

# Report and Recommendations of the Teaching Evaluation Finalization Task Force

August 5, 2015

## Members:

Craig Smith, Chair

Anita Mahadevan-Jansen

Andy Van Schaack

Liz Zechmeister

## Background

In October of 2104, the Final Report of the Committee to Reevaluate Evaluations (EvalReval) was released to the University Faculty, and a period of public comment within each of the four undergraduate schools ensued. The present task force was assembled at the request of Vice Provosts Tim McNamara and Cynthia Cyrus to formulate recommendations for the finalization and implementation of the new Teaching Evaluation Instrument in light of the feedback obtained from each of the schools. We have carefully considered, and extensively discussed the new instrument in light of this feedback, and have recommendations in the following three areas, each of which we discuss in turn: 1) The final structure of the Teaching Evaluation instrument; 2) Considerations surrounding the use of this instrument, and the presentation of the resulting evaluations; 3) Issues regarding the calibration and validation of the Teaching Evaluation instrument.

## Recommendations on the final structure of the Teaching Evaluation Instrument

Faculty across the four schools made a number of recommendations and observations regarding the items on the evaluation instrument. These included suggestions of additions to the instrument; concerns about the format of the response scales used on certain items (particularly the “administrative” items assessing overall quality of the instructor and course); as well as concerns and suggestions regarding the specific wording of particular items. Based on a careful consideration of these concerns and suggestions, we have the following recommendations for the final structure of the instrument. Attached to this report is a mock-up of the instrument as proposed by the original EvalReval committee (henceforth referred to as the “EvalReval” Instrument), and an updated mockup incorporating our recommendations (henceforth referred to as the “recommended” instrument).

- 1) *Add back an item assessing the amount learned in the course as a third administrative item*, as this is an item that one of the four schools (Peabody College) has heavily relied upon in tenure and promotion decisions. The item we propose adding is now item #13 on the recommended instrument: “Compared to other classes, the amount I learned in this course was:” This is a rewording of the original “amount learned” item to better fit the format of the new instrument.

- 2) *Revert to using the original rating anchors for the two administrative items assessing the overall quality of the instructor and course* (items # 11 and 12 on the recommended instrument). The scale anchors suggested by the EvalReval committee were criticized as being unbalanced because there was only one unambiguously negative response “poor” and at least three positive responses “Effective” or better. The Task Force recommends reverting to the scales originally used for this item “poor, marginal, average, very good, and excellent” because they are a more balanced set; this set of scales remains appropriate for the structure of these items; and using the same response scales as the previous instrument minimizes the probability that the response properties of these items will be greatly different from the previous items. This last point has relevance for the issues surrounding the calibration of the instrument noted below.
- 3) *Make minor rewordings to two of the items on the EvalReval Instrument (#1 and #13) to clarify their meaning.* In the course of our deliberations, the task force carefully considered each item, the construct it was intended to assess, its clarity, etc. Items that were critiqued in the feedback obtained from the Faculty were scrutinized especially carefully. Overall, the Task Force came away impressed with the quality of the proposed items, finding them to be well-structured and clear. However, minor rewordings were suggested for two items:
  - a. For item #1, “the” was deleted from the phrase “helped me understand the core ideas and issues in this subject” because it was noted that introductory-level students might not have a good understanding of what **the** core ideas and issues actually are. Although a very minor edit, it was felt that especially for the very first item on the instrument, the item meaning should be as clear as possible.
  - b. For item #13, assessing the assigned workload in the course, the phrase “required to do well” was changed to “assigned” because it was felt the former phrase would likely elicit more idiosyncratic responses as a function of a respondent’s abilities in the course, with the revised item being a cleaner assessment of the actual workload associated with the course. The task force believed that the reworded item could provide better formative feedback to instructors who were interested in calibrating the workload associated with their course to other courses in the university.
- 4) *Retain all of the close-ended items, making the instrument one item longer.* One faculty concern the Task Force took very seriously was that many believed the instrument to be too long. Therefore, the Task Force members carefully considered whether each close-ended item might be deleted. However, we found the instrument to be very well constructed with very little redundancy across items, with each item serving an articulable purpose. If the Task Force were to target one item for elimination it would be item #13, discussed above. This is because the desirability of the responses is much more contextualized for this item than for others (e.g., it was noted that a heavier than average workload might be highly desirable in certain courses, such as lab sciences,

whereas lighter workloads might be desirable in others, such as small discussion-based seminars). However, the overall consensus of the Task Force was that individual instructors could take this context into account in interpreting responses to this item, and that the formative feedback it offered justified retaining this item.

- 5) *Do not add any additional close-ended items beyond the “amount learned” one.* The Task Force also extensively considered whether any additional items should be added to the instrument. In the obtained Faculty feedback, there were multiple suggestions that an item assessing the the amount of effort the student invested in the course (e.g., the number of hours the student devoted to the course each week), and/or the degree to which the student was inspired to become intellectually engaged in the course material be added. Although the Task Force saw considerable merit to these items, the final consensus was that they did not contribute enough beyond what was already on the instrument to justify their addition, especially given the concerns expressed about instrument length. However, the Task Force does believe that it would be worthwhile to ensure that such items are included in the optional item bank proposed by the EvalReval Committee.
  
- 6) *Eliminate the first open ended item: “What advice would you offer to a student who plans to enroll in and hopes to do well in this course?”* This item was included on the instrument with the express purpose that the responses to this item would be made available to students, along with the quantitative data yielded by the close-ended items. The representatives from Vanderbilt Student Government on the EvalReval committee clearly placed high importance on this information being made available. However, the faculty have expressed a lot of concern about the publication of the open-ended responses. Opinion on the EvalReval Committee on this issue was mixed, and never reached true consensus, and the faculty feedback obtained in response to the EvalReval report indicated high levels of concern about publishing these responses. The faculty of the School of Engineering formally voted to recommend against this publication, and multiple faculty in each of the other three schools expressed strong reservations. If the responses to this item are not to be published, the Task Force does not believe the item should be retained on the instrument because its express purpose is to provide provide the information elicited by this question to students. We recommend against including the item, in light of faculty concerns, for the following reasons:
  - a. A broad consensus, both within the original EvalReval Committee, and across the Faculty providing feedback, is that if the responses are to be published, they will need to be screened and edited for inappropriate content before publication.
  - b. Such screening and editing will require significant resources.
  - c. Such screening and editing will necessarily be quite subjective.
  - d. The fact that only screened, and possibly edited, comments would be published likely opens the Administration to charges of censorship and suppression from the very students we would be trying to please by providing this information.
  - e. Elimination of this open-ended item would make the instrument appreciably shorter, at a functional level, than the originally proposed one.

Nonetheless, the Task Force does recommend that the quantitative data resulting from the close-ended items should continue to be released to students. We also recommend that in addition to presenting students with means and standard deviations, that we also provide students with more detailed information on the response distributions for the items. Because the close-ended items on the recommended instrument are substantially more descriptive than those on the currently used instrument, the quantitative data from the recommended instrument will be richer and more informative to the students, and thus will at least partially replace the more subjective open-ended responses we recommend eliminating.

### **Considerations Surrounding the Implementation and Use of the Revised Instrument**

- 1) *The Task Force strongly endorses the EvalReval Committee's Recommendations regarding working to change the culture of evaluations at Vanderbilt.* It is our considered opinion that beyond improving the evaluation instrument itself, as we believe that the EvalReval Committee, along with our recommendations, has done, the single most effective means of improving both the response rates associated with course evaluations, and the quality of the information provided by individual respondents will be to foster a culture in which participants (both faculty and students) understand the importance of teaching evaluations and their various informative, formative, and evaluative uses, and in which respondents are given good opportunities to take their time to complete the evaluations. Toward this end we recommend:
  - a. Instructors be formally encouraged to take some class time to explain the importance of the teaching evaluations in the context of their class, and to arrange for their students to complete the evaluations during class time. A couple of members of the Task Force have taken both of these steps in their own classes, and some of us have encouraged colleagues to also take these steps, and in every case we have been very pleased with the results. Response rates have been very high in these classes, and the thoughtfulness of the responses, particularly evident in the written comments, have been high.
  - b. A website addressing the issues of the importance and uses of evaluations should be developed, as recommended by the EvalReval committee. The Task Force sees such a website as being valuable to both interested faculty and students, and especially valuable for Faculty who want to follow our recommendation to explain the importance of evaluations to their students.
- 2) *The Task Force endorses the recommendation that faculty should be provided with the option to view the profile of the responses given by individual respondents.* Such profiles would allow faculty to differentiate between poor evaluations across items due to a small number of individuals who were generally dissatisfied with the course, or whether different individuals were dissatisfied with different aspects of the course, etc.

- 3) The task force also recommends that instructors be given the option of gathering mid-semester course evaluations in their classes using the same instrument, for purely formative feedback. The consensus of the task force is that many faculty, particularly junior faculty, will find this mid-term feedback to be quite helpful in improving their teaching, and in making mid-semester corrections to non-optimal aspects of their courses. The task force recommends that such mid-semester should be performed at the discretion of the instructor. More importantly, since the purpose of such an evaluation is purely for formative feedback, data from these evaluations should not be included in the instructor's dossier.

### **Issues regarding the calibration and validation of the evaluation instrument**

- 1) *Calibration.* In their feedback, a number of faculty voiced concerns that given the changes to the evaluation instrument it was quite possible that the response distributions to the administrative items used in promotion and retention decisions (overall quality of instructor and course, and amount learned in the course) might change, and that any such changes should be documented so that these changes could be taken into account when using the new instrument administratively. We note the following with regard to this issue:
  - a. This concern is strongest for the "learning" item because both the wording of the item and the response scale have changed.
  - b. Less potential change is likely for the other two items because their wordings are very similar to the previous wordings, and we have recommended that we retain the original response scales for these items.
  - c. Nonetheless, beginning with the first administration of the new instrument, and moving forward, this calibration issue should be empirically examined. Because there will never be a time-point in which both instruments are administered, and thus can be directly compared, we note that some sort of time series analysis, in which one looks for discontinuities in the ratings with the introduction of the new instrument are required.
    - i. Such analyses can examine overall trends in the instrument's behavior by aggregating across all courses, or certain classes of courses (lecture courses, seminars, etc.).
    - ii. More fine-grained and precise estimates of the effects of changing the instrument can be obtained by analysis of response profiles (either at an individual level or aggregated) for instructors who have taught the same course(s) both before and after the change in instrumentation.
  - d. Beyond simple calibration the Task Force believes that these analyses represent an important opportunity to examine the properties of the revised instrument with regard to rating biases that are known or believed to characterize teaching

evaluations, including gender biases, ethnic biases, as well as differences driven by the type of class (large lecture class vs small seminar; introductory class vs. more advanced class, etc.). In conducting the calibration analyses, the Task Force recommends that systematic variations in ratings associated with such factors, especially for the administrative items, be documented. Knowing such differences are important in interpreting the evaluations received for a particular instructor for a particular course of a certain type. One intriguing hypothesis advanced by a Task Force member is that, if we are successful in altering the culture of evaluation at Vanderbilt, and in eliciting more thoughtful responses from a broader cross-section of our students, we might expect overall evaluation differences attributable to factors like gender and ethnicity to decrease over time.

- e. The Task force wondered whether these analyses might comprise a research project suitable for an EdD student to take on. If not, it might be something that a Masters Level student with suitable quantitative skills could take on, within a part-time position, under the supervision of a small faculty committee.

2) *Validation.* In response to the EvalReval report several faculty noted that a Research I institution like Vanderbilt should systematically attempt to assess the validity of the teaching evaluation instrument. Questions to be addressed would be such things as whether the instrument reliably assessed anything beyond a general factor of student satisfaction. For instance, does the item assessing the amount learned in a course predict performance in more advanced courses in the evaluated course's field by students who pursue such advanced coursework, etc.? The Task Force endorses this observation and makes the following two recommendations.

- a. First, to address the issue of user satisfaction with the new instrument, and its perceived informativeness, the Task force recommends that a sample of users (both instructors and students) be surveyed this fall regarding how satisfied they are with the new evaluation instrument and how informative they find the information it yields to be. The EvalReval committee used focus groups to ensure that the new instrument would be informative, and informal surveys of colleagues and students by a member of the task force suggests a high degree of satisfaction with the new instrument and the information it yields. Nonetheless, in line with the EvalReval Committee's original suggestions, systematically obtaining data to examine this issue seems important.
- b. The validity of the instrument, both in terms of capturing information about teaching that goes beyond student satisfaction, and in predicting outcomes indicative of high quality teaching and learning, should also be systematically examined. Designing good validation studies is a very sophisticated affair that the Task Force believes extends beyond its resources. Therefore, we do not have firm recommendations about what such studies should look like. However, we

do recommend that after the new instrument has been in use for several cycles the Provost's Office put out a call for faculty proposals for studies to examine this issue, and then that it commission at least one such study.