DOCUMENT RESUME

ED 054 724 HE 002 517

TITLE Student Evaluation of Teaching. Presentations at a Conference.

INSTITUTION Pittsburgh Univ., Pa. Inst. for Higher Education.

PUB DATE Dec 70

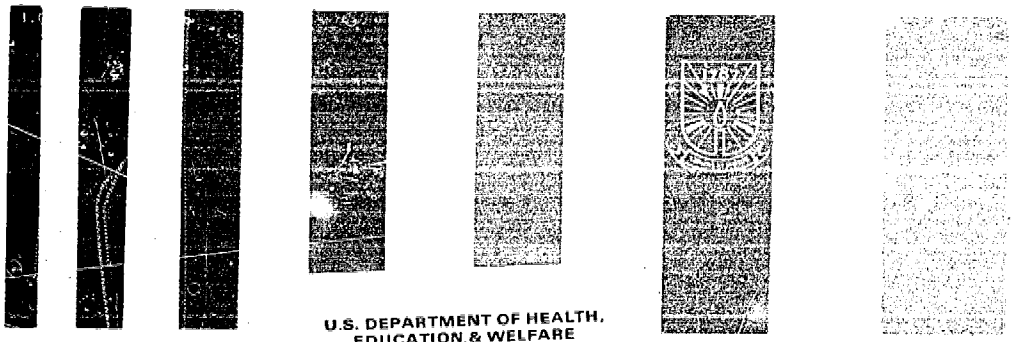
NOTE 23p.

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS College Faculty; College Students; *Effective Teaching; Evaluation; *Faculty Evaluation; *Higher Education; Learning; Rating Scales; *Teacher Evaluation; *Teaching Quality

ABSTRACT

This pamphlet presents two articles on student evaluation of instruction. The first article, "Research on Student Ratings of Teaching" by W. J. McKeachie, deals primarily with the reactions to his article in a recent AAUP Bulletin where he argued that systematic methods of collecting student evaluation of instruction should be used by faculty committees in evaluating teaching effectiveness. McKeachie thinks that the ultimate purpose of evaluating teaching is to improve learning, and that evaluation of either students or teachers should be based solely on learning. The second article, "Student Rating of Teaching, Some Questionable Assumptions" by George L. Fahey, examines some of the assumptions that are made in evaluating teaching effectiveness by means of student ratings and what these assumptions imply. (AF)



STUDENT EVALUATION OF TEACHING

Presentations at a Conference

INSTITUTE FOR HIGHER EDUCATION
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA

DECEMBER 1970

STUDENT EVALUATION OF TEACHING

Presentations at a Conference

Institute for Higher Education
University of Pittsburgh
Pittsburgh, Pennsylvania

December 1970

FOREWORD

One of the most persistent issues which has confronted higher education in the past decade has been that of evaluation. This has evidenced itself in many ways, not the least of which is that concerning student evaluation of instruction. Dr. Wilbert J. McKeachie, Chairman of the Department of Psychology at the University of Michigan, and Dr. George L. Fahey, Chairman of Educational Psychology at the University of Pittsburgh, have done much to clarify the issues concerning student evaluation of instruction in the two papers reproduced here.

The questions they pose, directly and indirectly, present a further problem to be addressed, i.e., after evaluation, what then?

Alex J. Ducanis
Director
Institute for Higher Education
University of Pittsburgh

RESEARCH ON STUDENT RATINGS OF TEACHING[1]

W. J. McKeachie

In preparing this paper I was in a quandary. To really carry out the implications of my title I should review much that was included in my AAUP Bulletin article; yet to simply repeat that hardly seems fair to those who have read it.

My resolution is a compromise. I'm going to use reactions to that article as a starting point and the research relevant to those comments and to Professor Fahey's "Questionable Assumptions."

Whatever else the article accomplished, it stirred up vigorous faculty reaction, both pro and con. Professor Borgatta, an outstanding sociologist at Wisconsin, wrote to the AAUP Bulletin as follows:

> ...What this raises is the question of the goals of education in the colleges and universities. It may be that some places will emphasize the form of presentation and the students' response to it as measured in a questionnaire such as proposed by McKeachie. On the other hand, some professors may feel that colleges and universities should be thought of as institutions of higher learning. One might very well say that the emphasis should be that students are there to learn, not to be taught. What does it mean to be taught? And, if the institutions are for learning, then it is of concern to emphasize having learned people on the faculty, and the criteria that are most relevant have to do with a person's scholarship and what he adds to the basis of knowledge at the institution. When this kind of criterion is raised, the relevance of student ratings pales.

Professor Borgatta's concern with learning is one we all share, I am sure. But is it enough to have learned people present on the faculty? Professor Borgatta says that students are "not to be taught." Can we assume that all students will learn well if they are in a college where there are scholars?

---

[1] Prepared for presentation at the University of Pittsburgh, December, 1970.

Borgatta and I are probably not as far apart as I imply. For I quite agree that colleges and universities are ideally conceived as communities of learners and that faculty scholarship is a component of the climate of learning. But I cannot believe that he would assert that there are no differences between faculty members of great scholarship in the degree to which they facilitate learning.

Certainly colleges and universities should be good places for faculty. But this is not our primary purpose. Surely an important part of a teacher's value is his contribution to student learning.

We have a great deal of evidence that teachers do differ significantly in their effect upon student learning. The increment produced by good teaching is measurable. Unfortunately it is not often measured. One of the reasons for this is that we have sometimes confused evaluating student achievement for purposes of assigning grades with evaluation of learning.

In assigning grades we have traditionally been concerned with the question, "What can the student do?" We give objective, essay, or problem tests measuring the student's memory of facts or concepts and his ability to use them; we construct our tests to discriminate the abler from the less able students. But we do not attempt to measure learning directly. If a student comes into the course with an unusually good knowledge of the field, he may receive a higher grade than the student who began with nothing and learned a great deal. And this is as it should be. A grade ordinarily is intended to represent a standard of achievement without regard to how the standard was reached and how much was learned in the course.

But our educational objectives go well beyond end-of-the-course achievement. We are not satisfied if what a student learns is forgotten as soon as the final examination is over. Rather we are concerned with what we have done that has influenced his <u>continued learning and performance throughout the rest of his life</u>. The ideal assessment of teaching effectiveness would measure the student's interests, skills, and knowledge before the course and at several points in time after the course. Such an assessment is not impossible, and I predict that some enterprising educational researcher will do it erelong.

But what about the average college teacher today? I think we can do better than most of us do. If we think of our exams not so much in terms of a sampling of what has been covered in the course and more in terms of assessing what foundation has been laid for continued

learning, our exams may change. For example, some of the facts and concepts that are interesting may change. For example, some of the facts and concepts that are interesting or conventionally taught may not be those most important for further learning; skills such as being able to read in the field of the course with some discrimination and thoughtfulness are important; ability to observe and evaluate one's observations may be important. Clearly we can't measure whether the student will continue to use these skills in every day life, but we can get some ideas of whether he's able to use them when the situation calls for it. With effort, objectively scored tests can be developed to do more with these objectives. At Michigan we've spent twenty years developing a test of psychological thinking. We have one that is not fully satisfactory but better than a typical achievement test. On our test students are given examples to analyze; they are asked to evaluate different hypotheses. They are required to apply concepts. The test is multiple choice.

But more often one needs to go beyond conventional tests. Thus to get at reading skill in psychology, I've tried bringing in journals and books to the exam room and asked students to select an article or a section of a book to read and to evaluate these readings as they have during the course in their ungraded reading logs. Since one of the skills I'm interested in developing is the ability to learn from others, I have tried having one question on which students were asked to pair off and to help one another prepare the best possible answer. Since I'm interested in developing observational skills, I have sometimes asked students to leave the examination room for fifteen minutes and to go out on campus to observe any behavior that interested them. They are then to return to the examination room and write an analysis of the behavior with hypotheses about relationships and methods for testing those hypotheses.

I'm not sure that these types of examinations provide more valid measures of learning than those conventionally used. I do think the effort to devise methods of evaluation that will stimulate student learning and go beyond straight memory is an important one.

If evaluating student learning is so difficult, it is clear that evaluating the effect of teaching on student learning must be even more difficult. As I pointed out in my article in the AAUP Bulletin, comparing achievement in mathematics with achievement in psychology is like comparing apples and oranges. We can't say precisely how many units of learning in mathematics are equivalent to how many units of learning in psychology. However, I apparently didn't get across to Professor

Borgatta and other critics that human beings are, nevertheless, able to evaluate such apparent incommensurables. We pay so much for an apple and so much for oranges. A baseball team will trade an outstanding batter for two pitchers--obviously feeling that they can make judgments comparing the value of these. If I am deciding whether to hire one professor rather than another, I can make a judgment. Professor Borgatta is probably able to judge whether one researcher is better than another even though they are working on different problems. In our own grading of students we may worry and vacillate, but we still are able to assign grades to students who differ in terms of how well they do on objective tests, how well they have written their term papers, or how well they have participated in class discussion. So we <u>are</u> able to make judgments about relative excellence even though the excellence may be achieved along different dimensions. The real question is "Are student judgments about teaching effectiveness related as they should be to other possible indices of effectiveness?"

Our best opportunity to answer this question comes in courses where at least some of the fundamental objectives are common to different teachers and some measures of learning by students taught by different teachers can be obtained. As I indicated in my <u>AAUP Bulletin</u> article the evidence here is positive but not overwhelming. In five conventional evaluation studies in different courses investigators have found positive correlations between student ratings of teacher effectiveness and average achievement of randomly selected classes.

What other sorts of evidence are there? One that appeals to me is in studies done by Elliott and by Russell showing that those students within a course whose achievement was higher than would be expected on the basis of their intelligence and previous background tended to rate the teacher as being more effective than the students whose achievement was less than expected. In both cases it was found that instructors who were particularly effective for the brightest students, for example, were rated as being more effective by bright students than by the students with less ability, while teachers whose effectiveness was greatest with the poorer students tended to be rated higher by these students than by the brighter students. This seems to me evidence that student ratings do reflect the teacher's effectiveness with them.

Another source of evidence would be the effect of the teacher on desire to continue learning. While I do not see recruiting students for majors in one's field or for advanced courses as being a primary objective of teaching, it does seem that teachers who are effective in stimulating students' desire for continued learning

would be those whose students tend to go on to elect more courses in the same field. Solomon and I showed that this seemed generally to be true of teachers of general psychology at the University of Michigan.

The question of validity, however, has to be countered for "Valid for what?"

Are student judgments better than faculty judgments? This is the question Professor Borgatta raises, and, I think, an appropriate one. But I think Professor Borgatta missed the point of my article. I was arguing that systematic methods of collecting student evaluation of instruction should be used by faculty committees in evaluating teaching effectiveness. I was not arguing that the student ratings should themselves determine administrative decisions about a teacher's salary or promotion. The question is not whether student ratings are sufficient in themselves but whether they can improve the validity of faculty and administrative judgments. I believe that most faculty judgments are already based upon student judgments of teaching. Faculty members tend to form their opinions of a colleague's teaching largely from conversation with students in their own courses or with those whom they are advising. The comments that come to their attention are often critical ones raised as complaints or rationalizations by a student who has not been successful. Very few institutions have systematic faculty observation of one another's courses or any other method by which faculty members can gain first-hand knowledge of a colleague's teaching. Even in the event that systematic visitation is used, it is rare that visitation is carried out more than once or twice, and one wonders how adequate a sample these sessions may be of the instructor's effectiveness over an entire term.

Certainly we should do better than we have in the past in getting additional information. One of the assets of television teaching has been that it has made teaching more public. The use of video tape enables teachers to get a picture of their own teaching and to share it with others. We should make use of these tools, but meanwhile we should do what we can to improve our present procedures.

The real difficulty in evaluating teaching is more subtle. When all is said and done, faculty judgments, student judgments and outcome measures of student achievement will probably not show a high degree of agreement at any global level. In fact we will probably find that different faculty members will evaluate teaching differently even when they have a good deal of information about teaching. The reason for this is that we weight different qualities and objectives of teaching

differently. For me the most important objective in teaching is to develop in students a motivation and capacity for continued learning. For others the most important characteristic of good teaching is transmission of a full and accurate picture of a particular body of subject matter. For still other teachers the most important objective is that of developing skills in solving problems in a particular field. All of us may agree that each of these objectives is important, but when we evaluate a teacher in terms of his achievement of all of the objectives of the course, the fact that we weight the objectives differently means that our ratings of overall teaching effectiveness are going to differ. One of the reasons that student ratings do not correlate more highly with outcome measures is that students' objectives in learning may be different from those measured by the typical outcome measures we have used in evaluating learning.

Now this does not discourage me. I believe that value judgments are fundamentals of human life and that a certain amount of disagreement is healthy. What I think we can do is to reduce the mystery about how we arrived at our value judgments and to strip out of the context of value judgments the empirical questions. I suspect that evaluation of teaching in the future will be more explicit with respect to the teacher's effectiveness in achieving different kinds of objectives. I believe that if we specify these objectives we will find that students do discriminate between different kinds of objectives. They may well rate one teacher as effective in stimulating interest in continued learning but not effective in communicating knowledge while another teacher may be effective in other respects. When we make these discriminations in our research we may find that the student, faculty and objective measures of learning are more highly correlated than has been revealed by past research.

In addition to getting evaluations of learning in the area of different goals, I think we need to look more and more at what goes on between the beginning and end of the course; in short we need to direct more attention to the process of teaching. Our research group at Michigan, led by Dick Mann, has been observing, recording, and categorizing every act in classrooms during a semester. The procedure is tremendously expensive and time consuming, but it is helping us understand teaching and learning from day to day as we would never have known by our traditional end-of-course measures. The University of Minnesota has been experimenting with periodic feedback from students during a course -- an approach that looks very promising.

But in addition to evaluating teaching effectiveness in terms of several goals and in terms of processes there is another sort of evaluation that is likely to be overlooked. One can also evaluate a teacher in terms of his choice of objectives and of techniques; i.e. we should not be satisfied simply with evaluation in terms of the professor's or students' goals. There is a higher level of evaluation in terms of evaluating teaching effectiveness in terms of the role of the course in the curriculum and even higher, in evaluating the curriculum itself. One of the reasons faculty resist student evaluation of teaching is the fear of collusion between student and instructor -- not the gross sort of collusion in which the instructor agrees to give high grades if the students give him high ratings but the subtle collusion in which the instructor sets minimal goals of memorizing basic knowledge and students accept this goal because it requires only skills which they understand. Neither student nor teacher is forced to think or to grapple with the material with real involvement. This is why evaluation is needed at several levels.

A final admonition. I believe student evaluation of teaching can be valid and useful; but let us remember that the ultimate purpose of evaluating teaching is to improve learning. Evaluation is not an end in itself. If a program of evaluation creates anxiety that interferes with good teaching, if it stimulates or reinforces hostility, if it simply takes so much time from learning that the net gain is negative, let's forget it. We must weigh the cost of evaluation against the gains. I believe there can be important gains but I would not overlook cost. The college is a <u>learning</u> community. Evaluation of either students or teachers should be forced to justify its existence in terms of <u>learning</u>.

STUDENT RATING OF TEACHING

SOME QUESTIONABLE ASSUMPTIONS

George L. Fahey

Two forces, at least two, seem to have merged to place an emphasis on the importance of evaluating the effectiveness of teaching. One of these is the expression of concern among students for an educational experience which is relevant and meaningful. No matter how sardonic we are with these terms, there is a primary concern that their effort and their tuition will result in self-development commensurate with the investment. Many students have said that they are exposed to poor teaching and some have proposed or initiated ways of identifying teachers who do not come up to the mark.

The other pressure to evaluate effectiveness may come from the same primary concerns but the mediating factors are different. In a diffuse sense there is the American focus on measuring everything in education by testing. This preoccupation among psychometricians started years ago -- Remmers at Purdue, Guthrie at Washington, Grinnell at Miami, and others. They developed teacher rating scales and fussed over reliability and validity but they never really started a widespread movement. Much more vigorous is the current concern over costs and with this a dalliance which may take on major importance -- the productivity index. The computers are busy on many campuses printing out ratios involving almost anything that can be counted -- clock hours, credit hours, student populations, faculty populations, committee hours, research hours, all in relation to dollars. The time-and-motion-study man is in the hallway. Even the most naive indexers know that these ratios will not tell the whole story. They glibly assert that what we must have first are baseline data and then we can cope with quality factors fed into the programs. How do we get measures of quality? The answer is obvious. We get the students to rate their teachers.

Since students are asking for quality and sometimes for ratings, and since administrators are nervously reaching for dramatic demonstrations before legislators of diligent efforts at quality-control, it seems likely that we will have an upsurge of rating procedures. It also seems safe to predict that many such procedures will depend on quickly obtained ratings by students.

I did not come today to try to stem this tide. I am not against the examination of quality. I think poor teaching is as bad or even worse than poor surgery. Ineptitude has no justification. I am not opposed to student ratings of teachers. I have repeatedly used such scales in my own classes and found the results helpful. The costs of higher education seem to be at or near their peak and we will have to find ways to get more for our money.

What I would like to do is point out some of the assumptions that we make in evaluating teaching effectiveness by means of student ratings and to examine what some of these assumptions imply. This is not with intent to deprecate the scales. It is with intent to keep us asking constantly, "What is it that we are doing?"-- to keep us from inferences unwarranted by the data.

It seems necessary to begin with a quick look at the way in which scales for students to rate teachers are put together. Neither the Bill of Rights nor the Book of Revelations specifies an exact procedure but, in general, it goes like this. We could start one today. Each of us could write ten cards, each of which contains an attribute of teaching. Each attribute must be one which in any given teacher could be observed as part of his teaching. On each card we would also enter our judgment of whether this trait is related to goodness or poorness in teaching. We will put all our cards into one list, discarding the duplicates. We distribute this list to a population of students to apply with reference to a teacher. (I am skipping over some purifying steps we might take along the way.) We combine the responses from individual students and find means and variances of the items.

This does not tell us whether the rated teacher is good or poor. We have to have a criterion. This could be our original judgment. When we wrote the items, we recorded some preconceptions of how good and poor teachers would appear. The most common approach to a comparison standard is not such speculation but correspondence with another assessment. This could be another scale. It could be ratings by observers. Most often the criterion is a global question attached to the scale. The students express their judgments on the separate items and also give an overall evaluation.

A scale which is "psychometrically good" shows each separate item to have a strong correlation with this global item but each separate item correlates at a low level with each other item. Items which fail these tests are discarded and we have left a dozen or so which make up our scale.

This scale can be refined until it has good reliability and ease of administration. It can yield one or more scores. These scores can be fed right into my productivity index and this makes me nervous. It is not that I am so tenderminded. It is that a variety of assumptions underlie this whole process. These assumptions, by the way, apply whether the productivity index demon is present or not.

Most scales up to the present have been developed and their use recommended as self-help devices. If I learn how my students see me, I can capitalize on my strengths and correct or avoid my weaknesses. Sometimes scales have been used by administrators seeking to coach teachers on how to teach better. Most scale developers have been opposed to any widespread use of their scales to determine promotions, tenure, or academic merit.

Now, let us look at some of the assumptions.

Assumption: That student raters of teaching effectiveness have had an opportunity to observe that which they are rating.

This assumption seems very safe, at least while we are thinking of conventional classrooms. Most students are there most of the time.

The only other person who sees more of a teacher's teaching is the teacher himself. He is always there when he is teaching. Oddly, there has been a relative dearth of research on self-rating. We have been preoccupied with population-sample ratings. Our statistics classes taught us how to deal with these but not with introspective data.

It seems very clear that "opportunity to observe," essential to rating procedures, is far more extensive with students than with supervisors or peers.

Assumption: That student ratings of teaching effectiveness are reasonably reliable.

It seems very safe to make this assumption when we deal with scales which have been carefully prepared. Naturally, the whimsical ones we throw together may not stand up, and shouldn't.

There is a great assemblage in the literature of reliability data on teacher rating scales. One of my students involved several thousand students and over one hundred instructors in successive terms -- same scale, same teachers, same subject, different students. The more than 100 repeated observation coefficients showed high reliability. There were only 3 below .9 and those were over .8. This is an almost frightening demonstration of

stability. Smedley, Zagorski, and Schmultz were seen by their students in April almost exactly as they had been seen by other students in November. I would like to return to this point later.

Assumption: That student ratings of teaching effectiveness can be obtained with relative ease.

This assumption also seems safe. Most rating scales take 10-15 minutes of student time, can be answered on a single sheet, machine scored and, with the right equipment, can be put through the computer with all sorts of statistical guides to inference.

This does not say that good scales are prepared with ease. It does not describe the level of ease with which the rated teacher can deal with himself on a computer print-out. But, these are other problems.

Assumption: That student ratings of teaching effectiveness are obtained with reasonable motivation for the rating task.

I had not thought much about this point until one of my students raised it. It sets off an uncertainty.

We know that some students have been demanding opportunity to rate their teachers. We also know that students will usually accept rating procedures and fill out the forms without dissent. We do not know what underlies the demands to rate. It is quite clear that there are some students who ask for anything which is disruptive. They may wish to reward or punish individuals or the system. They may be caught up in the efficiency-economy syndrome of our society. The ones who simply accept the task are doing what 13-plus years have conditioned them to do.

In neither case do we have proof that any large number of students really feel a need to rate their teachers. In fact, an occasional few will refuse.

Beyond this question of willingness is the more crucial question of what results are obtained under conditions of unknown motivation. We know that indifference causes unreliability, but we also know that student ratings tend to have high reliability. It is perhaps safe to assume that they are not completely whimsical. It is also quite clear that the majority are inclined to be quite generous. This can yield reliable results but reduce validity.

American society has been shocked in recent years to learn that a large proportion of American boys have no strong urge to go a-soldiering just because the politicians

have gotten involved in a war. Perhaps we have some other myths about what students want to do.

It is clear that a consensus in any direction can put a halo on one teacher and horns on another. We can infer but we cannot prove that these differential results are differences in teaching effectiveness.

Assumption: That the usual global question of overall effectiveness is more valid than single items and that scale validity is enhanced by holding or discarding items as they correlate with it.

In the absence of a criterion derived from external measures and possessed of the characteristics of reliability, comparability, and validity, we commonly utilize the global rating. It has certain distinct advantages -- being produced at the same time, by the same judges, presumably working from the same instructions, and same response sets. All this is good.

Using this criterion we kept in or threw out items from our initial pool as they correlated with it. We also kept in or threw out items which, although worded differently, appear to measure the same thing by having high intercorrelation.

Our residual scale is thus made up of items which are answered in the same direction as the global item.

This is neat logically, but does the fact that two measures correlate establish the validity of either one? Many statistics teachers delight in leading students through involved demonstrations of equivalence to show that correlation is not causation, or that concomitant variance may be coincidental.

We do have face validity in our scale since we simply asked the straight-forward question, "How well does he teach?" We also have equivalence reliability in the inter-correlations of our items and the criterion. We have a logical validity because we inserted items which we rationally relate to effectiveness in teaching. Whether we have predictive validity, we will not find out. We have no way of knowing when the prediction is achieved. We have to use our own scale, or one like it at both ends of the prediction.

This is one of the vagaries of this kind of measurement and no way out is evident. What is evident is a bootstrap connotation. Or, in another vein, a self-fulfilling prophecy. We prophesy what traits distinguish good and poor teachers. We build a scale which exemplifies

these traits, test it against itself, and demonstrate that our prophecy holds true.

Circularity in logic may be better than no logic if we limit our conclusions accordingly.

Assumption: That an initial pool of items samples true effectiveness and that refinement processes sharpen the instrument.

How good was our initial pool of items? This depends upon the breadth and the clarity of our discrimination of good and poor teaching. It also depends on our ability to articulate them in an inquiry which will cause someone else to respond in terms of the perceptions we began with. Remembering the many arguments we have heard and shared over what is goodness and poorness in teaching, there seem to be enormous differences of opinion. It really is amazing that any item ever survives the testing.

Some researchers have questioned students as to their perceptions of best and worst teachers and pooled these assertions to make up scales. Others have used their armchair speculation to assemble items. Actually, this may be the same process since the researchers were recently or remotely students themselves.

What does one put down when he is asked to contribute items to such a scale? It may be that he has studied with care teaching-learning processes and deduced from them general attributes or behaviors of teachers which move learners toward objectives. This would seem to be the safest of assumptions about sources of items. On the other hand, he may put down his casual, rather than studied, perceptions. He may put down what made him comfortable, what characterized a teacher he "liked," quite independent of effectiveness. If his own approach to intellectual experience depends on appeal to authority, he probably places highest authoritarian traits; while, if he enters learning experiences with inquiry, he most likely reports the teacher who helped him form questions.

The point is that the means and variances which emerge from our sampled populations and the intercorrelation of items may not really be a delineation of attributes of teaching effectiveness. They may express the distribution of learning styles in our respondents.

This is not to suggest that this problem is appreciably different from the problems of inferences from most psychological measures. Caution is required here more than elsewhere because we can be deceived by the seeming validity of our first question, "How well did he teach?"

An interesting observation came to me this week. One of my students is preparing a dissertation which is an examination of the experimental designs and statistical treatments in about 150 studies of teacher ratings by students. He tells me that most of these used home-made scales and all the scales involve, with varied wordings, the same 40 or 50 items. I don't know what this means, but either teaching is quite a simple task or there is much stereotypy in scale building.

Assumption: That a scale for rating teaching effectiveness can yield evidence beyond the quality of the items which make up the scale.

This assumption is involved in several others. It merely restates some of them in different content and points to a temptation in psychometrics. We assemble, perhaps somewhat casually, a list of items. We edit and try it out on population samples, make varied statistical gestures over obtained scores. By the time we have all the difficulty levels, discrimination values, intercorrelations, even the factors, the thing takes on a sort of holy glow.

If we have been precise, it is now a reliable measure of whatever it does measure. Maybe this was a ratio of hip and shoulder circumference with head diameter held constant, but now we begin to see that we really have a test of intelligence, temperament, and creative talent! Thousands of master's candidates have walked these garden paths and been given their degrees.

I have exceeded my point. We are not likely to get any more out of a test than we put into it.

Assumption: That a teacher rating scale can be developed in one frame of reference, e.g., teaching as a situation of teacher-students confrontation, over a fixed series of hours, on specified substance, with consensual rules of order, and applied in instructional settings in which any of these factors is varied.

I expect we have each taught classes which correspond reasonably to this model and we would accept a scale built around it. Probably we have also taught where it simply does not fit. In one set of values, the ideal kind of class is one in which the students do all the work. They prepare, execute, and evaluate themselves. If this seems overly avant garde for your tastes, I suggest you re-read some elementary lessons on the psychology of learning or go back to John Dewey.

Psychologists have been responsible for most of our teacher rating scales with a large disregard for the rest

of their science. They make teaching the center of the enterprise. Learning is the center and the teacher is a facilitator. Teaching which does not engage the learner may be beautiful art but it is not teaching.

Assumption: That the effectiveness of a teacher can be rated apart from the receptiveness and responsibilities of the students who rate him.

While there may be precedent, I have not seen a teacher rating procedure which obligated the rater to qualify his ratings according to the nature of his contribution to the teaching-learning situation, his motivation, his diligence, his readiness. I have seen scales which ask, "What grade do you expect?" There may have been one Bill Bendig used which asked, "What grade do you deserve?" but we admit grades to be imperfect symbols of involvement.

There is a Wonderland connotation in this, that teaching can be extracted from the teaching-learning situation and scaled by itself. It is like the play run through without audience.

It is almost certain that there will be a positive correlation between the actor at rehearsal and the actor with an audience, but it is by no means certain that the correlation will be high.

In a sense, this is the heart of our problem. We purport to rate <u>effectiveness</u> of an interaction by describing only one side of the interaction.

Assumption: That the teacher has a constant pattern of output unchanged by interaction with his students. (This assumption is second-order. It emerges from the data on reliability rather than from the initial rationale.)

You and I do not believe we are constant, but we keep coming out in a series of exact replications, as our students see us. It could be that we are stereotyped far beyond our insight. It could be that there is something built into the system of rating which taps only our constants and not our adaptations.

It seems as if the items which survive statistical treatment and constitute our scales tap deep-seated traits, predispositions to act in certain ways, not subject to short-time change. Take them away from us and we would be reeds blown in the wind. But, given only these traits we are automatoms. Neither extreme is realistic.

We have stable traits and we have adaptability. The scales seem to measure the stable.

Another kind of scale is feasible; one that picks up the adaptations. I haven't seen one, and I doubt it would look conventional, but it might be interesting. (Anyone in the market for a dissertation topic?)

Assumption: That knowing how a teacher stands in relation to a peer group actually tells anything about his effectiveness as a teacher.

Raw scores usually tell us nothing at all. If my profile shows me high here and low there by frequency count, I am none the wiser since the elevation of a score depends on the question. It may be that nearly everyone is high on this and low on that. If we want to know more, we establish norms which tell us how other people score on these items.

Here we have all the inherent problems with norms. A teacher is found to have all relatively high scores but perhaps he is being compared with a population of poor performers. Or, he looks bad because he is batting in the major leagues. There are great differences in campus folkways. Even on the same campus, math teachers have the boys at the chalk board, history teachers lecture, and the speech teachers concentrate on group process. Most norms will lump all these together.

I think there is no doubt that relatively relevant norms can be built so we can hold a teacher up against his peers. Unfortunately such norms are rare. They may be rare and also non-contributory. The speech man may be pared with speech men, but how does a speech teacher compare with a history teacher, and both with a math teacher? This is the kind of question asked as soon as we get beyond the department level or play with productivity indices.

We get an answer when we lump them and we have another chewing gum weld in the framework of our instrument.

Assumption: That being informed of how his students see him and how he will be rewarded or punished will cause an instructor to improve.

This is an excellent recipe in theory and perhaps in fact for neurosis. To be motivated to change can result in favorable change. Without potential for change, it can also result in rationalization, projection, distortion of values and a whole syndrome of defensiveness.

The trouble here is that the great majority of college teachers have had no training for the task. The Ph.D. has

done intensive research in a carefully restricted area. What he knows about teaching, about modifying the behavior of others or of himself, is usually based on the teachers who taught him and some homespun introspection.

I would suggest in deep seriousness that instituting a rating system while leaving improvement to chance is an invitation to catastrophy.

Assumption: That ratings obtained on a collection of items can be summed in some manner to yield a composite score and that this score can be inserted meaningfully into a promotion formula or a productivity index.

Let me note firmly -- this assumption is not always made. The preference of scale developers in the past has been the examination of profiles, not single scores.

When we attempt summation the result may not truly be the sum of the items. We can add apples and oranges if we wish. We then know how many pieces of fruit we have. Their species identities are lost. As I pointed out elsewhere, the scale items which survive the statistical process tend to be the more inflexible attributes of personal structure. When we lump them into one score, we have a symbol which conceals more than it reveals. It takes a very long inductive leap to say that it is an assessment of effectiveness in teaching. Perhaps it is but, as Scrooge said to the portly gentleman, "Pardon me, but I don't know that."

Assumption: That the rating of <u>teaching</u> effectiveness is a rational variable in the judging of the merit of a <u>teacher</u>.

We rate teaching effectiveness. We retain or dismiss a teacher. What I say here may seem like hair-splitting, but I think it relevant.

This is my own value system but, to me, the teacher is far more than one who performs before groups of students for his 12 hours a week, or his 12 hours plus his 30 of preparation and evaluation.

I may be far out on a humanistic limb here, and I will not belabor the point or ask you to accept it. To me, the teacher is a person who has a life of his own, habits and values of his own. These may have a significance in his role in the collegiate community but escape any system of rating.

Granted, I think he should have something more tangible than soul, but I worry about his soul being blotted

out by his relative standing in a statistic of social performance.

Assumption: That the rating of teaching effectiveness is in the best interests of student-teacher relationships.

My attitude toward my surgeon is going to be affected if he says, "That appendix will have to be removed. After I have taken it out I shall ask you how well you think I did it." I shall look differently, probably askance, at my preacher if he asks me at the church door, "How many inches toward salvation did I move you today?" At Howard Johnson's restaurant there is a card you can fill out, if you as an individual wish to, but your dentist says, "Now that feels better doesn't it?" and not, "How well do you think I did?"

Professors historically entrapped themselves into an evaluator role. If we think a bit about our origins, we are generally less than 100 years away from our clerical forebears, but we forgot something they knew. The minister doesn't tell anyone he is so many quality points away from salvation. He may go so far as to say, "If you go on the way you are now . . ." or he may indict the whole congregation. We made, achieving our lay status, a big thing out of scaling people. Now, we have turned the process around and the evaluee is evaluating the evaluator.

I think this phenomenon is amusing to speculate about but it is not the real issue. There is a basic principle in evaluation. Any modification of the attitudes of those evaluated or those evaluating can have an influence on the results. If I am a student in your class for a term at the end of which you surprise me with a request for me to rate you my response set is different from that which I would have if it is taken for granted all during the term that I can praise or blast you at the end. On the other hand, you must be a remarkably self-confident person if you move through the term completely indifferent to the fact that I shall have this privilege at the close.

This situation is complicated still further by knowledge by either or both of us that the results of my evaluation of you will reach a third person who has decisions to make about your future.

In other words, the process itself is altered by the very fact of its existence as a process. I think we know little or nothing about how it is altered or how much.

Assumption: That we can perpetuate the use of these scales, coupling them with systems of rewards and punishment, and avoid producing an excess of conformity behavior at the expense of innovation.

Widespread use of such scales, especially if they are used to decide promotion and merit increments, communicates clearly to an instructor how to get ahead. He can shape up or ship out. Efforts at innovation, creative approaches to teaching, cultivation of idiosyncratic patterns which made colorful some of the professors we remember -- for good or bad, these are gambles for the foolhardy.

If we succeed in making all professors alike, the future for higher education is indeed bleak.

There is another complication to this. If everyone is trying to be above the mean, the mean must inevitably change. It has to keep moving up. If we succeed in increasing motivation without any systematic processes for changing potential we simply can go about so far and we stall our motors.

Assumption: That the rating scales of the past and present have much of anything to do with the almost inevitable instructional designs of the future.

In my crystal ball, I see few remaining teachers in the design I mentioned of students-teacher confrontation, specified sequences, pre-determined substance, and standard rules of order. The teacher as purveyor of knowledge -- we have learned to use lecture as almost synonymous with teach -- has almost faded from my forecast.

I see two functionaries in the future. Hopefully, they will interact but probably not enough. One of these carries the knowledge purveyance role but his preoccupations range far beyond a faith in standing before a captive audience more or less at his convenience while he dictates the truth into their notebooks. He is designing programs, with many aids, for self-instruction.

The other busy professor, I foresee, is closer to teacher but he may be more like the reader's consultant in the library. He meets under very flexible situations with individuals or groups to stimulate and aid the integration and evaluation of knowledge and especially the socialization of knowledge.

The scales we now have will fit neither.

CONCLUSION

Again let me say, I am not against scales for students to rate teachers. I urge them on you as probably more objective, reliable, and perhaps valid than any alternative procedure we now use for such assessment.

I also urge that we use them with proper attention to their nature. They yield pooled reactions to those dimensions which were built in. They presume uniformity of conditions, styles, and purposes. They are shaped by statistical treatment not to be additive to single scores. They are interpretable only against ideals of which no model has consensus or against norms which may or may not be appropriate. They have an influence on teacher-student relations which is unknown. They have a built-in predisposition to establish models to be copied.