DOCUMENT RESUME

ED 110 484 TM 004 757

AUTHOR Nadeau, Gilles, G.
TITLE Validity of Student Ratings of Instruction: Validity for What Purpose and What Kind.
PUB DATE [Apr 75]
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Effective Teaching; Statistical Analysis; Student Evaluation; *Teacher Evaluation; *Teacher Rating; *Test Validity; *Validity

ABSTRACT

Four statements formulated after an extensive literature review on student ratings of instruction are proposed: (1) we must remember the basic definition of validity, (2) we must clarify what it is that a particular teacher is trying to do or proposes to do in a given classroom, (3) we must be clear about defining what we want to obtain from student ratings, and (4) we must make a greater effort to measure student performance as a result of or in spite of what the teacher intended to do and what actually was done. A Taxonomy of kinds of validity is presented. The identification of validity for a particular purpose, situation, and group is discussed. The need for a more sophisticated analysis of student rating data, utilizing more recent statistical tools such as discriminant analysis and multivariate procedures, is pointed out. (Author/BJG)

# VALIDITY OF STUDENT RATINGS OF INSTRUCTION: VALIDITY FOR WHAT PURPOSE AND WHAT KIND.

by

Gilles G. Nadeau, Ph.D.
Associate Professor
Université de Moncton
Visiting Professor
Centre for Learning and Development
McGill University





Paper presented as part of a Symposium entitled: "Conceptualizations of VALIDITY FOR Student Ratings of Instruction", American Educational Research Associations, Washington, D.C., April 1975.

# VALIDITY OF STUDENT RATINGS OF INSTRUCTION: VALIDITY FOR WHAT PURPOSE AND WHAT KIND OF VALIDITY?

by

Gilles G. Nadeau, Ph.D.
Association Professor
Université de Moncton
Visiting Professor
Centre for Learning and Development
McGill University

The title of my paper is: "Validity of Student Ratings of Instruction: Validity for What Purpose and What Kind of Validity?" My presentation amounts to four statements that I have formulated after an extensive review of literature on Student Ratings to be published in the near future by Clarke Irwin of Toronto.

STATEMENT I.

WE HAVE FORGOTTEN OR IGNORED THE BASIC DEFINITION OF VALIDITY - LET'S REMEMBER IT.

STATEMENT II.

WE OFTEN FORGET OR DO NOT KNOW OR IGNORE WHAT IT IS THAT A PARTICULAR TEACHER IS TRYING TO DO OR PROPOSES TO DO IN A GIVEN CLASSROOM. (Course Syllabi and Course Strategies).

LET'S CLARIFY THAT.

STATEMENT III.

WE HAVE TO DEFINE BETTER WHAT WE WANT TO OBTAIN FROM STUDENT RATINGS. (Reported events, attitudes, agreements, judgments, opinions or reactions or what?) LET'S BE CLEAR ABOUT THAT.

STATEMENT IV.

WE NEED TO MAKE A GREATER EFFORT TO MEASURE STUDENT PERFORMANCE AS A RESULT OF OR IN SPITE OF WHAT THE TEACHER OR PROFESSOR INTENDED TO DO AND WHAT ACTUALLY WAS DONE. LET'S GIVE OUR BEST EFFORTS TO THAT.

NOW Mr. Chairman WITH these four GIVEN'S I think I should sit down and open the floor for questions and discussion.

BUT maybe it would be beneficial to provide some explanations. SO back to Statement I. WHAT IS VALIDITY?

In my frame of reference "Validity is the ability of an instrument to measure what it was designed to measure." Now we know, don't we, that there is no such thing as a valid instrument. Usually we say that the results of a particular instrument are valid for a particular purpose, in a particular situation and with a particular group.

Here is a TAXONOMY OF KINDS OF VALIDITY .

- - - - - - - -

INSERT FIGURE I

- - - - - - - -

Now there are 24 definitions or kinds of validity which can be subsumed under four categories which themselves belong to either of two major types of validity. I ask you, what kinds of validity are we dealing with when we examine the host of studies done about and around student ratings. I propose to you that a study be done to classify reported and unreported studies according to these twenty-four kinds as a start to comprehend further the state of validity of student ratings. It is no easy task and I will give you some examples later. Which kinds of validity do people have in mind when constructing their questionnaires or rating forms?

VALID FOR A PARTICULAR PURPOSE:

Aleamoni et al. (1973) have shown that results of student ratings are different when instruction to students indicates that the purpose is for course improvement on the one hand and for P & T decisions on the other hand. What does this tell us about results of student ratings when the purpose has not been specified or is left vague or is otherwise ambiguous to the student? A similar study is now being done at McGill (Levy, 1975; Pascal, Nadeau, Shore) with four sets of instructions. We are anxious to see the results.

VALID FOR A PARTICULAR SITUATION :

What does this tell us about the validity of data on student ratings when one uses, uncritically, entire instruments developed elsewhere for purposes sometimes unknown? I believe this tells us that an instrument must have content validity and "Content Agreement" before it can be used with a particular professor and classroom, with the hope that the results are going to be valid.

I believe it also tells us that before an instrument is developed for use in a particular classroom for a particular group with a particular instructor, one needs to have a clear definition of the situation in which the instrument is going to be applied, namely, what is the classroom organization and management situation? (Is it a lecture, discussion, seminar, independent study, projects, group work; the list and combinations are endless). What are the teaching and learning activities and strategies? I believe we need to give answers to these before we can examine if results of ratings are valid.

VALID FOR A PARTICULAR GROUP :

Now, surely, there should be little quarrel about that. Recent work by Doyle et al. (1973) and others regarding student types is getting us to meet that particular requirement of validity. Also it should be clear that we

need to understand and measure, in increasingly better ways, the performances of students and student groups as they try to achieve specified objectives with particular strategies under specified constraints. To my mind the studies of relationships of student ratings and student performances are gravely lacking in those respects. What about the non-measurement or non-assessment of entry behaviours of students in courses? What does this do to the ratings and the post-achievement of students?

When I look at the available validity studies of student ratings, I believe that answers to the above questions will bring us closer to identifying specific teacher skills that need to be developed as part of a teacher's repertoire of teaching behaviours that will be related to specific performances of learners.

And finally, I also believe that we need to get a little more sophisticated in our analysis of student ratings data by making use of some of our more recent statistical tools such as discriminant analysis, multivariate procedures in order to get at some aspects of the validity problem. In this context longitudinal studies are of prime importance.

Table I. gives you some of the studies and their pre-occupations.

- - - - - - - - - -
INSERT TABLE I
- - - - - - - - - -

Investigations of validity of student ratings fall generally in three categories:

a) Rating form content validity.

b) Correlates of student ratings.

c) Comparisons of student ratings with ratings of other raters.

Table I gives you a start in your additional search for understanding of the validity question in the field of student ratings of instruction. I might add that out of the 123 validity studies listed 46 are dated in the 1972-74 period, 40 are in the 1968-71 period with the remaining 37 studies before 1968. Most of the validity studies are therefore relatively recent indeed.

## REFERENCES

Costin, F.; Greenough, W. T., and R. J. Menges. "Student Ratings of College Teaching: Reliability, Validity and Usefulness." Review Educ. Research. Vol. 14, no.5, 1971, pp. 511-535.

Doyle, K. O., and Whitely, S. E. "Reply to Rodin and Rodin." Measurement Services Center, University of Minnesota: November 1972.

McKeachie, W. J., et al. "Student Ratings of Teaching Effectiveness Validity Studies." American Educational Research Journal. Vol. 8, 1971, pp. 435-445.

Murray, H. G. "The Reliability and Validity of Student Ratings of Faculty Teaching Ability." Submitted to Journal of Educational Measurement. January, 1973. (a)

Nadeau, G. G. "Student Evaluation of Instruction: The Rating Questionnaire." In Knapper, Chris K., et al., (eds.), Scaling the Ivory Tower: Appraising Teaching in Higher Education. Toronto: Clarke-Irwin. (In preparation).

Sockloff, Alan L. Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia Penn.: Measurement and Research Center, Temple University, January, 1974. (Contains the fourteen papers of the Conference).

Touq, M. S.; Feldhusen , J. F., and Halstead, J. "Criterion Referenced Validity of Student Ratings of Instructors." Paper Presented at the Annual Meeting of the American Educational Research Association. New Orleans, Louisiana: February, 1973.

# FIGURE I

## KINDS OF VALIDITY

**Validity:** the state, status, or fact of being valid, sound, quality being grounded on truth or fact, truthful, in measurement, the extent to which a test or other measuring device does what it is supposed to do

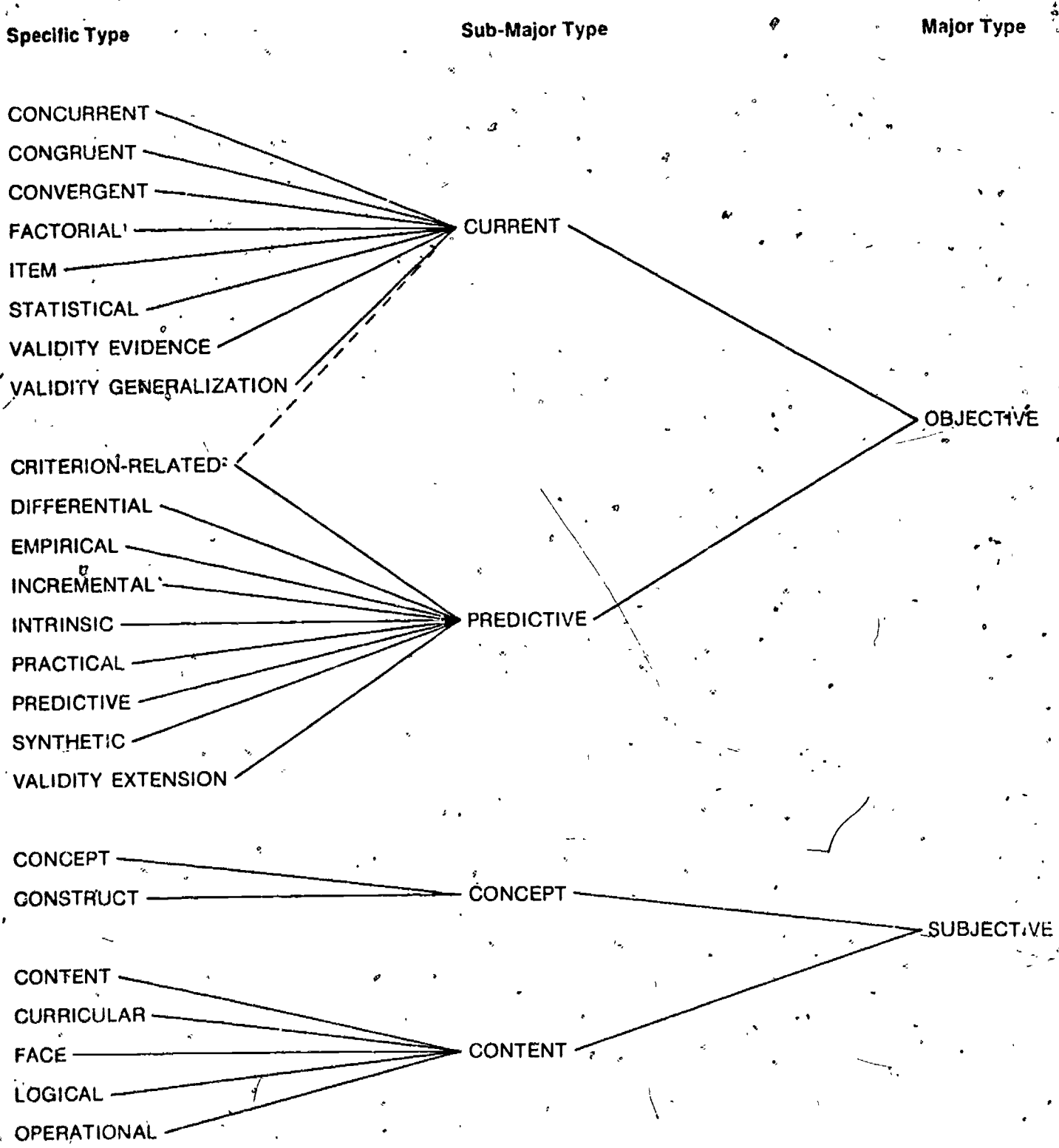
**Types of Evidence for Validity.** investigated by analysis of test content or by a study of relationships between test scores and criterion variables, independence of methods being a common denominator among the major types of validity excepting *content* validity

1. **Concurrent:** validity based upon correlation with a criterion variable that is measured at about the same time as the test administered
2. **Congruent:** evidence of validity obtained by correlating a test with an existing similar measure of the same function (e.g., correlation of a new intelligence test with an existing intelligence test).
3. **Convergent:** type of validity which requires a high correlation between a test and other variables which logically are related to the test, confirmation by independent measurement procedures
4. **Factorial:** a form of content validity, uses factor analysis to determine to what extent a test measures certain content areas, partitions true score variance into subcomponents which indicate the extent to which each factor is a subcomponent.
5. **Item:** discriminative value of an item; correlation between an item and some criterion of performance.
6. **Statistical:** evidence of test validity expressed numerically, usually as correlation between scores on the test and another set of measures such as scores on another test, teachers' marks, ratings by experts, etc
7. **Validity Evidence:** information gathered to determine exactly what kind of inferences can be made from test scores.
8. **Validity Generalization:** process in which additional information is obtained by checking the effectiveness of the test on a differently defined population but using the same criterion as in the original study.
9. **Criterion-related:** validity demonstrated by comparing test scores with one or more external variables considered to provide a direct measure of the characteristic or behavior in question, correlation between test score and criterion measure; test user wishes to forecast an individual's present or future standing on some variable of particular significance that is different from the test
10. **Differential:** validity which depends on difference between correlation of classification test (ideal test) with each of separate criteria to be predicted, with a two-criterion classification problem the ideal test would have a high correlation with one criterion and a zero or negative correlation with the other criterion.
11. **Empirical:** quality of test having definite and proved value for a given purpose; usually stated in terms of correlation; extent to which scores on a test agree with some outside criterion or future measure of success.
12. **Incremental:** amount the test will add to validity of predictions made on basis of data usually available, validity stated in terms of some increment in productive efficiency over information otherwise easily and cheaply available.
13. **Intrinsic:** validity evidence based on fact that items in a test are selected to stimulate the criterion item that the test is used to predict.
14. **Practical:** validity of a test as determined by its ability to predict within a certain sphere of behavior.
15. **Predictive:** (of a test) validity based upon correlation with a criterion variable that is not available until some time after testing (e.g., school grades)
16. **Synthetic:** validity for which each predictor is validated, not against a composite criterion but against job elements identified through job analysis; the validity of any test for a given job is then computed synthetically from the weights of these elements in the job and in the test.
17. **Validity Extension:** process by which test validity is checked against a new criterion as well as with a different population.
18. **Concept:** attempt to analyze the validity of broad concepts in subject areas.
19. **Construct:** validity evaluated by investigating what qualities a test measures by determining degree to which certain explanatory concepts or constructs account for performance on the test.
20. **Content:** validity demonstrated by showing how well the content of the test samples the subject matter about which conclusions are to be drawn, test user wishes to determine how individual performs at present in a universe of situations that test situation is claimed to represent.
21. **Curricular:** evidence of test validity indicated by agreement between test content and curricular content and test objectives and curricular objectives.
22. **Face:** validity referring to what a test appears to measure on basis of subjective evaluation, not what it actually measures; least justifiable of all evidences of validity.
23. **Logical:** estimate of content validity based on comparison of behavior demanded by the test with the behavior that, by a prior analysis, belongs to the variable to be measured.
24. **Operational:** ability of a test or measuring instrument to do some task, defined in terms of operations it actually performs (e.g., a yardstick is operationally valid for linear measurement).

Taken from: CEDR Quaterly, Phi Delta Kappa
Spring, 1974.

FIGURE I CONT'D...

TREE DIAGRAM DELINEATING TYPES OF EVIDENCE FOR VALIDITY INTO SUB-MAJOR AND MAJOR CATEGORIES




[1]Though defined in the Dictionary as a form of Content validity, it utilizes quantitative techniques for justification of evidence.
[2]Appropriate to both Current and Predictive one may wish to determine current standing or forecast future position

TABLE I.- Some Validity Studies in the Student Ratings Literature.

| AUTHOR (S) | DATE | AREA |
|---|---|---|
| Halstead | (1970) | A. Content Validity (contents agreement, factor structure) |
| Kent | (1967) | |
| Holmes | (1971) | |
| Davis, Hildebrand & Wilson | (1971) | " |
| Despande et al | (1970) | |
| Crawford & Bradshaw | (1968) | " |
| Coffman | (1954) | |
| Warrington | (1973) | " |
| Costin | (1968) | |
| French | (1957) | " |
| Tuickmann | (1973) | " |
| Gadzella | (1968) | |
| Hoyt | (1973) | " |
| Musella and Rush | (1968) | |
| Gagné & Chabot | (1970) | " |
| Perry | (1969) | |
| Perry & Baumann | (1973) | " |
| Downie | (1952) | |
| Aleamoni | (1973) | " |
| Mann | (1969) | |
| Langen | (1966) | " |

1. Adapted from Nadeau, G. G. "Student Evaluation of Instruction: The rating questionnaire" in Chris Knapper, George Geis, Charles Pascal and Bruce Shore (eds), Scaling the Ivory Tower: Appraising College & University Teaching. Clarke Irwin, Toronto, 1975.

TABLE I. - CONT'D

| AUTHOR (S) | DATE | AREA |
|---|---|---|
| | | B. Bias in responses (halo, leniency, popularity, hostility, etc.) |
| Kirchner | (1969) | |
| Sockloff | (1973) | |
| Hoyt | (1969) | " |
| Royce | (1956) | |
| Widlak, McDaniel, Feldhusen | (1973) | |
| Weavier | (1960) | |
| Remmers | (1959) | " |
| Sharon | (1970) | |
| Guthrie | (1954) | " |
| Centra | (1973) | |
| Aleamoni | (1972-73) | " |
| | | C. Correlates |
| | | 1. Student characteristics sex, age, class standing class size, grades, basis for judgements, achievement, etc. |
| Cohen & Berger | (1970) | |
| Lathrop | (1968) | |
| Frey | (1973) | |
| Lathrop & Richmond | (1967) | |
| Bentley | (1971) | " |
| McKeachie et al. | (1971) | |
| Nichols & Soper | (1972) | " |
| Mann | (1969) | |
| Shuh & Crivelli | (1973) | " |
| Kohlan | (1973) | |
| Elliot | (1950) | " |
| McClelland | (1970) | |
| Costin et. al. | (1971) | " |
| McKeachie and Solomon | (1958) | |
| Perry & Baumann | (1973) | " |
| Rodin & Rodin | (1972) | |
| Whitely & Doyle | (1973) | " |
| McKeachie | (1969,1973) | |
| Colliver | (1972) | " |
| Gransin & Painter | (1973) | |
| Whitely & Doyle | (1974) | " |

TABLE I. - CONT'D

| AUTHOR (S) | DATE | AREA |
|---|---|---|
| Mueller & Miller | (1970) | Correlates |
| Miller | (1972) | 1. Student characteristics sex, age, etc... |
| Remmers | (1960) | |
| Russell | (1951) | |
| Voeks & French | (1960) | " |
| Caffrey | (1969) | |
| Spencer | (1968) | " |
| Rubenstein & Mitchell | (1970) | |
| Treffinger & Feldhusen | (1970) | " |
| Walker | (1969) | |
| Yonge & Sarrenrath | (1968) | " |
| | | Course characteristics |
| Murray | (1973) | 2. (type of course, content difficulty, required vs elective, level, time of class, etc.) |
| Nichols & Soper | (1972) | |
| Remmers | (1959) | |
| McKeachie | (1973) | " |
| Marlman | (1973) | |
| Gage | (1961) | " |
| Villano et al. | (1974) | |
| Miller | (1972) | " |
| Guthrie | (1954) | |
| Eckert & Keller | (1954) | " |
| Lovell & Haner | (1955) | |
| Clark & Keller | (1954) | " |
| | | Instructor characteristics. |
| McKeachie | (1973) | 3. sex, age, rank, degrees, experience, grading standards, knowledge of subject, research, knowledge of teaching, personality traits, popularity and change after feedback etc. |
| Remmers | (1959) | |
| Costin et al. | (1971) | |
| Riley et al. | (1950) | |
| Stallings & Singhal | (1969) | |
| McGrath | (1962) | |
| Bressler | (1968) | " |

| AUTHOR (S) | DATE | AREA |
|---|---|---|
| Costin | (1968) | 3. Instructor characteristics |
| Guthrie | (1954) | |
| Hayes | (1971) | sex, age, rank, etc... |
| Voeks | (1962) | " |
| McDaniel & Feldhusen | (1970) | |
| Murray | (1973) | " |
| Richardson | (1973) | |
| Isaacson et al | (1963) | " |
| Clark & Blackburn | (1973) | |
| Sorey | (1968) | " |
| Sherman & Blackburn | (1974) | |
| Miller M. | (1971) | " |
| Thomas | (1969) | |
| Bentley | (1971) | " |
| Centra | (1972,1973) | |
| Aleamoni | (1973) | " |
| Hoyt | (1973) | |
| Tuchkman | (1973) | " |
| Tuckman & Oliver | (1968) | |
| Braunstein et al. | (1973) | " |
| Rayder | (1968) | |
| | | D. Student ratings versus other raters, alumn , colleagues, head of departments, obser rs etc... |
| Costin et al. | (1971) | |
| Murray | (1973) | |
| Sockloff | (1973) | |
| Hayes | (1971) | |
| Drucker & Remmers | (1951) | " |
| Costin | (1966) | |
| Webb & Nolan | (1955) | " |
| Perry | (1969) | |
| Wilson et al. | (1973) | " |
| Gaff | (1973) | |
| Touq et al. | (1973) | " |
| Centra | (1973) | |

TABLE I. - CONT'D

| AUTHOR (S) | DATE | AREA |
|---|---|---|
| | | D. Student ratings versus other raters, alumni, etc... |
| Braunstein & Benston | (1973) | |
| Guthrie | (1949, 1954) | |
| Maslow & Zimmerman | (1956) | " |