Memo

To: Lori Kletzer, Vice President for Academic Affair and Dean of Faculty

From: Course Evaluation Committee (Syed Tariq Ahmad, Audrey Brunetaux, Sahan Dissanayake, Paul Greenwood, D. Whitney King, Liam O'Brien, Ray Phillips, Kevin Rice, Stephanie Taylor, and William Wilson)

Date: May 2014

Re: Evaluation of Faculty Course Evaluation for possible bias.

The Course Evaluation Committee (CEC) has the responsibility to periodically review course evaluation data for evidence of bias. Over the past year, the CEC has reviewed the past reports prepared for the committee and has completed additional analysis of more recent teaching evaluations.

The past reports, each authored by Colby’s Office for Institutional Research, were submitted in 2005 and 2007 and are attached as appendices to this report.

Summaries of past findings:

- For the 2005 report, course evaluations from fall 2000 through spring 2005 for courses taught by tenure-track faculty were studied. The data, aggregated by course, were analyzed by faculty gender, division, and years since hire. While the data suggest there are no meaningful differences in course evaluations based on the gender of the faculty instructor, there were two patterns worth noting. In the natural sciences, female tenure-track faculty in their first and second years at Colby received somewhat lower evaluations than their male counterparts. However, this difference disappeared by year three. Analogously, in the humanities, males lagged their female counterparts for the first two years, but this difference disappeared for the most part by year three. There was no student gender information in these data as the data were aggregated by course.

- For the 2007 report, course evaluations from fall 2004 through spring 2007 for courses taught by tenure-track faculty were studied. The unit of analysis was the individual student evaluation - the data were not aggregated by course. There is some evidence that, in the Humanities, students assign higher ratings to female professors than to male professors. In the Social Sciences, males receive higher evaluations on the “teaching was effective item” (Q8). Other items reveal this pattern as well, but in no other cases was the difference statistically significant. This study showed some interaction effects suggesting that male students evaluate faculty differently than female students.
In preparing our evaluation, the Committee recognized three significant limitations in the past evaluations.

1. When the data are grouped by course (2005 report), all courses are weighted equally, ignoring different student experiences in large versus small courses. Differences in evaluations based on course size will be magnified by this method of analysis. The 2007 study solved this problem by looking at trends of all evaluations. Using individual evaluations also allows a study of the effect of student gender on evaluation.

2. Students evaluate courses by scoring faculty on a discrete five point scale (1, 2, 3, 4, and 5), with almost all responses being 1 or 2 ("Strongly Agree" or "Agree"). Past studies then evaluate differences between faculty groups by averaging the scores of many evaluations. In statistical terms, the analytic techniques utilized previously have assumed that the distribution of these scores was approximately normal. This is not the case since the responses follow a discrete multinomial distribution. Although proper techniques exist for the analysis of such a response, they have not been applied to the course evaluation data. Is an agree (a score of 2) twice as bad of a score of strongly agree (a score of one)? Simple averages may unduly amplify minor differences in the minds of student evaluators.

3. It is difficult, if not impossible, to evaluate for bias when we have no independent metric of teaching effectiveness. Many factors influence student responses on the course evaluations. Teaching effectiveness, even if we could agree on what that means, may not be the most significant factor in assigning scores on course evaluations. So what are we testing when we compare evaluation averages between groups of faculty? Certainly not only the presence of bias in student reporting of teaching effectiveness. Past evaluations used all tenure-track faculty in the analysis. We know that not all faculty in this group are considered equally accomplished teachers, given that some were asked to leave Colby due to poor teaching. We also understand that teaching improves, or at least changes, with experience. Given the narrow range of real responses on the evaluations (1 or 2 on the 5 point scale) it is also possible that poor performance of a few can skew the average results in significant ways. In other words, using course evaluation averages to score individual faculty or groups of faculty is an expedient, but a highly suspect form of analysis.

With these limitations in mind, the committee worked hard to evaluate a group of "accomplished faculty" for bias, or at least difference, in student evaluations of teaching based on gender. We limited our analysis to Associate Professors that have been at rank for less than seven years. We are assuming that all tenured faculty are effective
teachers and remain effective for at least seven years. We collected all evaluations from this group over the period Fall 2010 through Spring 2012 for a total of 4337 course evaluations, 2483 females and 1740 males. This number is reduced slightly by missing values for both gender and Question 8 ratings. Separate evaluations for laboratories and discussion sessions were also omitted. Histograms for all evaluations are shown in Figures 1A and 1B.

Figure 1 A. Percentage histogram of Question 8 scores for male and female faculty evaluated by male and female students. Plots of other questions show the same pattern and are listed in Appendix III. (The results for male faculty are in the forward row.)

As seen in Figure 1B, Female faculty have fewer total evaluations and there are more evaluations from female students for both male and female faculty due to slightly larger number of enrolled women and a higher evaluation response rate by female students. As expected, the figure shows that most evaluations fall in the first and second categories ("Strongly Agree" or "Agree").
Figure 1 B. Number histogram of Question 8 scores for male and female faculty evaluated by male and female students. Plots of other questions show the same pattern and are listed in Appendix III. (The results for male faculty are in the forward row.)

We argue that it is difficult to discern any difference between male and female faculty evaluated by male or female students in this presentation. Significantly, this is the same data format (without the colorful bars) provided to the Promotion and Tenure and Departmental Review Committees for their evaluation of candidates teaching evaluations (The evaluation summaries report percentages for each category).

We can, with significant reservations, also calculate evaluation scores by assigning a linear scale to each evaluation category and computing the means for different groups of evaluations. This parametric analysis is shown in Figures 2A and 2B and does indicate some small difference between the evaluation of male and female faculty by male students if viewed on a very expanded scale. This difference may be due to bias in evaluation, but will also be influenced by academic division, class size, and course level - all factors that are not equally sampled by faculty gender.
Figure 2A. Average of question 8 scores for male and female faculty evaluated by male and female students. This figure uses an expanded Y scale to highlight the differences in results. A score of 1 is better.

Of course, this analysis is subject to the same problems we raised earlier. Namely, we have used a parametric analysis to evaluate data that do not satisfy the assumptions of such methods. As a practical matter, the committee is not aware of calculated evaluation scores being used by any group at the college to evaluate faculty teaching effectiveness.
Based on our analysis we make the following observations/recommendations.

1. We are not able to evaluate Colby course evaluation data for the presence of bias by student evaluators because we do not have an independent measurement of teaching quality. We can observe that evaluation patterns for male and female faculty are similar for Associate Professors with a small gender difference in average evaluation scores for faculty evaluated by male students. This difference could be caused by many factors such as major, class size, and course level - as well as any biases associated with gender and age - all operating independently or in combination.

2. The evaluation has no method to actually assess what students have learned or the extent to which established learning goals have been met. At best, the evaluations broadly evaluate student satisfaction with the mechanics of teaching a course. Measuring student satisfaction is still important, and we observe that most students are very satisfied with their classroom experience as illustrated in the results histogram shown in Figure 1.
3. Computing a score for student satisfaction based on averages of the evaluation data is inappropriate and should be discouraged at all levels of evaluation. Presenting a histogram of evaluation responses for each instructor, course, division, and college could be an effective way to show broad trends in student satisfaction without overinterpreting the evaluation data. It is more appropriate to consider the evaluation data as a blunt tool to identify courses with significant student dissatisfaction rather than a refined instrument capable of testing subtle differences in teaching effectiveness.

4. The more the committee works with the evaluation data, the more we are discouraged by the content, utility, and use of the data for assessing teaching effectiveness. Systematically evaluating what is really possible to assess using Faculty Course Evaluations should be the first step toward an improved assessment tool. This analysis has not incorporated the additional comments that students may add to the course evaluation forms. Creating a mechanism that promotes thoughtful narrative instead of simple numerical scores would be helpful for the faculty instructor and others that use the evaluation for personnel decisions.