic Technology Services, 2005b, “Choosing the Correct Statistical Test,” www.ats.ucla.edu/stat/mult_pkg/whatstat/choosestat.html.

⁷ Referenced materials that are not available online can be accessed at UD’s Center for Teaching Effectiveness Library, 212 Gore Hall.

APPENDIX 1. WEB SHOTS OF UD'S ON-LINE COURSE EVALUATION SYSTEM

Figure 1. Create a Question Form

The screenshot shows a web browser window titled "Course Evaluations - Create a Question - Microsoft Internet Explorer provided by UGS". The address bar contains the URL: <https://chico.nss.udel.edu/CourseEvaluations/admin/question.jsp?searchParam=instr&instr=instr>. The page header includes the University of Delaware logo and the text "Course Evaluations". The main content area is titled "Create a Question - Instructor" and contains the following sections:

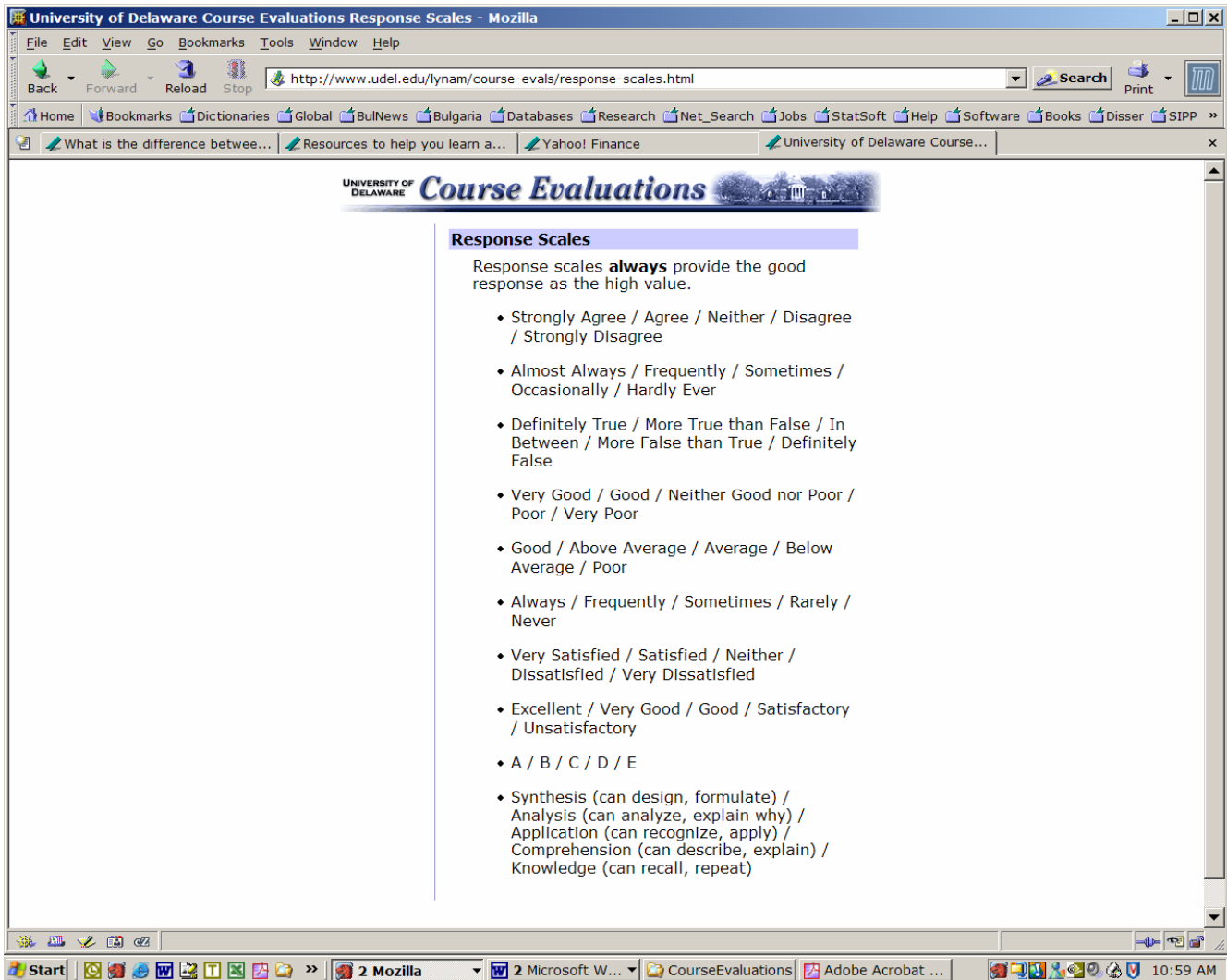
- Question**: A large text input field for entering the question text.
- Type question**: Radio buttons for "Course" and "Instructor".
- Question Status**: Radio buttons for "Display" (selected) and "Hide from view".
- Specify answer format. Choose one.**: A dropdown menu currently set to "Choose One". Below it are radio buttons for "Multiple Choice", "Essay (4000 character limit)", "Short Answer", "True / False", and "Yes / No". The "Multiple Choice" section includes five labeled input fields (A, B, C, D, E).

At the bottom of the form are two buttons: "Submit Form" and "Reset Form". The browser's taskbar at the bottom shows the Start button, several application icons, and the system clock displaying "10:46 AM".

Source:

<https://chico.nss.udel.edu/CourseEvaluations/admin/question.jsp?searchParam=instr&instr=instr>

Figure 2. Approved Rating Scales



Source: <https://chico.nss.udel.edu/CourseEvaluations/admin/instr.jsp>.