Report of the Teaching Evaluation Committee

The charge of this committee was to study possible improvements in the way that student evaluations of teaching are solicited, presented, and used at UCA. The committee would like to emphasize that standardized questionnaires are but one source of information about teaching; others are peer review, chair review, and alumni review. This committee strongly recommends that the process of evaluating teaching be as broad-based as possible with input beyond just student evaluations. As described in *Suggestions for Meeting Faculty Handbook Standards for Tenure and Promotion,* UCA (1996, p. 4):

> Additional documentation of contributions to teaching might include (but are not limited to): results of peer evaluations; portfolio evaluations; syllabi and course materials; outside recognition of the teacher and/or the teacher's students; instructional manuals or other pedagogical materials (including textbooks) the instructor has developed; theses and dissertations directed; and evidence of innovative and creative approaches to teaching improvement.

That said, we concur that regular student evaluations of teaching serve several important purposes. One is to improve student learning by improving faculty teaching. Another is to evaluate faculty for promotion, tenure, mid-tenure, merit pay and the like. A third is to allow faculty to reflect on their teaching. Our review of the literature suggests that some types of questions are best used for evaluative purposes, and others may be more appropriate for improvement purposes. Some information can be gathered from multiple choice questions; other information is best gathered from written questions.

FORMATIVE VERSUS SUMMATIVE

The literature distinguishes between two types of questions used for student evaluation measures: formative and summative. Formative questions are designed to inform the instructor about improvements that might be made. Summative questions are used for evaluative (personnel) purposes. There is substantial literature supporting the use of a small number of global questions for evaluative purposes and a larger selection of specific diagnostic items for improvement (Braskamp & Ory, 1994, Braskamp, Brandenburg & Ory, 1984, Cashin, 1999, Doyle, 1975, Feldman, 1989). There is some agreement in the literature that a small number of global questions are more indicative of teaching quality than many diagnostic questions.

The rationale for the use of global questions for summative evaluation is that they correlate more strongly with student learning, and student learning is the purpose of effective teaching. Braskamp, et al. (1984, p. 46) say "global ratings by students correlate more highly with student learning than do diagnostic ratings, therefore global ratings are recommended for personnel decision making." The underlying literature to support this claim is reviewed by Cohen (1981, p. 281-309) in "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multi-section Validity Studies." He studied results of 41 independent studies reporting on 68 multi-section courses. The data analyzed included responses to various teaching evaluation questions and student scores on independent learning evaluation instruments. The data included a wide range of disciplines. He finds that "The average correlation between an overall instructor rating and student achievement was .43; the average correlation between an overall course rating and student achievement was .47. . . other evaluative questions showed more modest relationships with student achievement. In 59 of 68 courses, overall instructor rating correlated positively with student achievement" (Cohen, 1981, p. 281).

Other studies that show a strong relationship between ratings and student learning are Centra (1977), Costin (1978), and Frey (1973). As Cohen says (1981, p. 283), "Even though there is a lack of unanimity on a definition of good teaching, most researchers in this area agree that student learning is the most important criterion of teaching effectiveness." According to Doyle (1975, p. 82), "probably the most

appropriate question to ask students for personnel purposes is 'How would you rate the general teaching ability of this instructor?' It is valid with respect to its reliability., and it also tends to infer qualities and outcomes that many would consider germane."

The committee suggests that the three questions below be used as global questions for summative purposes, and that the scores on these questions be the primary source of *student evaluation input* for making summative decisions about faculty teaching. Again we emphasize that student evaluation input is but one source of information about the quality of teaching. These three questions will replace the overall score that is calculated in the current document as an equally weighted average of responses to the other questions. **The committee could find no support in the literature for an equally weighted average of any set of typical student evaluation questions.** Theall is particularly critical of the calculated overall question (UCA seminar, 1999). Other arguments for the use of global questions are that the global questions are holistic and somewhat similar to a final grade in a course and they offer some degree of comparability across disciplines without specifying the components of effective teaching across disciplines.

The global questions we propose are shown below. We suggest that these questions be the initial questions asked on the student evaluation form, in the case that the length of the form we are proposing would cause some students not to complete what we consider to be the most important questions. Braskamp and Ory (1994, p. 178) find that the placement of the global questions on the form is not crucial to their validity.

1. How much do you feel you have learned in this class?

   Key: 5 = a great deal
   4 = more than usual
   3 = about the usual amount
   2 = less than usual
   1 = very little

2. How would you rate the instructor's teaching ability?

   Key: 5 = exceptional
   4 = very good
   3 = good
   2 = not very good
   1 = poor

3. How would you rate the course in general?

   Key 5 = exceptional
   4 = very good
   3 = good
   2 = not very good
   1 = poor

FORMATIVE QUESTIONS

Formative or diagnostic questions are designed to provide information about specific behaviors that occur in the classroom, and the reaction of the students to those behaviors. In our form, students are not asked to judge the appropriateness of the methods, but only how they feel about them. **As stated earlier, the literature does not verify that formative questions correlate with student learning.**

The literature distinguishes three types of formative questions: descriptive, general concept, and norms. The descriptive questions seek to identify what happens in the classroom. The general concept questions ask the student to make some inferences about course management, student outcomes of instruction,

student preferences for instructor's learning style and specific instructional setting (Braskamp et al., 1984, p. 43). The norms are used to distinguish differences in classroom experiences over which the instructor has no control, for example, "Is the course required? Is it a course in the major? Are the students freshmen or seniors?" The norms can be used to group or qualify the responses to the other questions.

The following set of formative questions was modified from questions from Theall. They were posted on the UCA web, and mostly favorable responses were received from faculty and administrators. A brown bag luncheon to discuss these questions was well attended by faculty, and additional changes were made in response to faculty suggestions. Those questions appear below, and are coded **D** for descriptive, **GC** for general concept, and **N** for norm.

For the following statements about the instructor, use this key:

5 = almost always
4 **=** more than half of the time
3 = about half of the time
2 **=** less than half of the time
1 **=** almost never

The Instructor

4. (D) communicates the purposes of class sessions and instructional activities
5. (D) speaks clearly and audibly when presenting information
6. (GC) uses examples and illustrations which help clarify the topic being discussed
7. (GC) shows meaningful relationships among the topics in this course
8. (D) inspires interest in the subject matter of this course
9. (D) relates course material to life situations when possible
10. (D) asks questions that challenge me to think
11. (D) provides opportunities for me to suggest or discuss issues related to the course
12. (D) develops an atmosphere of respect and trust in the classroom
13. (GC) manages classroom discussions so that they are useful
14. (GC) clears up points of confusion
15. (D) provides the opportunity for assistance on an individual basis outside of class
16. (D) gives me regular feedback about how well I am doing in the course
17. (GC) gives tests and assignments quickly enough to benefit me
18. (D) returns exams and assignments quickly enough to benefit me
19. (D) when necessary, suggests specific ways I can improve my performance in this course
20. (GC) makes effective use of class time
21. (D) is punctual in meeting class and office hour responsibilities
22. (GC) Rate how well the syllabus, course outline, or other overviews provided by the instructor helped you to understand the goals and requirements of this course.
23. (GC) Rate the usefulness of the assignments in helping you to learn.

The Course

(Each question has a key appropriate to it)

24. (N) the workload for this course is
25. (N) the difficulty level of the course activities and materials is
26. (N) of the following, which best describes this course for you
27. (N) your classification is

Written Comments

Scholarship in the field indicates that open-ended written comments should not be used for teacher evaluation concerning matters of promotion, tenure and merit pay. The quote below is representative of the reasons for not using written comments for evaluation:

> There can be hundreds or even thousands of comments. To assess them accurately we must do a content analysis, classifying every response as to content and also making a judgment about how positive or negative the comment is. This is extremely time consuming. My belief is that usually only the individual instructor has the motivation to do this and so the comments should only be used by the instructor for improvement. To have evaluators simply scan the comments to gain a general impression often leads to their only remembering the most sensational comments, not the more representative ones (Cashin, 1999, p. 38).

Written comments can, however, be very useful in helping teachers improve their teaching (Arreola, 1995, Braskamp & Ory, 1994, Cashin, 1999, and Centra, 1993) Below are questions this committee recommends (these comments to be seen only by the teacher, though they would be administered by a neutral party and kept closed until after grades have been turned in). The top of the questionnaire should make clear to the student that the comments are anonymous, confidential, and for the teacher's use only. It is recommended that the teacher can most benefit from the comments by reading question by question instead of form by form (a procedure that will make patterns clearer).

> Class, Section, Instructor's Name, Date
>
> Please write legibly in answering the following questions. Your thoughtful and responsible comments will be of great help to your professor. (These comments will be kept sealed until after the end of the term when grades have been assigned.)

1. Describe one or more things about the course that you found helpful. Please give examples and be specific.
2. Was there anything else the instructor could do, given time constraints, that you would find helpful?
3. How would you assess your own performance in this course?
4. Comment on how assigned work contributed to the learning process. What books, labs, drills, etc. were most useful to you? Which were least useful and why?

COMMENTS ON THE LENGTH OF THE FORM AND ALTERNATIVES

Some have criticized lengthy forms and have used factor analysis in order to reduce the number of redundant questions. The current UCA form is a variety of what Braskamp calls the "omnibus form" which asks questions in categories that have previously been defined by factor analysis. The factors that appear most often include communication skill, rapport with students, course organization, student rated accomplishments, course difficulty, grading and examinations. This format is thought to be less useful for summative purposes because of the possibility of disagreement about what constitutes effective teaching. Some have questioned the usefulness of lengthy forms claiming that earlier questions bias later questions. Doyle (1975, p. 41) reviewed early halo studies and concluded that "the halo effect is not so powerful that stable factor structures cannot emerge." In a lengthy questionnaire, with questions all designed to identify particular strengths and weaknesses in teaching, we would expect there to be strong correlation among a student's answers to questions. Earlier questions may influence later questions, although the literature on this point is not definitive. This is another reason to use only the first questions for evaluation.

For purposes of improvement, a recent view is that more information is better. An interesting variant of the formative form is the cafeteria program, which allows the instructor to choose some or all of the

questions that will be asked. An example of this is the University of Illinois Urbana Champaign model, which has 400 items. A drawback of this alternative is, as Braskamp and Ory (1994, p. 53) say, that "most teachers select items that assess their course management rather than their teaching skills or personal characteristics." Another drawback would be lack of comparability across campus.

An approach that focuses on student learning rather than instruction is termed the goal-based form which asks students to rate their performance on a number of stated course goals and objectives, such as gaining factual knowledge, developing special skills and competencies, developing appreciation for the subject matter, etc. An example of this is the IDEA system at Kansas State. An example of such a question might be, "How much would you say you've learned in this course?" Our questionnaire asks some questions along these lines. As outcomes assessment is further developed at UCA, it may provide better methods of evaluating faculty members, but at present, the students are one of our better sources of information about student learning.

PRESENTATION OF RESULTS OF STUDENT EVALUATIONS

- The committee recommends that percentile rankings *not* be presented for teaching evaluation purposes.
- It is also recommended that departmental and college means comparisons be calculated by lower- and upper-division groupings, depending on the division status of the course being evaluated.

The rationale for the first recommendation is that percentiles can be very sensitive to small changes in the instructor mean. A change of a few hundredths in the instructor mean can change a percentile ranking by tens of points. In a department or college where most responses to a question are excellent, a very good response may receive an undeserving low percentile rank. When comparing instructors' rankings across semesters, it is not difficult to find examples that illustrate the instability of percentile ranks. These ranks may change drastically even when raw score means are identical, and worse, move in a direction opposite to the means.

The rationale for the lower- and upper-division distinction in data presentation is that student motivation tends to show higher correlations with other student rating items than any other variable. Instructors are more likely to obtain higher ratings in classes where students had a prior interest in the subject matter or were taking the course as an elective (Cashin, 1999, p. 33). Upper-division and lower-division is a proxy for more detailed information along these lines.

RECOMMENDATION FOR GRAPHICAL DISPLAY OF FACULTY EVALUATION DATA

The committee recommends that results be displayed in the following graphical format (Figure 1):

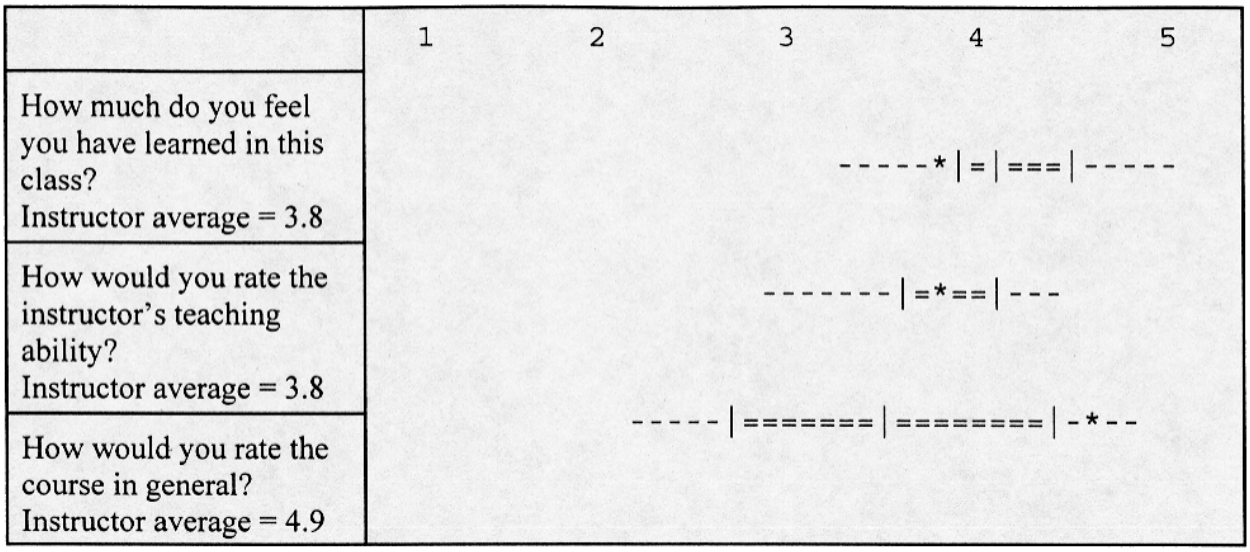

Figure 1: Graphical Display of Faculty Evaluation Data

Each graph is a box and whiskers plot for the group (e.g., upper or lower division, department/college) to which the individual faculty member is being compared. From left to right, the vertical lines represent the 25th, 50th, and 75th percentiles (quartiles), respectively. Dashes represent the upper and lower 25% of scores, and the equal signs represent the middle 50% of the scores. The asterisk represents the faculty member's score. If the faculty member's score falls at one of the quartiles, the asterisk will overwrite the vertical line (see second question). The resolution of the line printer (one character = 0.1 scale units) prevents overemphasis of extremely small differences in original scale values. Cashin (1999, p. 36) recommends reporting average scale values only to the nearest 10th of a point, exactly what is possible in the above example.

By providing a graphical representation of percentile ranks, this format provides comparative information, which is considered valuable for interpreting student ratings of faculty performance (Cashin, 1999, p. 32). At the same time, the box-and-whiskers plot helps avoid misinterpretation of percentile ranks by providing a clear indication of the instructor's performance on the original scale, as well as providing information on the comparison group. This is especially important when the data are significantly skewed (Cohen, Swerdlik, & Phillips, 1999, p. 120), as use of percentile ranks tends to overemphasize small differences between values around which the majority of scores are clustered. Extreme negative skew is typical of responses to faculty evaluation instruments (Tagomori & Bishop, 1995). The box-and-whiskers plot allows its influence on comparative measures to be seen with greater simplicity than other approaches that have been proposed (cf., Chin, Haughton, & Aczel, 1995).

The three examples here were chosen to illustrate some of the difficulties in using percentile ranks and the value of considering the original meanings of the ratings. For example, the hypothetical instructor here did approximately the same on the original scale for the first and second question. However, the percentile rankings would be substantially different. On the first question the instructor is below the 25th percentile, but on the second the instructor is right at the 50th percentile. On the third question, the instructor looks good by either measure. There is much more variability in the comparison group, yet despite this, the instructor received very high ratings and is in the upper 25% of the comparison group.

Incidently, the information on the comparison group may be used by departments and colleges to identify their strengths and weaknesses in teaching. In the above example, it is clear that responses are relatively consistent regarding instructor performance and amount learned in classes, but there is quite a bit of variation in responses to the overall value of courses taught by the comparison group. This might indicate a need for reconsideration of course offerings, or a clearer communication to students of the value and purposes of existing courses.

Although interpretation of this type of presentation becomes fairly easy after some experience, it is expected that some initial training of faculty and administrators will be necessary. Of course, this is true with any rating system, regardless of apparent simplicity of comparative ratings (Cashin, 1999, p.40)

USE OF THE RESULTS OF THE STUDENT EVALUATIONS

For reasons stated above, the committee recommends that the responses of questions designed to inform the faculty member (questions # 4-23 and the written comments) be sent exclusively to the faculty member, and that the responses of questions designed to be used to evaluate the faculty member be sent to the faculty member, chair, and dean (questions #1-3). Everyone would receive the matching norms to describe the characteristics of the class that the faculty member cannot control (questions #24-27).

To the question of how the chair could help the faculty member improve if he/she did not receive the formative responses, we suggest that processes are already in place to provide a conversation about the specifics of a faculty member's teaching performance. Those processes are as follows. During the annual performance review conference, the chair has an opportunity to discuss the teaching evaluation global scores with the faculty member, and it is anticipated that the faculty member will share the results of the formative questions with the chair if help in understanding what remedies would be appropriate is needed. Another opportunity occurs when the faculty member presents a faculty plan each year that lists what activities will be undertaken to try to improve in teaching, research and service to the university. In cases of merit pay, the use of the global ratings is the appropriate choice for ranking faculty members in a department, as that occasion is designed to reward performance, rather than promise. When faculty members prepare their mid-tenure review documentation and their applications for promotion and tenure, they may include copies of their formative ratings if they are good. If they are bad, but have improved, the faculty member will also have occasion to bring them to the table. These processes, rather than merely reviewing the individual formative scores, are more likely to provide improvement because they involve a conversation between the faculty member and the evaluator, in which improvement will be noticed and rewarded.

The committee recommends that the written comments be given to the faculty member only, with no expectation that they be sent forward. The rationale for the recommendation is that a written comment requires "extra effort" on the part of the student. Therefore these comments are most likely to come from students with extreme views. They are not representative of the majority of the class. In effect the use of written comments gives this unrepresentative group of students two opportunities to express their opinions, one on the numeric ratings and one on the written comments. Written comments can, and occasionally do, contain slanderous remarks that can have devastating consequences when peer committees participate in personnel matters. The anonymous nature of the written comments makes it difficult for the instructor to answer such charges. Also, since evaluators would have great numbers of such elements to scan, they are, as Cashin says, likely to remember the more sensational comments rather than those that are more representative (1999, p. 38).

CONCLUSIONS

The committee recommends that the administration consider implementation of the proposed teaching evaluation instrument. This version of the instrument was developed after discussing the input we received from the faculty relating to the initial draft proposal. From the beginning of our discussions we

were influenced by Theall's recommendations and examples, and many of the formative questions that appear are mere revisions of questions he has published. If some form of this instrument is to be implemented, Dr. Theall would need to be contacted concerning copyright and other logistical matters. The committee also believes that any new evaluation instrument should be phased in, possibly using both evaluations for a period of time.

Sources Cited

Areolla, Raoul A. (1995). *Developing a Comprehensive Faculty Evaluation System: A Handbook for College Faculty and Administrators on Designing and Operating a Comprehensive Faculty Evaluation System.* Bolton, MA: Anker.

Baez, Benjamin, and John A. Centra. (1995). *Legal and Tenure, Promotion, and Reappointment Administrative Implications.* Washington, DC: The George Washington University.

Braskamp, Larry A., and John C. Ory. (1994). *Assessing Faculty Work: Enhancing Individual and Institutional Performance.* San Francisco: Jossey-Bass.

Braskamp, Larry A., Dale C. Brandenburg, and John C. Ory. (1984). *Evaluating Teaching Effectiveness.* Beverly Hills: Sage.

Cashin, William. (1999). Student Ratings of Teaching: Uses and Misuses. *Changing Practices in Evaluating Teaching.* Ed. Peter Seldin. NY: Anchor.

Centra, John A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness.* San Francisco: Jossey-Bass.

Centra, J. A. (1977). Student Ratings of Instruction and their Relationship to Student Learning. *American Educational Research Journal,* 17-24.

Chin, L., D. Haughton, and A. Aczel. (1996). Analysis of Student Evaluation of Teaching Scores using Bootstrap and Permutation Methods. *Journal of Computing in Higher Education* 8, 69-84.

Cohen, Peter A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multi-section Validity Studies. *Review of Educational Research* 51, 281-309.

Costin, F. (1978). Do Student Ratings of College Teachers Predict Student Achievement? *Teaching of Psychology* 5, 86-88.

Doyle, Kenneth O. (1975). *Student Evaluation of Instruction*. Lexington, MA: D. C. Heath.

Frey, P. W. (1973). Student Ratings of Teaching: The Validity of Several Rating Factors. *Science* 182, 83-85.

Goodwin, Harold I., and Edwin R. Smith. (1983). *Faculty and Administrator Evaluation: Constructing the Instruments.* West Virginia University Department of Education Administration.

Seldin, P. (1999). *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions.* Bolton, MA: Anker.

Tagomori, H.T., and L. A. Bishop. (1995). Student Evaluation of Teaching: Flaws in the Instruments. *Thought & Action* 11, 63-78.

Theall, Michael. (1999). "Perspectives on Faculty Evaluation" presented in seminars to faculty and administrators at UCA, March 11, 1999.

Weimer, Maryellen, Joan L. Parrett, and Mary-Margaret Kerns. (1992). *How Am I Teaching? Forms and Activities for Acquiring Instructional Input.* Madison, WI: Magna.