F A I R L E I G H D I C K I N S O N U N I V E R S I T Y
David Flory, Professor of Physics

MEMORANDUM

A Position Paper on Student Evaluations[1]

Introduction

Student evaluations of faculty are generally conducted for one of three purposes: to allow the administration and faculty to evaluate the effectiveness of an instructor for faculty status or compensation review, to allow the department or the instructor to assess the classroom environment for the improvement of instruction, or to allow students to assess the quality or popularity of a course or an instructor. Each of these purposes is different and each requires a different approach to the evaluation process.

In this position paper I discuss the background and some history of student evaluations at FDU. I also discuss the conditions necessary to have safe, reliable evaluations; faculty concerns that must be addressed; dangers in using and comparing evaluations; the current instrument in use at FDU; and various issues to consider in using and evaluating the results. Finally, I attempt to address a series of concerns that have been raised regarding the *Endeavor* system that we are now using.

Background

In 1978 and 1979 a joint faculty-administration committee[2] spent over a year studying student evaluations of faculty. After reviewing the literature, consulting expert opinion, and deliberating for almost a year the committee concluded that the type of unproven evaluations we had been using should be discontinued, at least for use in the faculty status process. The reason for this recommendation was that all of the evaluation instruments then in use were formative questionnaires intended to provide diagnostic information for the professor. None had been

[1]This position paper represents the personal views and positions of the author.

[2]The AAUP-Administration Committee on Student Evaluation of Faculty submitted its report to the parties on August 8, 1979. The committee was jointly chaired by Fred Gaige, Dean, Becton College, and David Flory, AAUP tri-Campus President.

validated much less designed for the purpose of providing the information needed for summative use in faculty status review. The committee concluded that such forms were too dangerous. Faculty status review cannot be based on data known to be unreliable. Among the guidelines the committee proposed was that

> *Only proven instruments specifically developed for summative purposes should be used for faculty status review. Such development must include extensive testing to demonstrate validity, reliability and to develop norms. In general, formative questions are neither valid nor reliable for summative use.*[3]

This conclusion was incorporated by reference into Section 10.38 of the *1979-1982 Agreement*. The 1978 committee tested several commercially available instruments and selected the *Endeavor* system developed at Northwestern University. In 1986 Endeavor ceased commercial operation and for three years the University used a locally written instrument developed by an *ad-hoc* faculty committee. In 1989 a third committee sponsored by the Academic Senate revisited the matter. We obtained permission to use the Endeavor and incorporated it into our own form which is still in use.

There is an extensive body of research[4] and a deep reservoir of opinion on student evaluation of faculty.[5] Indeed, one of the problems with the subject is that every faculty member regards himself as an expert due to having spent a career testing students. There are, however, many faculty who have supplemented their innate expertise with some research. Further, there are centers at several major universities devoted to the evaluation and improvement of teaching. One body of research purports to show that no student evaluation instrument is valid. In a

---

[3]This clause contains several *terms-of-art* taken from the literature of student evaluation theory and used in their technical meaning. A *formative* evaluation is one used to provide information for the improvement of instruction. A *summative* evaluation is one used to assess ability or merit for faculty status or salary purposes. Note that the primary difference between summative and formative is the use made of the results. A *proven* instrument is one that has been validated and shown to be reliable for its intended purpose through field testing. It will also often have been *normed* so that the results have a standard known interpretation. A question is *reliable* if the responses are reproducible, stable over time, and generally independent of extraneous influences. A question is *valid* if it can be shown to reliably measure or allow assessment of some quantity of interest. Each of these terms is discussed in detail in the Endeavor Guide which is available from the author and on-line.

[4]The citations in this position paper are all from the 1970's when Endeavor was developed and first selected. There is a clear need for a more up-to-date survey of the literature.

[5]A brief review and discussion of research on the subject appears in "A Two-Dimensional Analysis of Student Ratings of Instruction", P. W. Frey, Research In Higher Education **9,** 69-91 (1978).

"classic" paper Rodin & Rodin[6] concluded that "students rate most highly instructors from whom they learn least." There are many faculty who support this view and it must be rebutted if student evaluations are to be accepted by faculty. The study by Rodin & Rodin is flawed[7] and their conclusion is not supported by most reviews of the literature which generally show that a properly designed instrument that is carefully administered can be reliable and valid[8], a position that I personally support. However, *the same research has also shown that a badly designed instrument can be misleading and even invalid.* Had Rodin & Rodin's results been used for faculty evaluation, incompetent faculty would have been rewarded and competent faculty penalized. The anti-evaluation research results are valid in that badly designed forms are not to be trusted. It is critically important that we not commit the errors in evaluation design that are known to produce unreliable results.

Safe Summative Evaluations

The key to a student evaluation form that is reliable enough to be used for summative faculty review is to use a form developed and validated specifically for that purpose. This point is made repeatedly in the advice of professionals in the field. From the Association of American Colleges[9]

> Much published work has established the reliability and some types of validity of student evaluations of teaching. There is no doubt that if the best known procedures are used, student judgments can provide an excellent source of first-hand data. How much faith can be placed in these judgments will depend on the quality of the instrument and of the procedures employed to collect them. ...

---

[6] M. Rodin and B. Rodin, Science 177, 1164 (1972)

[7] P. K. Gessner, Science 180, 566 (1973); P. W. Frey, Science 182, 83 (1973); and correspondence in Science 187, 555 (1975).

[8] F. Costin, W. T. Greenough, and R. J. Menges, Review of Educational Research 41, 511 (1971); A. M. Sullivan and G. R. Skanes, J. of Educational Psychology 66, 584, (1974); H. W. Marsh, H. Fleiner, and C. S. Thomas, J. of Educational Psychology 67, 833 (1975); P. W. Frey, Science 187, 557 (1975); J. A. Centra, American Educational Research J. 14, 17 (1977).

[9] *Evaluation of College Teaching: Guidelines for Summative and Formative Procedures*, Grace French-Lazovik, Director, Center of Evaluation of Teaching, University of Pittsburgh, published as an occasional paper by the Association of American Colleges.

> When Student judgments are to be considered in summative evaluations, a wholly different set of procedures [from those used in formative evaluation] is dictated in order to insure the comparability, accuracy, and consistency of the results necessary to their use in the academic decision process. A standard questionnaire, one which has been carefully derived and subjected to considerable refinement is necessary to provide comparability among professors. ...

Validation is *not* done by careful rereading of potential questions by their author(s). To validate a questionnaire it must be administered in a controlled environment under conditions that permit an independent judgment to be made about the quality of the responses. Reliability and repeatability must be verified, not speculated about. Finally, correlation studies must be done to show that the quantity being measured is related to other quantities of interest and to quantify that relation. This process is not simple. Professor Frey spent over five years developing *Endeavor,* producing at least four journal articles in the process. In one study he considered 26,787 responses from 1,298 class sections at Northwestern University.[10]

Faculty Opinions and Concerns

Faculty at most universities can be divided into two camps regarding the value of student evaluation. One camp fears and distrusts all student evaluations. They fervently believe that the only valid student evaluations are essays written by upper division honors students. They also often believe that no numerical average can ever be a valid measure of a teacher. These faculty cite the considerable body of research showing that many if not most student evaluations measure only the popularity or charisma of the faculty member and are easily influenced by grading policy or academic rigor. A second group believes that all student evaluations provide useful information. They are actively interested in the opinions of their students about the academic experience and seek what ever information they can obtain for the improvement of their class room performance. These faculty tend to favor detailed evaluation forms that ask

[10] P. W. Frey, Research In Higher Education **9,** 69-91 (1978)

many, variegated questions. Bolstered by their own skill and experience in designing and administering tests, they believe in the results of polling[11] students for their opinions.

A careful reading of the literature shows that, as is often the case, *truth* or, less dramatically, good practice lies somewhere between these two positions. The normal faculty-designed student evaluation form can provide much excellent information for an individual teacher seeking to improve instruction or for a department interested in course improvement or for a Chair trying to help a new instructor develop classroom skills. Such forms can be tailored to the needs and particular interests of a department or individual. The results can be interpreted by those with the detailed local knowledge necessary to understand them and identify possible anomalies. This type of diagnostic or *formative* evaluation can and should play an important role in a department's ongoing process of self-evaluation. It can aid in faculty growth and curricular improvement. It can help *form* behavior and structure by giving feedback. However, this type of form can also be very misleading as it is known to be neither valid nor reliable if compared across many instructors or for different types of classes.

Dangers in Comparing Evaluations

When the grades of students are compared, the raw score from a physics exam should never be directly compared to the points assigned to an essay by a reader. Although both may be based on the same 100 point scale, they are determined in totally different ways. The appropriate measure to compare is the course grade which is awarded to each by a faculty member that knows the meaning of the raw scores and can assess them in context. If the results from different exams administered by different faculty to different students in different classes are to be comparable it is necessary that the exams be carefully standardized. It is not enough that the same exam be administered, if the scores are to be compared and used like grades then the exams must be validated and normed. If the judgment regarding the worth of the numerical score is made by the Professor who made up the exam and taught the course then the exam can be home-made. If the raw score is to be compared to the score of different students by

[11] Polling is not the same as testing: They are different processes. In testing it is often appropriate to give a low grade to a student who answers incorrectly because the question was complex and not understood or who answers a question not asked. In polling either of these circumstances could invalidate the questionnaire. Faculty who are skilled testers may be totally naive pollsters. However, their skill and experience at testing will tend to make them believe they are skilled pollsters.

individuals not intimately involved with the actual course then a standardized exam must be used.

The observations of the preceding paragraph apply to student evaluations as well as to exams. If the results of the evaluation are analyzed and interpreted by those intimately familiar with the local circumstances and are assessed as part of a general program of course assessment, instructional improvement, or faculty development and evaluation then the judgments reached about the instructor or course may be properly communicated to others much like a grade may properly become part of the student's record. If the actual numerical averages from a student evaluation are to be used outside of a department and compared to similar averages for other faculty then it is essential that the evaluation instrument have been validated and proven. This is particularly important when the results become part of a merit evaluation or a faculty status review. In these cases the faculty member is being judged. The primary objective is to assess merit not to generate change. This *summative* use requires far more careful procedures than does the *formative* use mentioned above.

The Current FDU Evaluation Instrument.

The evaluation instrument developed by the 1989 Ad-Hoc Committee and the additional procedures for other student evaluations the Committee has drafted were designed to effect a compromise between the two camps of faculty mentioned above and to address the issues and problems of *formative* versus *summative* (read *diagnostic & developmental* versus *valuative & judgmental*) use. The forms and procedures recommended provide both the type of detailed local information desired by many faculty and departments and at the same time provide valid University wide standardized data for faculty evaluation. The standard form (reproduced at the end of this paper) has four sections.

> The *Overall Evaluation* section (Q1 through Q7) is an exact reproduction of a student evaluation developed at Northwestern University by Professor Peter Frey as part of an extended research project. The form was proven by Frey through twelve versions in a six year period. The final version, consisting of seven questions, has be shown to be reliable and valid in a large number of different circumstances. The questions have been carefully designed to avoid asking the student to *judge* the faculty member. The items can be

answered from direct classroom experience. The factors that do and that do not influence the results have been identified and analyzed.[12] The two factors, *rapport* and *pedagogy* that are computed[13] from responses to the seven questions are as reliable as any in general use in higher education for the *summative* purpose of evaluating faculty for merit or status purposes. The results of Part I of the form are appropriate for direct inclusion in faculty files. Inter-faculty comparisons may be made based on the data as long as direct comparisons are limited to generally similar courses and the known factors influencing the results are kept clearly at the forefront. It should be noted that these seven questions and the two characteristics that can be deduced from them are not intended to and do not provide much in the way of good diagnostic data. The best they will do is to flag a problem area or identify an outstanding instructor. They will not answer the question *why?*.

The *Course Evaluation* section consists of five questions (Q8-Q12) that tend to be more course oriented or to involve student's judgment. They have not been validated. They will provide some useful information beyond that from the first seven questions. The results from these five questions are to be provided to the departments where they are to be used by and available to those with a valid need for the results. The actual numerical values of the averages of the responses to *Course Evaluation* questions should not be used directly for inter-faculty comparison. Rather, they should be analyzed and evaluated by those with the detailed local knowledge necessary to validate and confirm whatever they indicate. The results of that evaluation but not the raw data may then be used as part of a faculty evaluation process and may be used for inter-faculty comparison.

The section headed *Instructor's Questions* (Q13-Q15) provides a limited vehicle for individual faculty to ask three questions of their own devising. The questions asked need not be made public by the instructor. All that is required is that a hand-out be prepared with questions A, B, and C listed. Clearly questions appropriate to the chosen (and generally reliable) response fields are preferable. However the instructor may, at his or her own risk, change the meaning of the seven fields.

---

[12]See Frey's research and the Endeavor Handbook.

[13]The formulas for these items, determined by factor analysis, are: Rapport = -0.2(advanced planning) + 0.5(dass discussion) + 0.5(personal help) + 0.2(grade accuracy), Pedagogy = 0.1(hard work) + 0.4(advanced planning) - 0.2(dass discussion) + 0.3(presentation darity) + 0.4(increased knowledge).

The *Written Response* section provides an opportunity for students to provide short written responses to three general questions. The part of the evaluation form containing the written answers is to be separated from the answer sheet by the administering department. The answer sheets will be returned to the central office handling the data analysis. The written responses should be retyped by someone other than the individual being evaluated so that the identity of the respondents is preserved. Handwriting is easily identified. The typed responses should then be given to the faculty member. Since these comments can be quite idiosyncratic they should remain confidential.

The evaluation should be administered around the tenth week of a semester. This is late enough that all faculty will (or should) have returned some graded or evaluated work and late enough that the semester's patterns will have been established. It is far enough away from finals that no conflict will arise with either remaining class time or final exam anxiety. The evaluation should never be administered with the faculty member present in the room. However, use of a reliable student from the class briefed by the faculty member is proper.

The current University evaluation form provides very limited diagnostic information. It is also, by the necessity of working for almost all situations, not tailored to any particular local need. Some departments or faculty may well desire more detailed or relevant information specific to their situation. In this event, individual colleges, departments, or faculty are authorized to administer their own local evaluations as supplements to the standard form should they so desire. If a local instrument is developed and administered the data from it shall be disseminated and used exactly like the data from the *Course Evaluation* section of the form. Only the conclusions resulting from a careful analysis of the raw data may be used for inter-faculty comparison or faculty evaluation.

Notes, Comments & *Caveats*

Polling versus Testing

Even if you think you are measuring the students' opinion of the faculty member you may not be. One problem is that students often do not answer the question that was asked. A professor reported a class that gave him a medium rating on punctuality when he had never been late. They were clearly responding to a general sense of his performance not to the specific

question asked. Research has shown that the results of student evaluations are actually determined by a relatively few group of *factors* regardless of the number or complexity of the actual questions asked. There are many known systematic effects independent of the actual student's faculty rapport that can distort the responses.

We must distinguish between the conclusions we draw from a single student's exam grade and the average score from an entire class. The former, examined with knowledge of the remainder of the class and the nature of the exam, can well form the basis for awarding a grade. The latter may measure the classes overall ability but it may also measure the difficulty of the exam or the ability of the teacher. It is quite complex to separate these.

Dangers in Unproven Instruments

We must distinguish between inter-faculty comparisons and single intra-faculty/class information. If we are going to compare average responses for a faculty member to department, college or university wide norms (or even to results for colleagues) we must know what systematic influences are present that effect scores. This requires a carefully proven instrument. If we are just going to ask about a single class its position relative to other classes is not important. Different orchestras tune to different pitches. This is generally not evident and will cause few problems if you listen to one orchestra. It could be very important if they play together at the same time. If evaluations of different faculty teaching different classes are to be compared a carefully proven instrument is required.

If we are going to look at detailed answers to a question in the context of a single faculty member and a particular section of a course and try to understand the dynamics of the professor's interaction with the class then these safeguards and caveats are not necessary. This type of diagnostic or "formative" information is an appropriate part of any peer evaluation process. It can and should become part of the department's evaluation of individual faculty. It can be a valuable element of curriculum analysis. However, the numerical averages should not be archived in personnel files because they are generally unreliable. There are severe dangers inherent in averaging things. For instance consider using the average annual temperature in Chicago to predict how to dress on any one day.

Questions that elicit a large range of answers from a single class regarding a single instructor are generally unreliable. Complex questions requiring reading skills are also apt to be unreliable—we are seeking information from students not testing them. This relates to the Polling/Testing question. The examination metaphor for student evaluations has major dangers. When we test we are looking to see if students understand. Failure to do so is their fault. If a student doesn't understand an exam question that *may* be appropriate grounds for lowering his/her grade. In such circumstances you certainly cannot trust the response. In student evaluations, or other polls, if a question is not understood the result is bad, unreliable data. If responses are averaged this may be impossible to detect.

One way out of some of these problems is to ask questions that do not require the students to give an *opinion* regarding the instructor. Rather, questions are asked that the student can answer *factually* based on his/her actual experience in the class room. This avoids the entire issue of the student's qualifications to judge the faculty member. The student is asked to be a *witness* not a *judge* or *jury member*. In a court of law witnesses may not express opinions unless they have been qualified as "expert witnesses". They may only testify as to the "facts". A common objection in a court is "the question calls for a conclusion by the witness". Unless the witness has been qualified "expert", the objection will be sustained. The analogous objection regarding student evaluations, "students are not qualified to judge me", can be avoided by asking for facts rather than judgments.

Why does *Discusssion* have negative weight in *Pedagogy*?

A great deal of discussion and confusion have arisen because of the appearance of negative weights (coefficients) in the computation of the *Rapport* and *Pedagogy* factors in the reporting of the *Endeavor* results. *Advanced Planning* has a negative coefficient in *Rapport* while *Class Discussion* is weighted negatively in *Pedagogy*. It is important to note that these weights do *not* represent value judgments on the value of planning or discussion in the classroom. Indeed, one of the virtues of the *Endeavor* system (at least in my opinion) is that is has no preferred or embedded model for “good instruction.” Rather, the coefficients used are derived using a technique call “factor analysis” which is a mathematical technique for extracting information hidden in complex systems.

There is debate about the value of factor analysis and I am not going to extend the debate now. The subject is complex and understanding it well enough to evaluate the methodology requires a solid background in linear algebra and statistics. The concepts of the method are not that difficult. Essentially, you are presented with a "large" number of responses that are determined by a "small" number of independent factors. Your job is to determine from the responses what the "best" set of factors are. The problem can be defined mathematically. The difficulty is that it often does not have a unique answer. However, it may have a "best" answer. This "best" answer is the origin of the coefficients used in computing *Rapport* and *Pedagogy*. The reason I am not going to try to defend factor analysis here is that, first, it is a very technical subject and beyond the scope of this discussion and, second, the value of these two factors can be shown by the extensive validation studies done in the development of *Endeavor*. You can never prove the validity of the application of any mathematical model to the social sciences–they are just too complex and have too many uncontrolled variables. (That is the not-so-humble opinion of a physicist.)

The negative coefficients can be understood intuitively as follows. It is quite reasonable that of the items that influence *Rapport* it is found that *Advanced Planning* has the lowest weight. If you then ask that seven questions rated 1-7 are used to compute one answer that is also required to be in the 1-7 range *and you require that the relative weights be preserved* then it may will happen that the lowest weighted answer may end up with a negative coefficient. *Pedagogy* is defined to be independent of *Rapport* (so that all the problems associated with charismatic or popular teachers does not cross over into its value) and it turns out that this implies that *Discussion* has the lowest coefficient with the same result as for *Planning* discussed above.

Safe Use of Formative Evaluations

Consider several examination examples that show the dangers in misusing unproven instruments. Sloppy exams (e.g.: too easy, too hard) or individualistic exams (125 points total, a class mean of 30 by design) or unusual exam conditions (too hot, too little time) where an entire class is subjected to an unstandardized circumstance are generally not unfair to a single class as long as the instructor is aware of the circumstances and the situation affected everyone in the

class evenly. However, such exams can be very unfair if administered to multiple sections resulting in one section receiving grades that are systematically 10 or 20 points different from another's. The effect of a hard or easy exam or the result of an extra half hour for completion is unimportant until the two classes are averaged together or compared. Then the difference can result in a major inequity. This is why ETS spends so much time "proving" it questions and standardizing its exams.

The conclusion to be drawn from these examples is that sloppy student evaluations given for a single instructor and "graded" (locally analyzed) for that one individual are OK if evaluated by a knowledgeable person. (However, note that an instructor insensitive to the fact that a particular exam was very easy or very hard will award skewed grades.) Such evaluations must not be merged into a general reporting system in their raw, un-curved form. They must be analyzed first. The raw exam scores should not be used for inter class comparison. They must be curved and analyzed first. A letter grade (or a normed numerical grade on a standard scale) must be assigned. Only then can the results be compared between classes or between instructors.

The essential idea is that the results must be curved or analyzed at the local level *before* inter-faculty comparisons are made. After a full peer/chair review of the instructor's classroom performance has been completed using, among many items, the results of a diagnostic student evaluation, then the conclusions reached may be compared for different faculty and summative evaluations made.

Important relationships to keep in mind when interpreting instructional ratings.

- Averages based on small samples are notoriously unreliable; when instructional ratings are based on 10 or fewer students they should be viewed with considerable caution.
- Evaluation decisions based on ratings from three or more classes are more reliable than those based on ratings from only one or two classes.
- Students who major in different departments have different backgrounds and different expectations and often use nonuniform standards in making their ratings; therefore, the instructional ratings for individuals should be compared only when similar courses are involved.

- People tend to become more tolerant of others as they grow older; research demonstrates that freshmen rate instructors more harshly than do upperclassmen.
- Ratings on items which ask about the personal relationship between instructors and their students are heavily influenced by class size; the larger the class, the lower the ratings.
- Instructors who grade stringently (i.e., many Bs and Cs) tend to receive lower ratings on items which ask about class discussion, student-instructor interaction, and satisfaction with grading.
- Experienced instructors tend to be rated higher than young instructors on presentation clarity and organizational skill but lower on student-instructor interaction.

Conclusion

Valid data on the ability of an instructor can be gotten from a well designed summative questionnaire. Important and useful information for the improvement of instruction and curricula can be gotten from a well designed formative questionnaire. However, a badly designed questionnaire will reward incompetence and mislead those who use it. These observations are based on a review of research on the validity of student evaluations; they are not personal opinion. Citations are given above.

David Flory
Professor of Physics

Original, August 20, 1996
Revised, April 7, 2002

E-mail: flory@fdu.edu
Mail stop: H-DH4-03
Phone: 692-7064.