DOCUMENT RESUME

ED 379 292 TM 022 637

AUTHOR Tagomori, Harry T.; Bishop, Laurence A.
TITLE Content Analysis of Evaluation Instruments Used for Student Evaluation of Classroom Teaching Performance in Higher Education.
PUB DATE Apr 94
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Faculty; Colleges; College Students; Content Analysis; *Evaluation Methods; Higher Education; Instructional Effectiveness; Performance; Responses; Student Attitudes; *Student Evaluation of Teacher Performance; Test Construction; Test Items; Test Use; *Test Validity

ABSTRACT

A major argument against evaluation of teacher performance by students pertains to the instruments being used. Colleges conduct instructional evaluation using instruments they devise, borrow, adopt, or adapt from other institutions. Whether these instruments are tested for content validity is unknown. This study determined how evaluation questions were presented in a sample of 200 evaluation instruments collected from 414 schools of education. Analysis of the evaluation questions indicated that there are questions of validity in many of the evaluation instruments used by these colleges. No particular instrument is accepted by all colleges and universities, and few tests of vaiidity have been performed for instruments in use. Flawed responses were skewed, ambiguous, or unclear, or else they did not correspond with the evaluation question. Some 58% of instruments were found to contain such flaws. Six tables and nine figures present study findings. (Contains 29 references.) (SLD)

CONTENT ANALYSIS OF EVALUATION INSTRUMENTS USED FOR STUDENT EVALUATION OF CLASSROOM TEACHING PERFORMANCE IN HIGHER EDUCATION

Harry T. Tagomori
University of Hawaii at Manoa
Laurence A. Bishop
University of San Francisco

A major argument against instructional evaluation pertains to instruments used. Colleges conduct instructional evaluation using instruments they devise, borrow, adopt or adapt from other institutions. Whether these instruments are tested for content validity is unknown. This study determined how evaluation questions were presented in evaluation instruments. Based on analysis of the evaluation questions contained in the instruments, a question of validity in many of the evaluation instruments used by colleges exists.

Introduction

Instructional evaluation has been recognized as an important function in education for centuries. It goes as far back in history as 350 A.D. (Doyle, 1983). During the early Christian era through the middle ages students paid their fees directly to their teachers where more esteemed teachers were paid a higher fee. This practice in essence was a form of student evaluation of instruction (Seldin, 1980).

Research on instructional evaluation is extensive and many times conflicting (Miller, 1987). The literature on instructional evaluation is broad and many conferences have addressed instructional evaluation. However, little is

known about the validity of the instruments used by colleges and universities.

Some institutions conduct instructional evaluation from instruments they develop themselves. Others conduct their evaluation from instruments they either borrow, adopt or adapt from other institutions. Whether those instruments are tested for content validity is unknown. As a result, faculty members in higher education may be evaluated with invalid instruments, conceivably leading to unfair assessment of their teaching performance.

A major argument against faculty evaluation is that the instruments used are incomplete or improperly phrased and may be subject to the student's interpretation (Morton, 1964). An evaluation instrument plays an important part in how objectively faculty members are judged. The instrument should link what students in the classroom observe with the quality of a faculty member's teaching skill. The strength of an evaluation instrument, therefore, depends on how valid the instrument is.

Purpose of the Study

The purpose of this study was to analyze the content of evaluation instruments used in student evaluation of classroom teaching performance to determine how evaluation questions were presented in the evaluation instruments. By examining the content of the instruments, conclusions can be drawn as to the validity of the evaluation instruments.

Inferences can then be made about the evaluation instruments being used by colleges and universities. Information gleaned from this study was also to provide new knowledge and direction for future research and development that may lead towards more valid faculty evaluation instruments.

Methodology

Population Sample

The population sample consisted of schools of education accredited by the National Council for Accreditation of Teacher Education (NCATE). NCATE is an independent agency recognized by the Education Department, Washington, D. C. and authorized by the Council on Postsecondary Accreditation (COPA) to accredit schools of education at colleges and universities colleges in the United States that prepare professional educators to staff preschool through secondary school programs (NCATE, 1988). At the time of this study, the number of schools accredited by NCATE was 517.

All 517 schools of education were contacted requesting a copy of the evaluation instrument they used to assess classroom teaching performance. Of the 517 schools contacted, 414 (80%) responded with a copy of their instrument.

Research Design

A descriptive design utilizing a content analysis research method was used to analyze the content of evaluation instruments used by schools of education

accredited by NCATE. Content analysis is basically a process where specific behaviors or actions are recorded. The behaviors can be in the form of written documents (Berelson 1952; Holsti, 1969; Krippendorf, 1980), such as instruments used to evaluate teaching performance.

The intent of the study was to systematically describe the composition of the content of evaluation instruments used by institutions to evaluate classroom teaching behavior. The basic goal was to convert nonquantitative units of analysis (evaluation questions) into quantitative data applying frequency-count recording and by a frequency distribution (Bailey, 1982).

The sample size for a given population was determined by using the significant sample size table developed by Krejcie and Morgan (1970). Based on this table, the sample for 414 instruments was determined to be 200 (N = 200). In transforming the content in the evaluation instruments into quantitative data, a systematic coding system was developed and applied to the study.

Procedure

The first step was to define the criteria with which to quantify the content of the evaluation instruments. This was done by a content analysis of randomly selected samples of the 200 evaluation instruments to be analyzed for this study. From the analysis of this sample, the criteria were defined and the type of flaws contained in the evaluation

instruments were identified. Thus, the criteria obtained from this study emerged or derived from the evaluation instruments being analyzed.

The criteria defined in this study are the type of flaws contained in the evaluation questions, the type of flaws in the responses to evaluation questions, and evaluation questions that did not correlate with classroom teaching performance. The flaws in the evaluation questions are defined as ambiguously stated, unclearly stated, and subjectively stated questions. The flaws in the responses to evaluation questions are characterized as ambiguous, skewed, and not clearly defined responses. Evaluation questions that do not correlate with classroom teaching performance are questions which describe behaviors not relevant to classroom teaching.

The second step was developing the means to quantify the content of the evaluation instruments. Two methods were used for this purpose. One was a simple binary coding (frequency-count recording) to indicate whether or not an item correlating with a specific criteria appeared in the evaluation instrument. The other method used was the frequency (frequency distribution) with which the criteria appeared in the evaluation instrument (Borg and Gall, 1989).

In the third step a coding scheme to transform and record the data into numbers (frequency-count recording and frequency distribution) was created. The coding scheme

developed consisted of three main parts: (1) The types of item flaw or criteria were indicated in the first column of the coding scheme. These were shown as ambiguous items, unclear items, and subjective items. (2) The schools of education from colleges and universities accredited by NCATE. The schools (hence evaluation instruments) were identified by code numbers 1 to 200 in the coding scheme. (3) The frequency of occurrence in evaluation instruments were the number of flawed items found in each instrument. These are the numbers shown under each school coded 1 to 200. These numbers also represent the data collected for each instrument. Table 1 illustrates a sample of the coding scheme developed for this study.

TABLE 1

ANATOMY OF THE CODING SCHEME USED FOR FREQUENCY DISTRIBUTION OF FLAWED EVALUATION ITEMS (N = 200; n = 4,028)

Type of Item Flaw or Criteria	Schools of Education from Colleges and Universities Accredited by NCATE (Evaluation Instruments)										
	Frequency of Occurrence in Evaluation Instruments										
	1	2	3	4	5	6	7	8	---->	200	Total
Ambiguous items	2	0	3	0	3	2	1	0	---->	3	910
Unclear items	1	3	3	3	4	5	3	0	---->	2	849
Subjective items	1	1	1	1	0	0	1	0	---->	0	440
Total	4	4	7	4	7	7	5	0	---->	5	2,199

N = total number of evaluation instruments
n = total number of evaluation items in N

The final step was testing the criteria to support the reliability of the criteria defined in this study. To test for criteria reliability, an inter-coder reliability test of the sample evaluation instruments was conducted. The testing for reliability was to determine the degree of agreement on the criteria between inter-coders. The correlation coefficient between coders was .90.

Data Collection

Once the coding system was established, each evaluation question contained in the 200 instruments was systematically analyzed and recorded into pertinent content criteria. A 12-column accounting ledger was first used as a recording device to manually quantify the raw data extracted from each evaluation instrument. The raw data were then transcribed into the final coding scheme, tallied and reported in formal tables of frequency and percentages.

The frequency distribution was the recorded number of evaluation questions present for a specific criteria (See Table 1 above). The frequency distribution, therefore, was the number of evaluation questions (n = 4,028) in the 200 instruments (N = 200) that were ambiguous, unclear, subjective or did not correlate with classroom teaching behavior.

The frequency-count recording was a simple binary count (0 and 1) used to record whether or not flawed responses were present in the evaluation instrument. The frequency-

count recording, therefore, was the number of evaluation instruments (N = 200) that contained response choices that were skewed, ambiguous, unclear or did not correspond with the evaluation question (See Table 2 below).

TABLE 2

ANATOMY OF THE CODING SCHEME USED FOR FREQUENCY-COUNT RECORDING (N = 200)

Type of Flawed Responses By Criteria	Schools of Education from Universities and Colleges Accredited by NCATE (Evaluation Instruments)										
	Frequency of Occurrence in Evaluation Instruments										
	1	2	3	4	5	6	7	8	---->	200	Total
Positively skew	0	0	0	0	1	0	1	0	---->	0	26
Negatively skew	1	0	0	0	0	0	0	1	---->	0	04
Ambiguous	1	1	1	1	0	0	1	0	---->	1	35
Unclear/vague	1	0	1	1	1	0	0	0	---->	1	43
Do Not Reflect	0	0	0	0	0	0	1	0	---->	0	08
Total	3	1	2	2	2	0	2	1	---->	2	116

N = total number of evaluation instruments

Each evaluation question contained in the evaluation instruments was systematically examined and classified according to pertinent content criteria. The nonquantitative data was then transformed and recorded as quantitative data into the coding scheme and reported in formal tables of frequency and percentage.

Criteria Sample and Analysis

The criteria for each type of flaw contained in the evaluation instruments analyzed were ambiguities in the wording of the evaluation questions, vagueness (unclearness) in the wording of the evaluation questions, and subjectively stated evaluation questions. The other type of questions considered flaws in the evaluation instruments were evaluation questions that were not relevant to classroom teaching behavior. The content analysis also investigated the content of responses used for each evaluation question. Criteria for flawed responses were defined (evolving from a sample analysis) and applied to this study.

Ambiguous Evaluation Question

Ambiguity in an evaluation question is one that can be classified in two or more categories. The item is stated in such a way that it allows two or more meanings in the communication. Thus, the wording of the evaluation question can be understood in more than one way or refer to two or more things at the same time. Because the evaluation question concerns more than one characteristic of teaching behavior, the characteristic being assessed in the question is not known. The respondent most likely will be confused as to which variable is being evaluated.

."How clear were the goals, aims and requirements of the course?"

."Students' knowledge, thinking ability, and/or skills were extended as a result of this course."

."Asks thought-provoking questions and encourages students to ask questions, disagree, express ideas, etc."

."Instructor's speech, appearance and poise enhance teaching."

Figure 1. Examples of ambiguous evaluation items taken verbatim from evaluation instruments

In Figure 1, it is not explicit which teaching characteristic (variable) is being evaluated. In the first item, for example, it is not known which characteristic of the course goals, aims, or requirements is being evaluated. Of 4,028 evaluation questions contained in the 200 instruments analyzed, 22.6% of the questions were ambiguously worded.

Unclear or Vague Evaluation Question

An evaluation question that is unclearly stated is one that is not clearly expressed. The item is stated in general or indefinite terms and does not have an exact or precise meaning. The wording of the evaluation question most likely will confuse the respondent. This type of question may lead the respondent to guess (subjectivity) what is being evaluated. The evaluation question is not clearly defined and is confusing or uncertain in content.

."The total experience under the control of this person was very worthwhile."

."Makes use of fair and equitable tests or other means of evaluation."

."What percent of course material my instructor covered I learned?"

."Level of degree of difficulty of the course."

Figure 2. Examples of unclear evaluation items taken verbatim from evaluation instruments

The wording of the example evaluation items in Figure 2 is confusing in content, and the items are not clearly written. In the first example of Figure 2, the teaching characteristic being evaluated is confusing because it is not clear what is meant by either "total experience" or "very worthwhile?"

The second example is also confusing because it is unclear what is meant by "makes use of fair and equitable tests." The manner in which these questions are presented may lead a respondent to guess, at most, what is being evaluated. Of the 4,028 evaluation questions investigated, 21.1% were unclearly or vaguely worded.

Subjective Evaluation Question

A subjective evaluation question is presented in such a way that it assumes the respondent has knowledge of what others in class feel about the instructor's teaching performance. In essence, what the evaluation question asks is to assess the knowledge of other students in the class. As an

example, the first evaluation question in Figure 3 asks the student to determine if others in the class understand the lectures. Other types of subjective evaluation questions rely on the respondent's personal feelings rather than asking an observable and more objective teaching characteristic. The manner in which the evaluation question is stated assumes the respondent has the knowledge on how others (students and the instructor) in the class feel.

."Main points of lectures were clearly understood by students in class."

."The teacher's apparent familiarity with the subject matter."

."How well does the instructor understand you?"

."Students use their mistakes as opportunity to learn."

Figure 3. Examples of subjective evaluation items taken verbatim from evaluation instruments

The examples in Figure 3 ask the respondent to evaluate characteristics that may not be observable. In the first example, how does the respondent to this question know whether other students in class clearly understand the "main points of the lectures?" The subjectiveness of the evaluation items in Figure 3 will most likely influence a biased response because they ask for the respondent's personal opinion, whether or not he or she has knowledge of the characteristics being evaluated. The items lack objectivity forcing the respondent to make judgment calls.

There were 10.9% of the total number of evaluation questions worded subjectively.

Table 3 summarizes evaluation questions that were ambiguous, unclear or subjective. Of the 4,028 evaluation questions investigated, 54.6% were characteristic of these types of flaws.

TABLE 3

PERCENT OF EVALUATION ITEMS THAT ARE AMBIGUOUS, UNCLEAR, OR SUBJECTIVE (n = 4,028)

Type of Evaluation Item Flaw (Criteria)	Number of Evaluation Items	Percent
Ambiguous	910	22.6
Unclear	849	21.1
Subjective.	440	10.9
Total	2,199	54.6

n = total number of evaluation items

The number of flawed items for each criteria in Table 3 may seem small compared to the total number (4,028 items) of evaluation questions analyzed. However, collectively, the high number (54.6%) of ambiguous, unclear, and subjective items contained in the evaluation instruments significantly illustrates the discrepancies found in the instruments analyzed.

Uncorrelated Evaluation Question

An evaluation question not correlated with classroom teaching behavior is one that is not characteristic of effective or ineffective classroom teaching performance.

The evaluation question has no relevance to classroom teaching.

."How well are you able to take notes?"
."The location of this class was convenient."
."How well do you like this instructor?"
."The classroom was comfortable."

Figure 4. Examples of uncorrelated evaluation items.

The sample evaluation items in Figure 4 are not representative of classroom teaching behavior. They ask students to assess situations which are not relevant with teaching. The location of the classroom, whether or not students take notes, or the comfort of the classroom has no bearing on the quality of teaching. One problem with evaluation instruments concerns whether the instruments measure what they are supposed to measure for the purpose they serve. One issue of student evaluation of faculty instruction, therefore, is whether the items in the evaluation instruments really characterize effective teaching performance (Whitman and Weiss, 1982). Validity, when applied to instruments used to assess teaching performance, is defined as the degree to which the instrument measures what it is supposed to measure (Borg and Gall, 1989). Therefore, the question asked is do the evaluation questions posed in the instruments characterize observable classroom teaching behavior? It is obvious that

the samples above are not characteristic of classroom teaching behavior.

Of 4,028 items, 987 (24.5%) evaluation questions did not correlate with classroom teaching performance. Table 4 is a summary of evaluation questions that do not characterize classroom teaching performance.

TABLE 4

PERCENT OF EVALUATION ITEMS NOT CORRELATED WITH CLASSROOM TEACHING BEHAVIOR
(n = 4,028)

Type of Item Flaw (Criteria)	Number of Items	Percent of n
No correlation items.	987	24.5

n = total number of evaluation items

Evaluation Responses that are Flawed

Flawed responses to evaluation questions are those which are skewed, ambiguous, unclear or do not correspond with the evaluation question. A skewed response offers either more positive than negative response options or more negative than positive response options.

Positive Evaluation Response

In a positive response, the response options lean more towards a positive evaluation. In Figure 5, the evaluation item offers four positive response options while it offers only one negative choice. The evaluation will likely be biased towards a positive rating.

"My interests were broadened to"

A. the very highest degree
B. a very high degree
C. a high degree
D. an average degree
E. a poor degree or not at all

Figure 5. Example of a positively skewed response option, most likely rating will be more positive than negative.

Negative Evaluation Response

The example in Figure 6 resembles the positive response options except that the response choices lean towards a negative evaluation. The evaluation item offers three negative response options to only one positive choice. The evaluation will likely be biased towards a negative rating

"To what extent did the instructor use examples to help clarify the material?"

A. Frequently
B. Somewhat slow
C. Seldom
D. Never

Figure 6. Example of a negatively skewed response option, most likely rating will be more negative than positive.

Ambiguous Evaluation Response

An ambiguous response offers more than one response option that can be classified in two or more categories at the same time. The response is stated in such a way that it has more than one meaning in the response options. The response is confusing and unclear as to which response option the evaluation item refers. In Figure 7, the

response options to the evaluation item are confusing because they ask for more than one area of assessment simultaneously, such as, the course being "too difficult" along with "too elementary" and the course being "merely repetitious" along with "too difficult." The response options refer to two things at the same time and therefore can be interpreted in more than one way.

"The level at which course is taught."

9 8 7	6 5 4 3	2 1 0
Level of material right for this particular course.	Course too difficult much of the time. Course too elementary much of the time.	Course merely repetitious of previous courses. Course far too difficult.

Figure 7. Example of ambiguous response options ask to respond to two different values at the same time.

Unclear Evaluation Response

An unclear evaluation response to an evaluation item is one whose response options are vague, uncertain, indistinct or confusing. The response is not clearly defined. The response values 2, 3 and 4 in Figure 8 are not defined and are therefore confusing. Also, there should not be more than two responses (yes or no) in a dichotomous choice.

"In your opinion, did your professor/instructor make good use of class time?"

1	2	3	4	5
Yes				No

Figure 8. Example of unclear response options. The value between the "Yes" and "No" response choices are not clear.

Evaluation Response Not Correspondent to the Evaluation Item

The evaluation response that does not correspond with the evaluation question is one that has no correlation to the evaluation statement. These types of response options seem to relate to some other type of evaluation question. In Figure 9, the response choices to the example "The instructor's attention to starting and stopping class on time..." are "outstanding," "above average," "average," "below average," or "poor." These response choices do not seem to be compatible with what the statement is asking.

"The instructor's attention to starting and stopping class on time was"

1. Outstanding	4. Below Average
2. Above Average	5. Poor
3. Average	6. Not Applicable

Figure 9. Example of response options that do not correlate with the evaluation item

Of the 200 evaluation instruments analyzed, 58% contained responses that were skewed, ambiguous, unclear or did not correspond to the evaluation question. Table 5 summarizes the number of instruments that contained these type of responses.

TABLE 5

PERCENT OF EVALUATION INSTRUMENTS CONTAINING FLAWED RESPONSES TO EVALUATION ITEMS (N = 200)

Type of Response Flaw Criteria	Number of Instruments	Percent
Positively skewed.	26	13.0
Negatively skewed.	4	2.0
Ambiguous.	35	17.5
Unclear.	43	21.5
Do not correspond with item. .	8	4.0
Total.	116	58.0

N = total number of evaluation instruments

A large number of evaluation instruments in the sample analyzed contained flawed questions and responses. Table 6 summarizes the three areas investigated in this study: flawed evaluation questions, such as, ambiguous questions, unclear or vague questions, and subjective questions; flawed responses to evaluation questions that are ambiguous, unclear, skewed, or do not correspond to the question; and evaluation questions that do not correlate with classroom teaching performance.

TABLE 6

SUMMARY OF FLAWED ITEMS AND RESPONSES IN EVALUATION INSTRUMENTS (N = 200; n = 4,028)

Criteria	(N) No. of Instr.	(N) % of N	(n) No. of Items	(n) % of n	Ave. Per Instr.
Ambiguous items. . . .	185	92.5	910	22.6	4.9
Unclear items.	180	90.0	849	21.1	4.7
Subjective items . . .	152	76.0	440	10.9	2.9
No correlation items .	190	95.0	987	24.5	5.2
Flaws in responses . .	116	58.0	-	-	-
Total.	-	-	3,186	79.1	17.7

N = total number of instruments
n = total number of evaluation questions

Table 6 shows that more than 90% of the evaluation instruments contained evaluation questions that were ambiguous and unclear or vague in content. There were 76% of the instruments that contained subjectively stated questions. More than 90% of the instruments contained evaluation questions that did not correlate with classroom teaching behavior. Almost 60% of the evaluation instruments contained responses to evaluation questions that were either skewed, ambiguous, unclear or did not correspond to the question asked.

The total number of evaluation questions that contained these flaws seem quite high (79.1%). The average number of evaluation questions per instrument, excluding comments and open-ended questions, was 20.14. The average number of flawed evaluation questions contained in each instrument was 17.70 as shown in Table 6.

Summary

The content of evaluation instruments used by schools of education accredited by the National Council for Accreditation of Teacher Education (NCATE) were analyzed. The criteria derived for this study emerged from a sample analysis of evaluation instruments. The procedure followed in defining the criteria was to first analyze the content of randomly selected instruments. From this preliminary investigation, the criteria used in this study were obtained.

The content analysis of evaluation instruments focused on three major criteria: 1. The frequency distribution of evaluation questions that were stated ambiguously, unclearly, and subjectively. 2. The frequency-count recording of evaluation instruments containing responses to evaluation questions that were skewed, ambiguous, unclear or did not correspond to the evaluation questions. 3. The frequency distribution of evaluation questions that did not correspond to classroom teaching performance.

Content analysis of 4,028 evaluation questions contained in the 200 evaluation instruments analyzed revealed 54.6% of the questions were ambiguous, unclear or subjective in content. Another 24.5% of the questions did not correlate with classroom teaching performance. A total of 79.1% of the questions were either flawed or did not identify with teaching behavior. This study also revealed that 58% of evaluation instruments contained responses to evaluation questions that were ambiguous, skewed, unclear or did not correspond with the question. Of the 200

evaluation instruments analyzed, an average of 164.6 (82.3%) of the instruments contained one or more types of flaws mentioned in this study.

Implications

No literature indicates that any particular evaluation instrument is universally accepted by colleges and universities. Furthermore, it has not been demonstrated that evaluation instruments were tested for validity, although there are thousands of instruments being used today. This study identified the nature and extent of the flaws in evaluation instruments used in the population sample. Institutions need to focus on these errors and systematically institute corrective measures.

The reliability of an evaluation instrument is principally concerned with random errors in the data. One such error comes from poorly phrased evaluation questions (Doyle, 1983). Reliability is the measure that provides consistent and stable indications of the characteristics being investigated (Anderson et al. 1975). Validity is the extent to which the measures correspond to the characteristics under investigation (Dressel, 1978).

Validity, when applied to instruments used to evaluate college teaching, is defined as the degree to which the instrument measures what it is designed to measure. There is more than one kind of validity, however, which this definition does not take into consideration. Therefore, the question asked should not be "Is this a valid evaluation instrument?" but "Is this evaluation

instrument valid for the purpose to which it is intended?" (Borg and Gall, 1989)

This definition when applied to student evaluation of teaching is very important because without standards for validity, evaluation of teaching instruments can be misused and can have a deleterious effect on the faculty being evaluated (Borg and Gall, 1989). As an example, if educators unscrupulously develop evaluation instruments without the benefit of supporting evidence that the instrument assesses classroom teaching effectiveness and take the scores from these evaluations at face value, the results could hurt the faculty being evaluated when these results are used to make personnel decisions.

Validity, then, represents not only the degree to which the instrument measures what it intends to measure, it is also the "extent to which student rating forms serve their purpose" (Miller, 1987). In other words, do they measure what they are supposed to measure for the purpose for which they are intended. Valid evaluations take all relevant variables into account and judge them objectively (Miller, 1987). The questions asked concerning validity of evaluation instruments would be "What does this item of information signify? What implications, or indirect meaning, does it carry? What information should an evaluation include?" (Doyle, 1983).

Content validation has to do with the "attribution of meaning through expert judgment about the importance of the questions to be asked" (Doyle, 1983). Content validity involves someone's

inspection of the items and deciding whether they are sufficiently consonant with the content. There is obviously a heavy reliance on human judgment in using this approach, and it is sometimes referred to as a "judgmentally oriented approach" (Popham, 1975). Content validation, therefore, relies on the selection of experts on whose judgment the estimates of item importance rest (Doyle, 1983). The evaluation system should emphasize effective evaluators, that evaluators are technically knowledgeable, and that they are well trained in conducting evaluations (Wells, 1982). Students are considered effective evaluators on the premise that because they spend many hours with the teacher, they know better than anyone else how the course and professor's teaching characteristics affect them (Doyle, 1983).

Student evaluation of teachers is one of several means for evaluating college teaching. This method reflects the level of student satisfaction with a professor's teaching performance in the classroom, and is a widely used method. Many faculty, however, believe this method is widely misused and is even threatening to some faculty (Gage, 1974). Some of the arguments concerning faculty evaluation instruments are that they are incomplete or improperly phrased and may be subject to the student's interpretation (Morton, 1964). This study supports Morton's findings.

Students are an important factor in evaluating teaching, because their ratings have certain advantages in that they are both logical and empirical. No one sees, hears, reads, or experiences

the professor's work as fully, directly, and personally as the students in the classroom (Gage, 1974). Students are in the best position to judge a professor's work and compare them with others with whom they have taken courses (Miller, 1975). Their importance, however, "depends in large measure on whether they are asked the right questions" on the evaluation instrument (Seldin, 1980). Although research of student evaluation is extensive, particularly of various rating forms being used today, very few conclusions can be drawn regarding the impact of student evaluation of faculty behavior (Smith, 1976). There is no study that could be found that analyzes the content of evaluation instruments used by colleges and universities.

The literature on the reliability of student evaluations of classroom teaching is extensive and many times conflicting. It can be cautiously generalized, however, that students can provide acceptably valid and reliable opinions on good teaching if asked questions on the evaluation instrument that are on subjects within their experience and scope that are clearly and concisely written, and that refer to commonly accepted aspects of good teaching (Miller, 1987). This study raises serious doubts about how clearly written current evaluation instruments are as well as how well they refer to "accepted aspects of good teaching."

Educational Importance of the Study

The arguments presented by both opponents and proponents on the issue of faculty evaluation are hard to deny. They both seem to have merit. But there is no denying also that higher education

needs to make further progress in studying the details necessary for describing competent teaching. This factor becomes especially crucial when faculty evaluation is used for personnel decisions.

How institutions of higher education appraise faculty members' teaching performance has emerged as a sensitive and important subject since a professional career and personal well-being may depend on it. Colleges and universities should search for solutions to this pressing problem. In doing so, institutions should address the controversy regarding decisions made for tenure, promotion, and retention of faculty which continues to plague many colleges and universities (Seldin, 1985).

No matter how opponents and proponents view teaching evaluation, numerous studies have indicated that more and more faculty members and deans favor student evaluation of teaching. One study showed that 72% of professors surveyed said they favored a formal procedure to evaluate teaching. From the same group, 82% also felt that students should be involved. In that same study, 85% endorsed a formal program of faculty evaluation be used in making decisions about such matters as salary, promotion, and tenure (Gaff, Wilson & others, 1970).

A nationwide survey of 616 private and public institutions showed that 98.8% of private college academic deans and 99.0% of public college academic deans indicated that classroom teaching performance and effectiveness were the most important factors in faculty evaluation. In that same study, students and faculty

members both agreed that student input should be included in the evaluation of faculty performance (Seldin, 1985).

A national survey conducted by the American Council on Education (ACE) showed nearly 70% of the faculty members agreed that there should be a formal student evaluation of their teaching. In another study by the ACE, 72% of responding college freshmen felt they should evaluate faculty performance (Centra, 1982).

Research evidence indicate that students can make valid and reliable judgments about classroom teaching performance if asked the right questions. The argument is that students are professional teacher-observers by the time they reach college, having observed teachers since preschool. The argument goes further in that if students are asked relevant questions that are within their experiential background, they can make fair and sound judgments about teaching (Miller, 1987). In order to make relevant observations and to interpret those observations in a valid manner, however, students need to be using valid, objective instruments and the evaluation process needs to be completed systematically.

Although some studies conclude that good teaching is not validly measured by student evaluations in their present form, a substantial majority of studies have shown evidence that students can evaluate fairly and perceptively (Miller, 1974). However, the existence of adversaries on the issues of faculty evaluation and the continued controversy surrounding students' involvement in faculty evaluation are strong indicators that problems still exist that need to be further studied.

As long as these problems regarding teaching evaluation persist, continued studies of faculty evaluation are needed. When there are "professors" in higher education that are ineffective in the classroom; when evaluation is considered for personnel decisions; when student learning is at stake; when taxpayers support the institutions; when the governing body of a university desires to instill excellence in the institution; when government and the community demand accountability from faculty members and collective concerns; when institutions seek contributions and support from alumni, private agencies, and the community in general; when institutions seek grants from the government; and when questionable evaluation instruments are still being used to measure teaching performance; further research to evaluate the effectiveness of classroom teaching is necessary.

There is no denying that in order to foster a quality institution, much depends on the administration, the governing body of the institution, excellence within the faculty ranks, and to some extent the student themselves. One means of attaining institutional quality would be for these constituents to actively refine the evaluation process and carefully scrutinize existing evaluation instruments by which to assess classroom teaching performance.

Although much has been written about teaching evaluation, the literature search for this study did not disclose any content analysis of evaluation instruments used in colleges and universities. This study revealed major problems in the

instruments used by the population sample. This study also described the extent of those problems.

Educators are held responsible for providing questions within the scope of students' experience and knowledge. Students should then be able to correlate their professor's teaching behavior with characteristics stated in the evaluation instrument and respond appropriately. Effective as well as ineffective teaching can be observed, studied, discussed, analyzed, and assessed. But the quality in an evaluation instrument depends on its validity and reliability. This study identified the nature and extent of the problems in evaluation instruments. Institutions need to focus on these problems and systematically institute corrective measures. Students are not responsible for how evaluation instruments are presented nor are they responsible for administering the evaluation process. If educators ask students to evaluate their teaching performance using flawed instruments and students misconstrue what the evaluation question is asking, educators should at least be partially blamed for abuse of student evaluation in higher education.

Conclusion

Based on the findings of this study, evaluation instruments used in their present form need serious investigation. This analysis of evaluation instruments revealed major problems in the content of the instruments analyzed. This study also identified the nature and extent of flaws in the evaluation instruments.

Institutions need to focus on those errors and systematically institute corrective measures.

The development of an evaluation instrument is a long process that needs to be carefully formulated. Developing a valid instrument involves three things: describing or defining the characteristics of effective teaching, phrasing the questions applying to these characteristics, and selecting the most appropriate responses to each question. The instrument, therefore, should be developed around the experience and scope of the students, minimizing as much as possible the chance of subjectively worded items.

Much research on student evaluation of teaching performance conclude that evaluation from responsible students achieve a very high degree of reliability. Studies done by McKeachie (1979), Aleamoni (1980), Cohen (1980), Marsh (1980), and Millman (1981), on the reliability and validity of student evaluation support its usefulness as a measure of instructional effectiveness. Other studies by Kulik and Kulik (1974), Centra (1979), and Aleamoni and Hexner (1980) also support the reliability and validity of student evaluation (Stevens, 1987).

Based on the definition of validity, reliability and error analysis, implications concerning this study can be drawn. The high percentage of flawed evaluation questions and responses and evaluation questions that are not relevant to teaching behavior indicate major flaws in the content of the 200 evaluation instruments analyzed. The findings in this study suggest there are

serious questions about the validity of the instruments used by colleges and universities to assess classroom teaching performance.

The whole process of devising an evaluation instrument and testing them for validity and reliability is time-consuming that is often rushed for action by the administration, resulting in the whole process not being guided by tested principles of operation. The end result is using flawed instruments to evaluate classroom teaching performance.

References

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), Handbook for teacher evaluation. Newbury Park, CA: Sage, 116.

Anderson, S. B., Ball, S., Murphy, R. T. & Associates (1975). Encyclopedia of educational evaluation. San Francisco: Jossey-Bass.

Bailey, K. D. (1982). Methods of social research. New York: Macmillan.

Berelson, B. (1952). Content analysis in communication research. New York: Free Press.

Borg, W. R. & Gall, M. D. (1989). Educational research: An introduction (5th ed.). New York: Longman.

Centra, J. A. (1982). Determining faculty effectiveness. San Francisco: Jossey-Bass, 17.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. Research in higher education 13, 4.

Doyle, K. O. Jr. (1983). Evaluating teaching. Lexington: D. C. Heath.

Dressel, P. L. (1978). Handbook of academic evaluation. San Francisco: Jossey-Bass.

Gaff, J. G., Wilson, R. C. & others. (1970). The teaching environment: A study of optimum working conditions for effective college teaching 63, Berkeley: Center for Research and Development in Higher Education, 29.

Gage, N. L. (1974). Student ratings of college teaching: Their justification and proper use. In N. S. Glasman & B. R. Killait (Eds.), Second UCSB Conference on Effective Teaching, Santa Barbara: University of California at Santa Barbara, 72-86.

Holsti, O. R. (1969). Content analysis for the social sciences and humanities. Reading, MA: Addison-Wesley.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. In W. S. Gehman & G. R. Thomas (Eds.), Educational and psychological measurement: Vol. 30, No. 3. VA: Educational and Psychological Measurement, 608-612.

Krippendorf, K. (1980). Content analysis an introduction to its methodology. Beverly Hills: Sage.

McKeachie, W. J. (1969). Student ratings of faculty. AAUP Bulletin 55, 439-444.

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

Miller, R. I. (1987). Evaluating faculty for promotion and tenure. San Francisco: Jossey-Bass.

Miller, R. I. (1975). Assessing teacher effectiveness. In B. Massey (Ed.), Proceedings of the First International Conference on Improving University Teaching, Heidelberg: University of Maryland, 31-32.

Miller, R. I. (1974). Developing programs for faculty evaluation. San Francisco: Jossey-Bass.

Millman, J. (1981). Handbook of teacher evaluation. Newbury Park, CA: Sage.

Morton, R. K. (1964). Evaluating college teaching. In H. A. Estrain & D. M. Goode (Eds.), College and university teaching. Dubuqe, Iowa: W. C. Brown, 521.

NCATE 35th Annual List of Accredited Programs/Units. (188-89). National council for accreditation of teacher education. Washington, D. C.

Popham, W. J. (1975). Educational Evaluation. Englewood Cliffs, NJ: Prentice-Hall.

Seldin, P. (1985). Changing practices in faculty evaluation. San Francisco: Jossey-Bass.

Seldin, P. (1980). Successful faculty evaluation programs. New York: Coventry.

Smith, A. B. (1976). Faculty development and evaluation in higher education. AAHE-ERIC/Higher Education Research Report No. 8. Washington, D. C.: The American Association for Higher Education, 35-36.

Stevens, J. J. (1987). Using student ratings to improve instruction. New directions for teaching and learning: Techniques for evaluating and improving instruction. San Francisco: Jossey-Bass, 33.

Wells, R. G. (1982). Guidelines for effective and defensible performance appraisal system. Personnel Journal, 61 10, 776-782.

Whitman, N. & Weiss, E. (1982). Faculty evaluation: The use of explicit criteria for promotion, retention, and tenure. AAHE-Eric/Higher Education Research Report (No. 2). Washington, D. C.: American Association of Higher Education, 2.