**Attachment #1 to Resolution #13-05/06-SA**

# Report of the

# Course Evaluation Subcommittee

# Submitted to the Academic Senate
# October 24, 2005

**Summary of the Report of the Course Evaluation Subcommittee**

1 Purpose:
   a) To connect student background variables to course responses.
   b) To identify dimensions of teaching and learning that are valued within our University community and the literature on teaching in higher education.
   c) To provide relevant information to candidates and reviewers in the Retention, Tenure and promotion process.
   d) To characterize student experience in a course and provide instructors with information helpful in refining course characteristics and instruction.

2 Recommendations are based on
   a) review of relevant literature, multiple assessment center websites, and forms currently in use on campus,
   b) input from individual faculty members and Senate members (Spring 2005),
   c) feedback from nationally recognized workshop.

3 Two types of assessment forms developed:
   a) Behavioral Profile – not a rating form; produces a profile of teaching practices; included in response to feedback received in Senate that some faculty wanted descriptions of teaching, not numeric codes.
   b) Rating – model most commonly used since the 1980's so there is more research available related to the use of this approach in higher education than is available on the behavioral profile approach.

4 Both forms have three part structure: background, core items and supplemental items.

5 Development followed recommended steps; items based on analysis of characteristics of good teaching.

6 Domains are consistent with literature.

7 Literature supports local piloting of instrument in all cases, even if purchasing nationally recognized tools.

8 Literature recommends that the pilot study should include more items than will ultimately be used. This is especially appropriate for piloting of the ranking approach. Final form recommendation is 8 items per domain.

9 Pilot report with recommendations will be provided to Senate by March. Recommendations will include recommendations for implementation and contextualization within an overall program of evaluation. Implementation planned for Spring 2006.

**Report of the Course Evaluation Subcommittee**
**October 25, 2005**

## Introduction

In this report, the work of the Academic Senate Subcommittee on Course Evaluations is summarized and a recommendation is forwarded to the Senate in support of piloting two newly designed course evaluation forms at the end of Fall semester 2005. This pilot use of the forms would involve a small number of classes taught by tenured professors representing all three colleges. The committee would request faculty volunteers as participants. The Committee further recommends that the pilot include short surveys of the faculty, staff and student participants' perceptions of the course evaluation instrument and its administration and processing. Once this pilot study is completed, the subcommittee will submit a revised course evaluation form along with its recommendations for implementation of standardized use of the form across the campus. The course evaluation form developed for use in this pilot is composed of three sections: background information, common core questions and supplemental questions selected by the department or instructor as described below. A rationale for the structure and content of the evaluation form is provided as part of this report.

## Background to the proposed University-Wide Course Evaluation

Starting late in the fall of 2004, a subcommittee of the Senate Student Affairs Committee accepted the charge of developing a proposal for University-wide course evaluations. The committee began by collecting and examining all of the course evaluations currently in use on campus and collecting and examining course evaluations from across the CSU and from data bases in campus assessment centers around the country. We also read reports on course evaluation research and solicited input from colleagues on campus who have been involved in course evaluation and survey research. Simultaneously we were reading about and discussing characteristics of good teaching. By the middle of the spring semester 2005, the committee had settled on a basic structure (background, core and supplemental questions), identified domains to be assessed and begun examining the pros and cons of various question structures. This approach to developing an evaluation tool follows the guidelines provided in the literature, particularly Arreola, 2000. In late spring, the committee brought a document to the Senate describing the proposed structure and identifying some background questions and some examples of core question formats. Input received from Senators was incorporated into the documents, and the committee began discussion of the possible development of two different models for consideration. In a meeting in August, the President and Provost requested that the committee finish its work to bring a completed proposal to the Senate by mid-October.

## Proposal

The committee has developed the two attached course evaluation instruments and proposes to pilot them at the end of the fall 2005 semester in the classes of a volunteer

group of tenured faculty (preferably full professors). The committee recommends participation of classes representing departments in all three colleges. Faculty participating in the pilot would have the Pilot Course Evaluation Forms administered in at least one class with different members of the class using the different forms. This process would be facilitated by other faculty members or staff support persons (depending on the departments usual procedure) using a protocol provided by the committee. Forms would be submitted for scoring and transcription of written comments following current procedures. Faculty, department staff and students participating in the pilot would be asked to complete a follow up survey asking about their perceptions of the Pilot Course Evaluation Forms and their experience of the process. Results of the pilot and follow up survey would be evaluated and used to 1) make revisions to the forms and 2) produce a document comparing the results from the two forms. Based on the data received, the committee would recommend one of the two forms for implementation. A report of this process along with the recommendation and revised evaluation form would be submitted to the Academic Senate before spring break with the intent of being able to implement the new course evaluation process for Spring 2006. There are additional recommendations with regard to the implementation process that appear at the end of this report.

**Purpose**

The purpose for developing a University-Wide course evaluation form is four-fold:

1 To connect student background variables to course responses.

Research reviewed by the committee documented the importance of gathering data on background variables known to influence student responses so that we can use that data to better interpret the information we receive from course evaluations (see for example, Civian & Brennan, 1996; Schlenker & McKinnon, 1994). The committee examined research on background variables and data bases in assessment centers nation-wide to identify those variables that were most consistently deemed to be influential.

2 To identify dimensions of teaching and learning that are valued within our University community and the literature on teaching in higher education.

Based on review of course evaluations currently in use at HSU and literature related to teaching effectiveness and evaluation (for example Centra 1993; Cress 2000; Dominowski 2002; Goldsmid & Wilson, 1980; Seldon , 1984; ), the committee identified a list of characteristics of good teaching. As work continued evaluating assessment approaches, this list was refined to identify four dimensions of quality instruction that were used to guide the development of a set of core items to be used in all course evaluations across the University. The standardization of a set core items used campus wide would provide data that would identify for students, faculty and administrators dimensions of teaching and learning that are valued within our University community and that reflect current research on teaching in higher education. Care was taken to restrict the content of the items in the core to practices that could be legitimately expected to occur in

all lecture/discussion and seminar classes. The expectation of the committee is that items related to labs, discussion sections and other specialized forms would be added under the supplemental questions. This model is supported by most of the major assessment centers (see for example, PICES at Purdue, Multi-Op at Indiana State, IDEA at KSU).

3 To provide relevant information to candidates and reviewers in the Retention, Tenure and promotion process.

All materials reviewed on faculty evaluation clearly identify the importance of a comprehensive plan for faculty evaluations that includes peer evaluations, self-evaluations, portfolios and other possible forms of evaluative material as well as student evaluations (see for example Arreola, 2000; Cangelosi, 1991; Flannery, 2000). At several points the committee discussed the importance of seeing the student course evaluation component as one piece of this larger evaluative process. In that context it appeared to the committee that the evaluation form should provide concrete information that would assist instructors in further developing and refining their courses and their instructional practices as well as providing evaluative information for the RTP process. To provide for both these needs, the evaluation form must gather data that assesses both generally accepted standards of good practice and data that assesses unique characteristics of particular contents, course structures and instructional approaches. This need further supported the adoption the two-part structure: "Core" items reflecting the dimensions of teaching described above for use in the RTP process and additional supplemental items to be selected or developed by the department or instructor that would be used primarily by the instructor for course development purposes. The committee has some further recommendations in this regard that appear at the end of this report with the recommendations for implementation.

4 To characterize student experience in a course and provide instructors with information helpful in refining course characteristics and instruction.

Review of the literature and the combined expertise of various faculty including committee members, brought to light several critical problems that often arise in student evaluations. Some of these are:

(a) Evaluation questions sometimes ask students to assess dimensions that they do not have the background to assess. For example, students can tell you whether or not an instructor was able to provide answers to questions that went beyond the information given in the text, but they cannot judge whether or not a professor is knowledgeable in his or her field. This later evaluation must be made by academic colleagues.

(b) Evaluation questions often ask students to assess global dimensions such as organization without providing information about what constitutes the characteristics of the dimension. The student is thus left to rate an individual's level of organization against whatever construct the student has developed for this dimension. Students responses will naturally reflect their prior experience more clearly than their knowledge of particular dimensions of instructional organization.

While it is true that relatively high correlations have been found between faculty self ratings and student ratings on some dimensions of this type, the psychometric problem can be avoided by asking students about specific organization related practices.

(c) Primarily two types of scales appear in the literature, behavioral descriptors and rating scales. Behavioral descriptors ask students to select all responses that describe their experience with regard to an item. Such instruments provide a pattern of responses around a core activity. This type of instrument can provide very specific information about student experiences. While this provides information about specified dimensions of interest, the instruments often take substantial time to administer and the information gained is somewhat complex to analyze and interpret. This type of model does not lend itself to numeric rankings – a characteristic that is seen as an asset or a liability depending on one's view of the evaluation process. Since the 1980's rating scales have gained popularity because they are easily administered and scored and do quickly yield a numeric code. Although their use for course evaluations has been refined to very high levels, the problem of comparative rating is inherent in the construction of the items. The most effective way to attempt to minimize this is to write the items as descriptions of student experience (rather than projection of teacher characteristics) and use an "agree – disagree" format. Rating scales also invite ranking or comparative rating of instructors in the RTP process rather than evaluating instructors on the basis of demonstration of particular behaviors considered important in teaching. There are some current authors (Arreola, 2000 as an example) who recommend designing the entire evaluation process so that it results in rankings on a single common scale with the result that the entire evaluation process can be represented with a final numeric outcome.

(d) Questions that ask for overall ratings or assessments give the appearance of measuring a summary impression, but, in fact, are unreliable and more open to unidentified variance than questions that ask about the frequency of particular behaviors. This questions also fail to yield any information useful in course or instruction refinement.

(e) In order to make appropriate use of student evaluations in the overall context of faculty evaluation, it is important that the nature and source of this data be very clear. One structural way to maintain clarity is to have the instrument clearly focus on the fact that the data is a report of the student's experience in a particular class.

The committee developed a set of criteria for designing and evaluating items that take into consideration what we learned about the difficulties with item structure and content. The criteria list is provided below.

**Structure**

The two proposed course evaluation forms are divided into three sections: background information, core items and supplemental items. The committee proposes adoption of the background questions and core items across the campus and development of supplemental items by departments and/or faculty as desired. An item pool could be developed to provide potential items addressing a) special instructional settings such as labs, distance learning courses, and discussion sections, and b) special pedagogical dimensions of interest to particular departments and/or individuals. The committee has had some discussion on this point and could provide items to begin to build an item pool. If the behavioral descriptor approach is taken, it will take time to develop such items. If the rating scale approach is taken, items can be drawn from models provided in a number of established item pools. The committee did not pursue development of a pool of supplemental items due to time constraints.

The proposed course evaluations are both designed to be used with scantron forms. In the future, the evaluation could be readily transformed to an on-line format if the Senate decided to recommend that approach. The committee chair was advised at the meeting in August that the committee would not need to make a recommendation on that issue given the time constraints and level of concerns surrounding this process.

Both of the proposed forms provide space for comments. The behavioral descriptors form provides comment space at the end of each question. Psychometrically this is a highly desirable approach because it elicits more specific data. From the point of view of processing the data, this may be a major challenge. The rating scale form follows the accepted practice of providing open ended questions and space for comments at the end of the form. The committee plans consultation with staff on the issue of transcribing comments as a part of the follow up to the Pilot. The committee therefore strongly recommends that staff be involved in the piloting of the instruments and the evaluation of the pilot results.

**Core Items**

Items for each of the instruments were designed to probe particular student experiences related to four dimensions of teaching: Organization, Knowledge/Content, Instructional Practices (including teacher/student interactions), and Feedback/Assessment. Rather than dealing with issues related to diversity as a separate section, the committee included relevant items within the four categories.

The following criteria were used to evaluate items for the Behavioral Profile:

- Items provide a relevant range of responses
- Items encourage students to report their experiences rather than project a judgment
- Statements are specific and utilize straightforward construction
- Statements focus on student's own experience, not the experience of others
- Statements avoid a sense of progression in responses, especially avoid suggesting a judgmental progression from good to bad or vice versa
- Statements avoid rankings or comparisons

- Statements avoid words that appear to have relevance but are actually judgmental
- Statements avoid language that requires personal interpretation (words like "respect")

The following criteria were use to evaluate items for the Rating Scale:

- Each item refers to only one issue
- The response scale matches the language of the items
- The response scale is balanced
- There are at least eight items for each domain

**Considerations with regard to the Pilot**

Piloting the two different instruments would allow the committee to provide concrete information to the Senate with regard to five important characteristics: validity, reliability, administration, scoring and transcription and reporting of results.

1. Validity

For both types of instruments, the Behavioral Profile and the Rating Scale, validity is generally established using a combination of face or content validity (the degree to which the items appear logically to refer to the identified dimensions of teaching) and concurrent or empirical validity. Concurrent validities are usually established by comparing the results of the evaluation instruments with peer evaluations and/or self-evaluations structured on the same dimensions. While undertaking a thorough investigation of concurrent validity is probably not within our resources, follow-up surveys of participating faculty and students could collect information about their views of the face validity of the instruments.

2. Reliability

Two types of reliability are generally examined in the construction of any psychometric instrument: internal consistency and test-retest reliability.

Internal consistency can be viewed in two ways. One approach is to construct an instrument that has paired items so that responses on the items can be compared to demonstrate that the paired items are being answered in consistent ways. Split-half reliability is an example of this type of approach. Another approach is to conduct an item analysis to examine whether or not items related to a domain are functioning in similar ways. This can be done using either correlational or factor analytic approaches. The correlational approach is more appropriate to the rating scale as these represent a continuous variable model and a factor analytic approach is more appropriate to the behavioral profile model as this is a discrete variable approach. The instruments are not designed to be evaluated using a paired item analysis, but can and should be subjected to item load analysis.

Unfortunately, there probably is not sufficient time to conduct a test-retest reliability study. Such a study would involve doing the evaluation twice in the same classes using essentially the same forms with only minor variations to reduce possible double sampling effects. There are also some problems with test-retest reliability for course evaluations. The most pressing problem is that the literature and research is not consistent with regard to how the timing of the evaluation during the semester effects the outcome.

3. Administration of the instruments

Since the structure of the two instruments is quite different, information should be collected on how the students respond to each of the structures. Students may find one form more approachable than that other. The behavioral profile approach may require more time and reflection, since the student must read each response and decided whether or not it applies to her or his experience. This may be viewed as an asset or a liability by faculty and students, and their views on this issue will be solicited on the surveys. Information brought to the committee from the Evaluation Workshop indicated that course evaluations should require no more than 20 minutes to complete.

4. Scoring and transcription

Scantron forms will be read and comments transcribed through the usual processes. The behavioral profile instrument will produce frequency counts of the responses on each item. These can be converted to percentages to indicate what percent of the members of a class gave each of the possible responses on each item. The rating scale approach adds up the rating and produces an average for each item.

Transcription of the comments on the two instruments could potentially be quite different. This is an area where it will be critical for staff to provide feedback on the comparative complexity and demands of the task presented by each of the instruments.

5. Reporting of the result of the evaluation

The pilot will also provide faculty and administrators the opportunity to examine the types of reports that are possible from the two different evaluation approaches. The behavior profile approach includes within each item responses that indicate teaching strengths and responses that indicate potential problems. Results can be presented in three ways. Percentages can be provided for all responses; the report can list only those responses given by an identified percentage of the class, of the report can list only the top three or four most common responses. For purposes of the pilot, it would be appropriate to provide percentages for all responses, but after viewing the reports there may be recommendations for use of a summary form of report. The reports from the rating scales are the typical numeric summaries that we have been using.

The advantage to standardizing the evaluation form would be that everyone would be using the same scale so the summaries could all be interpreted in the same way.

Based on the analysis of the evaluation results and the surveys, information can be provided to the Senate about the characteristics and usability of each of the forms. A recommendation would be made based on how each instrument performed in each of the categories described above.

**Specific Recommendations**

The committee recommends the following plan.

1. Conduct a pilot study using both the behavioral and rating scale formats at the end of fall semester 2005 including surveys of students, faculty and staff.
2. Present report of pilot and final recommendations to Senate by March of 2006.
3. In collaboration with Associated Students, development of an approach to helping students better understand and utilize the student course evaluation process.
4. Develop an integrated view of the evaluation process including specifications of the dimensions that are appropriate to each of the types of evaluation information included in the WPAF: student evaluations, peer evaluations, self-evaluations, sample syllabi and any other documentation, and the relationships between these evaluation sources for each of the areas of evaluation (Teaching, Scholarship, University and Community Service).
5. Examine the linkage between evaluation of teaching and faculty development.

**Report of the Course Evaluation Subcommittee**
**Reference List**

Arreola, R. (2000). *Developing a Comprehensive Faculty Evaluation System.* (2nd Edition) Bolton, MA: Anker Publishing Company, Inc.

Cangelosi, J. (1991). *Evaluating Classroom Instruction.* New York: Longman Publishing Group.

Centra, J. (1993). *Reflective Faculty Evaluation*. San Francisco: Jossey-Bass Publishers

Civian, J.; Brennan, R. *Student and Course Factors Predicting Satisfaction in Undergraduate Courses at Harvard University*. Paper presented at the Annual Meeting of the American Educational research Association (New York, NY, April 8-12, 1996) (ERIC Document Reproduction Service No. HE029134)

Dominowski, R. (2002) *Teaching Undergraduates*. MahWah, NJ: Lawrence ealrbaum Associates Publishing.

Doyle, K. (1975). *Student Evaluation of Instruction*. Lexington, MA: Lexington Books.

Flanner, K (Fall 2000). Contextualizing Course Evaluations: Using Student' Self-Evaluation. *ADE Bulletin 126* pp. 53-57 downloaded from http://www.mla.org/ade/bulletin/n126/126053.htm 10/11/2005

Goldsmid, C and Wilson, E. *Passing on Sociology: the teaching of a discipline.* American Sociological Association.

Laurillard, D. (2002). *Rethinking University Teaching* (2nd Edition). New York: RoutledgeFalmer.

Loftis, RR. (2003). *Evaluation of an Instructional Effectiveness Instrument*. Unpublished master's thesis, Humboldt State University, Arcata.

Schlenker, D. and McKinnon, N. (1994). *Assessing Faculty Performance Using the Student Evaluation of Instruction.* Atlantic Baptist College. (ERIC Document reproduction Service No. HE027508)

Seldin, P. and Associates. (1995). *Improving College Teaching*. Bolton, MA: Anker Publishing Company, Inc.

Seldin, P. (1984) *Changing Practices in Faculty Evaluation*. San Francisco: Jossey-Bass Publishers.

Theall, M and Franklin, J. Student Ratings of Instruction: Issues for Improving Practice.*Monographs of New Directions of Teaching and Learning.* , 43 (Fall 1990).