Report to CEPP on Quantitative Student Ratings of Faculty
May 3, 2010
Rik Scarce

Summary
While the current student quantitative faculty ratings instrument (the "Dean's Card") is appropriate for personnel purposes, the literature indicates that if those evaluations are intended to be used by faculty to improve teaching, a new instrument needs to be devised. In addition, faculty will require additional institutional support to take full advantage of the benefits flowing from a new ratings instrument.

Background
For several years, members of CEPP have expressed curiosity, and even concern, about the quality and effectiveness of the Dean's Card. At the Committee's Spring 2009 retreat, I volunteered to research the state of the art in faculty evaluations for the Committee. Through the course of this research, I communicated with Paty Rubio, Jeff Segrave, and Catherine Ross (Associate Director, Institute for Teaching and Learning, University of Connecticut). In addition, I undertook an extensive review of the research on faculty evaluations, and I consulted multiple relevant websites.

A Word about Terminology
Although we commonly refer to the Dean's Card results as "evaluations," more and more scholars prefer the phrase student "ratings" of instructors. Arguably the foremost authority on the subject, William E. Cashin (1995), wrote, "I suggest that the term 'student ratings' is preferable to 'student evaluations.' 'Evaluation' has a definitive and terminal connotation; it suggests that we have an answer. 'Rating' implies that we have data which need to be interpreted. Using the term 'rating' rather than 'evaluation' helps to distinguish between the people who provide the information (sources of data) and the people who interpret it in combination with other sources of data (evaluators).

"Viewing student ratings as data rather than as evaluations may also help to put them in proper perspective. Writers on faculty evaluation are almost universal in recommending the use of *multiple* sources of data." Cashin's argument appears to have gained considerable "traction" in recent years, with increasing numbers of publications referring to student ratings rather than to evaluations. As such, I use "ratings" throughout.

Creation of the Current Dean's Card
The Dean's Card instrument most of us are familiar with was approved for use in 1992 as part of a sweeping set of tenure and promotion policy changes recommended by the Committee on Promotion and Tenure (CAPT, 1991). Previously, a six-item instrument—also known colloquially as the Dean's Card—was used in all courses; on a five-point scale ranging from Poor to Excellent, students were asked to rate each instructor's "Mastery of the subject matter," "Degree of preparation for class," "Ability to convey the material," "Enthusiasm/Interest in material," "Availability for outside help," and "Overall rating of the instructor."

In our conversation, Jeff Segrave—who, with Ralph Ciancio, was one of two CAPT chairs involved in the CAPT policy changes that ushered in the current Dean's Card—noted that there was considerable consternation regarding what the numbers from the old evaluations meant, what was actually being measured, and whether the instrument was appropriate for the task at-hand. An effort in the 1985-86 academic year to revise the instrument failed (CAPT, 1991). The current Dean's Card resulted from the critique of the prior Dean's Card in the early '90s and from an extensive examination of the literature.

Theoretical Tensions and Practical Concerns

The literature on faculty ratings is vast and goes back decades. It is a constantly evolving field, but some of its tenets are well established. Among the most important, and most accepted, observations in the literature is the identification of two closely-related tensions that frame the use of student ratings. One is between the uses of ratings for evaluation and development, the other between the formative and summative purposes of the ratings. When applied to the Skidmore setting, these tensions in the literature point to a small but weighty number of practical concerns.

Evaluation versus Development

Within the literature, "evaluation" refers to the process of amassing information about teaching from all possible sources and judging the quality of teaching based upon that information. Evaluation, then, is the province of personnel committees and others charged with evaluating a candidate's file for continuation, promotion, or tenure. In a frequently-cited paper, Cashin (1990) argues that "a few global or summary items or scores" should be used in student rating instruments intended to evaluate instructor performance because "student rating items tend to correlate more highly with student learning than do more specific items.

"Suggested summary items are:
1) Overall, how effective was the instructor?
2) Overall, how worthwhile was the course?
3) Overall, how much did you learn?"

Those items are, of course, the three used in the Dean's Card. Cashin goes on to write that "such items would serve the purpose of evaluation, which is to decide how well the instructor taught (*not* what he or she might do to improve—which is the focus of development)" (Cashin, 1990). Cashin's "evaluation"-"development" distinction is a crucial one. Personnel committees' information needs are narrow, Cashin suggests: of the sort that those three questions may satisfy.

In contrast, as it relates to student ratings, faculty development is a more complex matter, one that three global questions cannot address and—according to Cashin (1995) and others—they were not intended to address. Faculty development requires additional, comparatively specific questions that will aid faculty in the improvement of their instruction but that may be too specific to be of interest to evaluators.

Summative versus Formative

The evaluation-development tension addresses the *purposes* of student ratings of instructors. The summative-formative tension essentially adds a *timing* component evaluation-development. Instruments with a summative purpose provide key data for personnel decisions and *are administered at the end of a course*. On the other hand, instruments with "formative" purposes are often *administered during the term*, usually only once at roughly the mid-point, but conceivably at any time or multiple times. Moreover, formative efforts are usually embedded in broader faculty development efforts—"systems," as I refer to them, below.

In describing the summative-formative distinction, Edward Nuhfer writes, "Summative evaluations given at the end of a course are direct measures of student satisfaction. 'Satisfaction' is the sum of complex factors that include learning, teaching traits, and affective personal reactions that are products of both what happens in a class and what an individual has brought with him or her to the class in form of bias and motivation. Formative evaluations given during the ongoing course, usually about mid-term, ask detailed questions that provide a profile of pedagogy and strategy being employed. There is plenty of evidence to show that the functions of the two kinds of evaluation must be clearly separated" (Nuhfer, 2003).

A UCLA website presents that distinction more starkly: "Teaching evaluations are commonly considered for summative purposes, including tenure, merit increase, retention for non-tenured faculty, promotion, and course assignment decisions…. Using evaluations to inform instructors of their teaching effectiveness and to aid them in improving or enhancing their teaching constitute the formative purposes of teaching evaluations" (UCLA OID, N.d.).

Conclusions

Two conclusions emerge from my reading of the literature. First, the Dean's Card is an inadequate tool for the job, if that job is improving faculty teaching. The Dean's Card *may be* appropriate for evaluation and for summative purposes. Indeed, that was its explicit intent when it was put in place in 1992 (CAPT, 1991). (Viewed another way, *the current Dean's Card was never intended to be used by faculty to highlight the strengths and weaknesses in their teaching.*) However, none other than Cashin (1995) noted, "Although there is general agreement that student ratings should be used *when their purpose is to improve teaching*, there is disagreement about how many, or which, dimensions should be used *for personnel decisions*" (emphasis in original). As such, Cashin noted that scholars had called into question his own conclusions from 1990; recall that those conclusions are reflected in the current Dean's Card.

Second, Skidmore does not appear to address the core of the summative-formative distinction. We *do* have the beginnings of a formative *system*—such as pedagogy sessions, mentoring, and substantial support for first-year faculty thanks to recent initiatives by the First-Year Experience. However, other aspects of such a system along the lines mentioned in the literature, whereby faculty might be *institutionally* encouraged and supported to obtain in-term student ratings and reflect on the results of those ratings

with faculty outside of their department, are lacking at the College. Such support might include making available statistically reliable and valid student rating instruments (and the computation of the results from those instruments) to faculty who wish to use them, as well as other support like expert mentoring to assist younger faculty in interpreting in-semester ratings.

It is also important to note that *final* course ratings by students—not only the in-semester ratings mentioned in the previous paragraph and noted in the literature cited above—may include more items than narrowly constructed, global indicators such as the "Overall…" questions now in use at Skidmore. An article in *College Teaching* observed, "In addition to items designed to target specific dimensions or behaviors, the end-of-course evaluation forms also tend to include global items on the overall effectiveness of the teacher or the quality of the course" (Hobson and Talbot, 2001: 27). In addition, in an e-mail exchange with Catherine Ross, Associate Director of the University of Connecticut's Institution for Teaching and Learning, she reacted to Skidmore's current student ratings instrument by writing, "I am looking for how this instrument would help faculty to improve their teaching, and I don't see any way they could really from the three questions here and this raises another whole set of issues around the purpose of evaluating teaching when there seems to be no help or support built into the system."

Finally, faculty need guidance as they attempt to grasp what they see in in-term and final course evaluations. In the summary of a 1987 U.S. Department of Education literature review, Kathleen Simon put the matter succinctly: "To be used effectively, evaluation forms must be analyzed systematically and the results communicated to the faculty in a way to enhance their professional development. The most important use of student evaluations is to improve classroom instruction" (Simon, 1987). UCLA's Office of Instructional Development suggests a strategy for that communication, writing, "In particular, studies also indicate that mid-semester evaluations and feedback accompanied with consultation from a faculty developer or peer are more effective than traditional practices that leave the instructor to interpret end of semester findings by him/herself" (UCLA OID, N.d.).

Thus, the College's final course ratings may, and almost certainly should, include both global items of the sort that are now in place (even if not those precise items) and the sort of specific items that are absent from our instruments. The advantage of such a combined approach is that the data from the same instrument may be used for summative/evaluative purposes (personnel matters) *and* for developmental/formative purposes (instructor improvement). My review of the literature identified no shortcomings in combining summative and formative items in the same instrument.

Recommendations

--Change our rhetoric. We should begin using "ratings" to refer to student-generated information about instruction and "evaluation" to refer to peer-generated information about instruction.

--Change our forms. The literature indicates that a three-item instrument such as the Dean's Card is inadequate for informing instructors about where they need to improve their teaching and how they may do so. As such, CEPP should seriously consider devising (or purchasing) an appropriate instrument and should consider validity and reliability issues (of the sort extensively discussed in the literature) in doing so.

--Be aware of differences. The instructor's gender and race may be correlated with student ratings of instruction (Smyth, 2009). Keeping this point in mind when developing (or selecting) student rating instruments and, especially, when evaluating the data from new and more elaborate instruments, may improve instructors' understandings of their ratings and may assist all who are involved in personnel decisions in fairly interpreting the data.

--Develop a formative evaluation *system*. CEPP should consider recommending building upon the foundation for a formative evaluation system now in place for early-career faculty members (and perhaps others as well). In addition to final course ratings and peer reviews, such a process might include structured (institutionally-supported) mid-term student ratings of instruction, in-depth analysis of mid-term and end-of-term student ratings, consultations with expert instructors (such as present and past Ciancio Award recipients), and orientation sessions to help faculty grasp the role of ratings and evaluation in their instruction and in the College's personnel processes. (I should note that mid-term ratings and all consultations with expert instructors should probably not be incorporated into personnel decisions.)

--Consult broadly. The current Dean's Card was researched and put into place by CAPT, and CAPT clearly needs to be consulted should changes like those I envision be pursued. The Faculty Development Committee should be consulted as well, as should the Dean of Faculty, the Vice President for Academic Affairs, all departments and programs, on-campus experts on evaluation (Holly Hodgins, for example), and Jeff Segrave.

Resources

Below, I include two lists. The "References" are those items cited above in my report. Most items in the "Bibliography" are from a helpful list that was posted in early 2010 to the POD Network listserv (for faculty developers) that Katherine Ross at the University of Connecticut shared with me; it was originally posted by Ron Berk, a respected scholar of faculty evaluation. My research indicates that there exist numerous helpful compilations of the extensive literature of faculty evaluations, only a few of which appear below.

References

CAPT. 1991. "CAPT Report on Standards and Methods for Evaluating Faculty Performance." Skidmore College.

Cashin, William E. 1996. "Developing an Effective Faculty Evaluation System." IDEA Paper 33. Manhattan, Kansas: Center for Faculty Evaluation and Development. Available on-line at: http://www.theideacenter.org/sites/default/files/Idea_Paper_33.pdf .

Cashin, William E. 1995. "Student Ratings of Teaching: The Research Revisited." IDEA Paper 32. Manhattan, Kansas: Center for Faculty Evaluation and Development. Available on-line at: http://www.theideacenter.org/sites/default/files/Idea_Paper_32.pdf .

Hobson, Suzanne M. and Donna M. Talbot. 2001. "Understanding Student Evaluations: What All Faculty Should Know." *College Teaching* 49(1): 26-31.

Nuhfer, Edward B. 2003. "Of What Value are Student Evaluations?" Available on-line at: http://www.isu.edu/ctl/facultydev/extras/student-evals.html.

Simon, Kathleen 1987. "Using Student Evaluations To Improve Teaching and Make Personnel Decisions." Available on-line at: http://www.eric.ed.gov:80/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED290377&ERICExtSearch_SearchType_0=no&accno=ED290377.

Smyth, Bettye. 2009. "Student ratings of teaching effectiveness for faculty groups based on race and gender." *Education* 129(4): 615-624. (Also available on-line at: http://findarticles.com/p/articles/mi_qa3673/is_4_129/ai_n32067856/).

UCLA OID (Office of Instructional Development). N.d. "Literature Review." Available on-line at: http://www.oid.ucla.edu/publications/evalofinstruction/eval6.

Bibliography

Abrami, P.C. (1989). How should we use student ratings to evaluate teaching? Research in Higher Education,30, 221-227.

Abrami, P.C. & d'Apollonia. (1991). Multidimensional students' evaluations of teaching effectiveness-generalizability of "N=1" research: Comments on Marsh (1991). Journal of Educational Psychology,83, 411-415.

Anonymous. (N.d.) "Career Development: Using Student Evaluations." University of Medicine and Dentistry of New Jersey. Available on-line at: http://cte.umdnj.edu/career_development/career_student_evals.cfm. *Note: a comparatively brief but rich linked bibliography.*

Arreola, R. A. (2007). Developing a Comprehensive Faculty Evaluation System. San Francisco: Jossey-Bass.

Braskamp, L.A. & Ory, J.C. (1994). Assessing Faculty Work: Enhancing individual and

institutional performance. San Francisco: Jossey-Bass.

Cashin, W.E. (1988) Student ratings of teaching: A summary of the research. IDEA Paper No. 20. Center for Faculty Evaluation, Kansas State University.

Cashin, W.E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin & Associates, Changing practices in evaluating teaching. Bolton, MA: Anker Publishing Company.

Centra, J.A. (1993) Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness. San Francisco: Jossey-Bass.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51, 281-309.

Cohen, P.A. (1990). Bring research into practice. In M. Theall, & J. Franklin (Eds.), Student Ratings of instruction: Issues for Improving Practice: New Directions for Teaching and Learning, No. 43 (pp. 123-132). San Francisco: Jossey-Bass.

Feldman, K.A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Research in Higher Education,30, 583-645.

Marsh, H.W. & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart (Ed.), Higher Education: Handbook of theory and research, Vol. 8 (pp. 143-233). New York, NY:Agathon.

Marsh, H.W., & Dunkin, M.J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry & J.C. Smart (Eds.), Effective teaching in higher education: Research and practice (pp. 241-320). New York, NY:Agathon.

Mckeachie,W. J. (1999). Teaching Tips: Strategies, research ,and theory for college and university teachers. (10th ed.). Boston, MA: Houghton Mifflin.

Ory, J.C. & Ryan, K. (2001). How do student ratings measure up to anew validity framework? In M. Theall, P.C. Abrami, & L.A. Mets (Eds.), New Directions for institutional research: No. 109. The student ratings debate: Are they valid? San Francisco: Jossey-Bass.

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