Comments on the Faculty Evaluation Process
for the College of Education at Valdosta State University

Bill Huitt, Department of Psychology and Counseling
September, 1998

As we discuss making changes in the student evaluation of college faculty I believe it is important to critically review relevant literature pertaining to this topic. The fact that multiple sources of data are used in the evaluation process is certainly supported by most researchers in this field (see Seldon & Angelo, 1994) and is to be commended. However, the following comments pertain to the instrument used by students to evaluate faculty as this is an important component of the faculty evaluation process. These comments focus on four critical issues:

1) Using data for formative versus summative purposes;
2) Validity of student ratings;
3) Criterion- versus norm-related standards; and
4) Does "one size" fit all.

The first issue regards the use of data for formative and/or summative purposes. As reported by Gordon (1998) there is unanimity on the view that teaching is complex and multidimensional. The major issue for evaluation purposes is how to investigate and report on a faculty member's teaching effectiveness. Marsh and his colleagues (e.g., 1983, 1984, 1993) believe data should be collected on multiple items and scores reported in terms of composite scores generated from a weighted analysis of those items. Unfortunately, other research does not verify the large number of factors advocated by Marsh (see Cranton & Smith, 1990); at most two and sometimes three factors emerge through factor analysis. This small number of factors generally accounts for 70% to 90% of the variance among faculty.

Abrami and his associates (e.g., 1989, 1990) advocate the use of global statements for summative evaluation purposes. Cashin and Downey (1992) suggest the following two statements (using a 5-point Likert scale of 1 = definitely false and 5 = definitely true):

1) Overall, I rate this INSTRUCTOR an excellent teacher;
2) Overall, I rate this an excellent COURSE.

Cashin and Downey's research shows that these two statements account for over 50% of the variance among instructors. In any case, I find no literature to support the use of data from unweighted single items for summative evaluation purposes (or formative, for that matter) as is done in the COE.

Therefore, I suggest that for summative evaluation purposes the COE should either develop weighted item composite scores based on factor analysis or simply use the two items mentioned above. At present, neither of these items is listed on the COE instrument. If factor analysis is to be used, the level of aggregation should be class means by instructor; this method highlights differences among instructors rather than differences among student raters (see Cranton and Smith, 1990).

The formative/summative discussion raises a related issue: how does one judge teacher effectiveness? McKeachie (1973, 1990, 1997) has consistently pointed out that verifying student ratings relative to learning outcomes is very difficult and has not been done with any great success. This is one reason for the push towards departmental and even individual faculty- or student-generated evaluation instruments; educators might be able to state with some clarity what they believe faculty should be doing in the classroom, but whether that is related to student learning is somewhat unclear. Abbott and his colleagues (1990) report on a process whereby students in a specific course vote on items to be included on the student evaluation form. This may be one alternative to pursue in helping faculty develop formative evaluation data.

Another issue relates to the use of criterion- or norm-referenced evaluation approaches. There appears to be unanimity among researchers in this area that norm-referenced evaluation methods should not be used (e.g., Rifkin, 1995). Rather evaluation should be criterion-referenced with the criterion established by correlating classroom behavior with outcome measures. As previously stated, this is quite difficult to do. In any case, the norm-referenced system used to report data in the COE only compounds the problem of using individual items as previously discussed. Again, I find absolutely no literature that recommends the COE procedure. In fact, the vast majority of the literature I have reviewed suggests just the opposite of the present policy.

A fourth issue relates to the “one size fits all” approach to student evaluation of instructors. The work of Murray, Rushton, and Paunonen (1990) shows how students in different level psychology courses valued different teaching processes. At a minimum, alternative factor structures need to be explored for Freshman/Sophomore-, Junior/Senior-, and Graduate-level courses. In addition, faculty and students do not always agree on what constitutes good teaching (see Feldman, 1988). This highlights the need to have behavioral-anchors for rating items and research demonstrating the correlation between classroom processes and student learning.

In light of the difficulties involved in producing reliable and valid data from student evaluation of faculty, it seems imperative that faculty accept the data provided. Otherwise, they will not use it to make changes in their own teaching or as members of personnel committees. North Carolina State University (1994) has been one of the pioneers in making the evaluation process more relevant to faculty by involving them in the construction of the instrument. Their on-line handbook is a ready reference for the process they have implemented.

In summary, the fact that some student evaluation instruments have demonstrated reliability and/or validity does not mean that all instruments will. It is up to developers of any new instrument to demonstrate its relationship to other known instruments or desired outcome measures. It is generally acknowledged that developers will investigate the psychometric properties of an instrument prior to its actual use. I encourage the College to undertake such a study prior to data being used for faculty evaluation purposes. As professional educators, I dare say we would not tolerate a student using data that had not been subjected to a study of its psychometric properties to be used in a graduate thesis or dissertation. We should tolerate no less when it comes to our professional practice.

References

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A., & Hess, C. W. (1990). Satisfaction with processes of collecting student opinions about instruction: The student perspective. Journal of Educational Psychology, 82(2), 201-206.

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? Research in Higher Education. 80(2), 221-227.

Abrami, P. C., d'Apollonia, S., & Cohen, P. (1990). Validity of student ratings of instruction: What we know and what we do not. Journal of Educational Psychology, 82(2), 219-231.

Cashin, W. E.,& Downey, R. G. (1992). Using global student ratings items for summative evaluations. Journal of educational Psychology, 84(4), 563-572.

Cranton, P., & Smith, R. A. (1990). Reconsidering the unit of analysis: A model of student ratings of instruction. Journal of Educational Psychology, 82(2), 207-212.

Feldman, K. A. (1988). Effective college teaching from the students'and faculty's view: Matched or mismatched priorities? Research in Higher Education, 28(4), 291-344.

Gordon, P. (1998). Student evaluations of college instructors: An overview. Available online: [http://www.valdosta.edu/~whuitt/psy702/files/tcheval.pdf or http://www.valdosta.edu/~whuitt/psy702/files/tcheval.html]

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. Journal of Educational Psychology, 75(1), 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76(5), 707-754.

Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluation of teaching effectiveness: A profile analysis. Journal of Higher Education, 64(1), 1-15.

McKeachie, W. J. (1973). Correlates of students' ratings. In A. L. Sockloff (Ed.). Proceedings: The first invitational conference on faculty effectiveness evaluated by students (pp. 213-218). Temple University.

McKeachie, W. J. (1990). Research on college teaching: The historical background. Journal of Educational Psychology, 82(2), 189-200.

McKeachie, W. J. (1997). Student ratings: The validity of use. American Psychologist, 52(11), 1218-1225.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. Journal of Educational Psychology, 82(2), 250-261.

North Carolina State University handbook for advising and teaching. (1994). Available online: [http://www2.ncsu.edu/ncsu/provost/info/hat/current/ch10/0105.html].

Rifkin, T. (1995). The status and scope of faculty evaluation. (ERIC Reproduction Service No. ED385315).

Seldon, P., & Angelo, T. A. (1997). Assessing and evaluating faculty: When will we ever learn? (To use what we know). Proceedings of the AAHE 1997 Conference on Assessment and Quality Assessing Impact: Evidence and Action.