When asked to examine the reliability and validity of measures of faculty achievement, one must ask for what purposes the measures are to be used. The reliability and validity of a measure exists only with respect of some other variable that is implied by a definition of faculty achievement. If faculty achievement is defined in terms of teaching effectiveness, research, and public service, a measure of faculty achievement would be considered reliable and valid if, and only if, it is significantly correlated with that definition. More specifically, if research is defined in terms of numbers of articles, someone else may quarrel with the appropriateness of that definition, but a count of the number of articles is reliable and valid with respect to that definition. Similarly, if teaching effectiveness is defined in terms of student academic gain, a measure such as a survey of student opinion or other information about individual faculty members is reliable and valid only if that measure is correlated with student academic gain.

Because we do not have a clear and unequivocal definition of teaching effectiveness, it is controversy that any measure is reliable and/or valid. Even with a definition in terms of student academic gain, the question of whether the survey of student opinion (or other often used "measures") is reliable and valid -- as of yet -- is unanswered. Based on a review of the relevant literature, it is the position of the Task Force that the degree of correlation between current measures of teaching effectiveness and student academic gain has not yet been adequately documented.

If, instead, teaching effectiveness is defined in terms of customer satisfaction, and IF we accept the assumption that students both reliably and validly reflect their satisfaction, then a survey of student opinion is appropriate by definition. Even under this second definition, however, other questions still remain. If faculty behaviors can be changed in accordance with student opinion about such behaviors, will the predominant faculty behaviors that result be consistent with our purpose associated with the mission of this University? IF a survey of student opinion and student academic gain are significantly correlated, then the survey is reliable and valid and, further, changes in faculty behavior would be expected to be consistent with our mission.

The Task Force recommends that, IF measures of teaching effectiveness can be constructed that are correlated with academic gain, then and only then should they be used for administrative decisions such as merit pay, promotions, or tenure. Since teaching effectiveness is of such critical importance to the mission of this University, the Task Force recommends that appropriate resources be allocated to develop a measure of teaching effectiveness that correlates with student academic gain.
If those who make decisions such as merit pay, promotions, and tenure chose to continue the use of student opinion and other measures, then they should be aware of the problems associated with them and take corrective actions when possible. The remainder of this report focuses on these problems.

In the absence of this first choice, the Task Force suggests that questions that concern the relationship between student opinion and variables such as the size of the class, the time of the day, the amount of equipment used, the level of the class, and other environmental variables be examined at this University. Such an effort will have costs, but questions about these and other relationships are considered to be important.
INTRODUCTION:
After much deliberation, the Task Force for the Evaluation of Teaching has developed responses to these queries:
1. WHY should we try to evaluate teaching performance?
2. WHO should actually appraise teaching?
3. WHAT are appropriate criteria for evaluating teaching?
4. HOW should information regarding teaching performance be obtained?

PART 1: WHY SHOULD WE EVALUATE TEACHING PERFORMANCE?
Obviously, since our primary objective is to educate, we should endeavor to maintain at least satisfactory levels of teaching effectiveness. To be more precise, the evaluations should be done for two major reasons:
(A) Developmental Purposes--to assist faculty in their development as teachers; and (B) Administrative Purposes--to aid in making personnel decisions such as merit pay, promotions, and tenure.

RECOMMENDATIONS:
1. We recommend that separate "items" be developed for each function. In other words, one measured for faculty development and one measure for administrative purposes.

2. Each faculty member shall maintain control of the use of the data gathered for developmental feedback. The department chair/head shall not automatically receive copies of the information without the faculty member's consent. However, faculty members shall have access to all information used for administrative purposes.

3. The measure used should be validated against appropriate criteria. As previously noted, the best criterion is a measure of academic gain.

COMMENTS:
Not all faculty will have the same developmental needs. To evaluate faculty on a composite of items related to the various areas of teaching would be like comparing apples to oranges. Further, if faculty or administrators knew that the "ratings" could affect personnel decisions, they would not explore potential problem areas thoroughly.

A second warning concerns using any method of evaluation for making personnel decisions. One must realize that no evaluation procedure is 100% valid or reliable. Therefore, unless departments can develop ways to measure teaching effectiveness which are perceived to be fair, then one should not be overly confident in the ratings or rankings. At best, we can make gross distinctions of what is either satisfactory or unsatisfactory performance. Attempting to make fine discriminations of what is "meritorious" will probably result in reduced morale. Remember, an extreme few will view themselves as "average."
The latter caveat touches on a problem ubiquitous to all aspects of our performance evaluations. We have major problems defining and measuring meritorious performance, yet we continue to allocate resources as if we can. While this paradox is beyond the scope of the Task Force, we would be remiss to neglect it. We are indirectly recommending that we only distinguish satisfactory teachers from unsatisfactory ones. However, such a recommendation may serve to perpetuate the under-reward inequities perceived by those who teach well. In other words, rewards would be based primarily on scholarly activity and service while ignoring our main goal of teaching. Thus, the entire merit system needs to be modified.

PART 2: WHO SHOULD ACTUALLY APPRAISE TEACHING?

The key criterion for selecting "raters" involves answering the question: "Who has the best opportunity to observe or acquire knowledge about the appraisee's performance?"

RECOMMENDATION:

The department chair/head is in the best position to gather and synthesize information from a variety of sources. He/she should use other valuable resources such as student evaluations. Even self-evaluations by faculty could provide information regarding teaching performance for developmental purposes only.

Whoever is involved making evaluations (i.e., chairs/head, students, peers, self), should be trained regarding potential appraisal errors and limitations. Possession of this knowledge should not be assumed simply because one is a "professional" or in an administrative position!

COMMENTS:

(A) Student evaluations may be considered in the appraisal but must not be the sole determinant. Students have the greatest exposure to a faculty member's performance and have an opportunity to evaluate communication skills and rapport. These may be necessary, but not sufficient, requisites for effective teaching. Other sources regarding subject matter competence are needed.

Using student input can also serve a public relation function provided they (the students) receive positive feedback that their opinions are valued.

(B) Peers usually have limited information regarding a colleague's performance. Thus their input to a dept. chair/head should be accompanied with the warning that it is based on incomplete data.
PART 3: WHAT ARE APPROPRIATE CRITERIA FOR EVALUATING TEACHING? and
PART 4: HOW SHOULD THESE CRITERIA BE MEASURED?

The goal of teaching is to increase the knowledge and/or skills of the students. An obvious criterion would thus be the "value-added" concept whereby the amount students knew after a course minus the amount prior to enrollment is computed. (A review of the relevant literature indicates that, to date, student evaluations have not been validated against this criterion.) Since such "result" type criteria is problematic, we have been using "behavioral" criteria -- measuring those behaviors which are considered instrumental to effective teaching. These include: knowledge of subject matter, providing explanations, speaking clearly, helping students, challenging students, encouraging analytical thinking, and motivating students to learn.

This should be considered a temporary procedure until a validation study has been conducted to see what measures correlate with learning criteria. (As a footnote, without substantiating data regarding their validity, the University could be placed in a precarious legal position if student ratings only are used for appraising teaching effectiveness.)

RECOMMENDATIONS:

It is recommended that each faculty member reach an agreement with his/her dept. chair/head regarding what measures of teaching performance are appropriate for individual needs. This mutual goal setting and measurement plan could make provisions for possible options such as:

A. Student Opinion
B. Samples of Course Content
C. Class Visitations
D. Direct Measures of Academic Gain

(A) Student Opinion -- Such surveys usually have at least face validity as a measure of consumer satisfaction. The one area where they may be weak is their ability to evaluate subject matter competence.

For developmental purposes -- The Purdue Instrument or clones are quite adequate. It should be modified to encourage more student comments, but it can still help faculty identify problem areas.

In either case, instructors should not administer or receive student evaluations until all results have been summarized and prepared for review.
In terms of feedback of results each faculty member should receive the following information:

1. Mean scores
2. Median scores
3. Standard errors of measurement
4. Departmental norms
5. College norms
6. University norms

For administrative purposes—The literature suggests that student input may be sufficiently summarized by one global item such as:

How would you rate this teacher's overall effectiveness?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>VERY POOR</td>
<td>FAIR</td>
<td>MODERATE</td>
<td>GOOD</td>
<td>VERY GOOD</td>
<td>EXCELLENT</td>
<td></td>
</tr>
</tbody>
</table>

Common departmental or college core items are probably unnecessary.

In terms of feedback of results each faculty member should receive the following information:

1. Mean scores
2. Median scores
3. Standard errors of measurement
4. Departmental norms
5. College norms
6. University norms

COMMENT: Again, it should be emphasized that student evaluations should not be the sole criterion for appraising teaching effectiveness. Student ratings may be influenced by the level of the course, course requirements, whether or not the course is an elective or a requirement, and other extraneous factors. Therefore, other information regarding the content of a course is needed.

Furthermore, since we do not have appropriate validity and reliability information on these measures, extreme caution should be used when making discriminations among faculty. Only scores falling in the extremes may be useful. Due to the error of measurement associated with the item used for administrative purposes, faculty scores falling within ± standard error of measurement may be indistinguishable. Any discriminations of individuals on this measure are therefore subjective.
(B) **Samples of Course Content** -- This could include items like syllabi, textbook choices, tests, assignments, etc. One could look for evidence concerning such factors as:

- Do the tests demonstrate acceptable rigor?
- Are textbooks and other readings current and appropriate to the course?
- Are written assignments appropriate for the course?
- Are innovative methods of instruction used?
- Are classes met on a regular basis?
- Are office hours maintained?
- Have course objectives been clearly communicated?

This list of questions is not intended to be all inclusive. Prior to any evaluations being made, specific definitions and examples of what will constitute satisfactory performance need to be developed by each department.

(C) **Class Visitations** -- Class visitation is a potentially explosive activity and one that requires extreme care. The Task Force does not recommend that all departments adopt the procedure. It is one possible option that is contingent on a supportive, collegial departmental environment. They may be done on an unannounced basis provided that the visits are requested by the faculty member. However, several such visits are necessary and may be impractical in some departments.

Again, prior to any visits and subsequent evaluations, the department needs to reach a consensus regarding specific behaviors they perceive to be correlated with effective teaching. **Everyone** must clearly understand and agree to these standards.

(D) **Academic Gain** -- Direct measures of academic gain are available in many courses in the form of differences in pre- and post-course tests of knowledge. These differences are "results" measures of teaching effectiveness. They are relatively expensive, but more consistent with the definition of teaching effectiveness in terms of student academic gain.
PART 5: WHEN AND HOW OFTEN SHOULD TEACHING BE EVALUATED?

RECOMMENDATION:
Each faculty member should have his/her teaching performance evaluated for each course he/she teaches at least annually.

COMMENTS:
The goal is to evaluate performance in each course an instructor teaches at least once per period of record (therefore annually). Currently, courses taught only in the Spring are not subject to evaluations. This is unfair both to the faculty member and to the students. An instructor's "best" course may be in the Spring semester yet it presently does not "count" in the evaluation.

Faculty may, for developmental purposes, choose to have all their courses evaluated. Informal evaluations and solicitations of student comments can be made as often as deemed necessary by the faculty seeking to improve. These latter evaluations should not be used against a faculty member when decisions regarding promotions, tenure, or raises are made.