STUDENT DESCRIPTIONS OF INSTRUCTIONAL CHARACTERISTICS AS RELEVANT INDICES OF TEACHING EFFECTIVENESS

by

Roger Gene Hoffman

Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1976

APPROVAL SHEET

Title of Dissertation: Student Descriptions of Instructional Characteristics as Relevant Indices of Teaching Effectiveness

Name of Candidate: Roger Gene Hoffman
Doctor of Philosophy, 1976

Dissertation and Abstract Approved:

C. J. Bartlett, Ph.D.
Professor and Chairman
Department of Psychology

Date Approved: 7/28/76

ABSTRACT

Title of Dissertation: Student Descriptions of Instructional Characteristics as Relevant Indices of Teaching Effectiveness

Roger Gene Hoffman, Doctor of Philosophy, 1976

Dissertation directed by: C. J. Bartlett, Professor and Chairman, Department of Psychology

The Check-List of Instructional Characteristics (CLIC) is a student response questionnaire designed to provide university faculty with feedback concerning their instruction. The CLIC contains six factor analytically derived scales. Knowledge and Skill, Consideration, and Critical Demands refer to characteristics of the instructor. Coordination refers to the way readings, examinations, and class presentations are related. Student Involvement and Overall Satisfaction refer to student reactions. The purpose of this research was to assess the relevance of these scales to the assessment of teaching effectiveness.

A model was constructed with Course Size, Course Format, Students' Initial Interest in the Course, and reported SAT as antecedents of the CLIC ratings, and Students' Outcome Interest, Self-Reported Learning Progress, and Student Course Performance as outcomes. Students' Satisfaction with the Instructor as a Teacher (Satisfaction/Teacher) and Students' Satisfaction with the Instructor as a Person (Satisfaction/Person) were assessed. Expected Grade, Grade Point Average, and reported SAT were examined as potential contaminants of the CLIC ratings.

Results were analyzed for four groups of classes, 155 assorted Regular classes, 75 Math, 28 Speech, and 10 German classes. The Math,

Speech, and German samples were each composed of sections of the same course, taught by different instructors, but using common student performance measurements.

The following conclusions were drawn:

(1) The CLIC scales, except for Consideration and Critical Demands, are related to Satisfaction/Person as a result of a common association with Satisfaction/Teacher. The relationships between Consideration and Critical Demands with outcome criteria are not contaminated by their relationships with Satisfaction/Person.

(2) Reported SAT, and Grade Point Average are not related to the CLIC ratings. Expected Grade is correlated with the CLIC ratings, however, at least part of this relationship can be attributed to the relationships that expected grade and the CLIC ratings share with students' perceived Learning Progress.

(3) Students' Initial Interest, Course Size, and Format may bias the level of the CLIC ratings, but do not appear to invalidate the ratings.

(4) Coordination and Consideration are the instructor related scales most closely associated with Student Involvement. Furthermore, their relationships are independent of students' Initial Interest. The relationships may hold only in discussion classes.

(5) Knowledge and Skill, and Student Involvement seem to be the scales most highly associated with students' Overall Satisfaction.

(6) Student Involvement, Overall Satisfaction, and Consideration may be the most closely associated with Outcome Interest, after the effects of Initial Interest are removed. Knowledge and Skill, and Critical Demands also seem to have some relevance for this criterion.

(7) In studying student learning as a criterion, the effects of

student ability were statistically or methodologically controlled in each setting. The effects of Initial Interest were partialled out in Regular and Speech classes. None of the CLIC scales were related to the performance criterion in Speech classes. The Speech setting was the only setting in which the classes sampled were discussion sections which shared a common lecture. Instructors in the other samples were fully responsible for course presentations.

Knowledge and Skill, and Overall Satisfaction were related to learning criteria in the Speech, Math, and Regular classes. In addition, Consideration, Coordination, and Student Involvement were related to learning in Math and Regular classes. The sample size restricted any generalizations from the German sample, although the correlations were in the expected directions.

It is concluded that evidence was found to support the CLIC as a relevant criterion for evaluating teaching effectiveness, and that it may be useful as a guide for improving instruction. Overall Satisfaction, Knowledge and Skill, Consideration, and Student Involvement showed highly consistent relationships with the various outcome criteria.

ACKNOWLEDGMENTS

A number of individuals have been involved in this project even before it became a dissertation effort. The following faculty served on a guidance committee during the initial construction of the questionnaire which became the focal point of this project. These included: Dr. Gilbert Austin, Bureau of Educational Research and Field Services; Dr. C. J. Bartlett, Department of Psychology; Dr. Chancey Dayton, Department of Measurement and Statistics; Dr. Jerry DeBarthe, Department of Animal Sciences; Dr. Claude Kacser, Department of Physics and Astronomy; Dr. Terry Kuhn, Department of Music, and Dr. Robert E. Shoenberg, Dean for Undergraduate Studies. Dr. S. Sorenson, Department of Math; Dr. M. Moore, Department of Speech and Dramatic Arts; and Dr. G. Pfister, Department of Germanic and Slavic Languages lent their support by offering their special course settings for use as research samples for this project.

The following faculty served as advisors on my dissertation committee: Dr. C. J. Bartlett, Dr. Robert E. Shoenberg, Dr. Benjamin Schneider, Dr. Irwin L. Goldstein, and Dr. Ellin K. Scholnick. Each of them has offered valuable comments, suggestions and support. Special appreciation must be extended to Dr. Bartlett and Dr. Shoenberg, Dr. Shoenberg for his trust in my technical capabilities and his patience with my use of the written language, and Dr. Bartlett for his efforts in building my technical abilities to the point of being trustworthy and his enthusiastic support of my efforts.

Finally, I would like to thank Peggy, my wife, for her assistance in reading, editing, typing, and retyping and her wisdom in sustaining my

ego all while growing in her own career.

Research funds were provided by the Office of the Dean for Undergraduate Studies, University of Maryland. Computer time was furnished by the Computer Science Center, University of Maryland. The Center for Evaluation and Development in Higher Education, Kansas State University, granted permission to use their ten student learning progress items.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

STUDENT EVALUATIONS OF TEACHING EFFECTIVENESS: ISSUES CONCERNING THEIR VALIDITY

Student evaluations of teacher effectiveness have been used for over fifty years. Countless rating forms of one kind or another have been compiled. Numerous articles and books, both empirical and philosophical have been published and work in the area remains active.

There are a number of discernable lines of research. The more traditional lines of investigation seem to include issues related to (1) reliability of the ratings, (2) components or dimensions of teaching and (3) biasing factors such as class size, grade received in the course, sex of the instructor, etc. A number of reviews of this research are available (e.g., Costin, Greenough, & Menges, 1971; Kulik & Kulik, 1974; Wetrogen, 1970). Based on these reviews, the following conclusions seem permissible:

Reliability. Stability and internal consistency reliabilities tend to be satisfactory (Smock, 1974; Costin et al., 1971).

Dimensionality. Kulik and Kulik (1974) report that four basic dimensions tend to appear. These are Skill of presentation, instructor's Rapport with the students, amount of Structure in the course, and course work load or Difficulty. The Skill dimension tends to be the strongest and is composed of items related to clarity of presentation, stimulation of students and interesting presentation. The dimension is also the one most highly related to global ratings of the instructor or the course as a whole (Kulik & Kulik, 1974).

Bias. The effects of biasing factors such as grade in the course generally are not strong (McKeachie, 1973; Hoyt, 1973). However, results are somewhat mixed. For example, Costin et al. (1971) report a number of studies in which size of the class correlated with ratings, but they also indicated that the relationship is often not found. Furthermore, they explain that size of the class may be differentially related to selected aspects of the class. These results are even more inconclusive considering the number of different rating forms which have been examined. One might conclude that these biasing factors tend to be situationally specific.

Recently, a fourth topic, the issue of the validity of student evaluations, has become a major research focus. That issue has generated a number of questions regarding the appropriateness of evaluations for different purposes, the validity of evaluations as indicators of other criteria that might have more relevance, contaminants of ratings, and alternative types of rating forms. Research regarding these issues is fragmented and not well replicated so that generalizations are not warranted at this time. However this research does indicate the kind of considerations that need to be made when assessing the validity of a student evaluation questionnaire.

The purpose of the present research is to examine the validity of the Check-List of Instructional Characteristics (CLIC), a student response rating instrument developed at the University of Maryland (Hoffman, 1975). In this chapter some of the issues regarding the validation of such a rating instrument will be reviewed. In the following chapter, a model of teacher evaluation is proposed as a guide for validation of the CLIC.

Elementary Concepts of Criterion Development

To provide a framework for the presentation of the research concern-

ing the validity of student evaluations, some elementary concepts of criterion development will be discussed. Criteria are standards used to appraise the success or worth of a program or the performance of a person. There are five basic concepts relevant to appraisal systems. The term ultimate criteria is commonly used to refer to the total set of outcomes determining success. However, it is often difficult to define and measure precisely all of the elements of the ultimate criteria. The actual criterion refers to what is actually measured by the appraisal system. The extent to which the actual criterion represents the ultimate criterion is termed criterion relevance. Goldstein (1974) indicates that relevance is "the fundamental requirement that transcends all other considerations related to criterion development" (p. 53). Relevance is limited by two factors. One is criterion deficiency which refers to the extent to which the actual criterion fails to assess elements of the ultimate criterion. Criterion contamination, on the other hand, refers to the extent to which the actual criterion measures elements that are not considered part of the ultimate criterion. Because there may be uncertainty regarding what constitutes the elements of the ultimate criterion, assessment of the relevance of an actual criterion tends to be a matter of judgment. However, that judgment can be facilitated by examining criterion contamination and deficiency.

Actual criteria can vary in a hierarchical fashion (Seashore & Yuchtman, 1967), with the highest levels being more closely associated with the ultimate criterion. A number of ways of conceptualizing these levels have been presented. For example, with regard to the evaluation of training programs, Kirkpatrick (1959) has identified four levels of criteria. First, there is the reaction of students to the program.

Second, is an assessment of the learning outcomes of the program. The third criterion concerns the actual behavior of the students in the transfer setting. The fourth level is results which refers to outcomes of behavior such as production rates, turnover, etc. Lindbom and Osterberg (1954) presented similar classes of criteria. Their categories were developed specifically for the evaluation of supervisory training and include three levels: (1) supervisors' classroom behaviors, (2) supervisors' on-the-job behaviors, and (3) the behaviors of supervisors' subordinates.

Implied in both of these classification schemes is a chain of relationships between the levels. Using measures of each of the levels, the extent of the relationships can be determined. For example, Lindbom and Osterberg's first level can be regarded as an assessment of supervisors' knowledge. That knowledge can be examined for its relatedness to actual on-the-job behavior. In turn, on-the-job behavior can be examined for its effect on the behavior of subordinates. One could argue that subordinates' behavior is the most relevant criterion for assessing a supervisors' ultimate success. By beginning with subordinates' behavior, one could work back through the chain, examining the strength of each relationship, in order to provide a basis for judging the relevance of the lower level criterion. Thus if supervisors' behaviors were related to subordinates' behaviors but not to knowledge, one might conclude that the particular measure of knowledge being used is not relevant to supervisors' success. However, that conclusion might not be the most appropriate one.

The behavior of subordinates is probably not completely determined by supervisors' behaviors. Therefore, while subordinate behavior may on

the surface be regarded as the most relevant, it may be contaminated by factors other than the supervisor, thereby reducing its relevance for the ultimate criterion of supervisor success. Likewise, supervisor behavior may be constrained by factors outside the control of the supervisor (e.g., Fleishman, Harris, & Burtt, 1955) which would reduce the relevance of that criterion. When contaminating factors operate in this manner the relationship between the various levels of criteria are attenuated. Thus, in order to decide why hypothesized criteria do not relate to each other it is helpful to have some assessment of possible contaminating factors. An examination of possible contaminants can provide some aid in making judgments about the relevance of the criteria at each level.

The ultimate criterion also tends to be multidimensional in the sense that there are distinct aspects of success (Ghiselli, 1956). Likewise actual criteria tend to be multidimensional with the requirement that each dimension be relevant to some aspect of ultimate success. Criterion deficiency can be reduced by generating a set of actual criteria which maximizes the number of aspects of the ultimate criterion which are estimated. A number of authors (e.g., Guion, 1961; Dunnette, 1963) have discussed problems related to multiple criteria.

Another type of criteria, process criteria, has recently received attention (Guba, 1969; Weiss & Rein, 1970; Cronbach, 1975). While the types of criteria previously discussed are concerned with the consequences or outcomes of a program, process criteria are concerned with describing the program itself. That is, process criteria focus on the events that occurred during the program. However, just as the various levels of outcome criteria are related to each other, process and outcome criteria may be related in means-ends fashion. Similarly, process criteria may vary

in their relevance and they may be multidimensional.

The process-outcome distinction has been applied to faculty evaluation systems (Hartley & Hogan, 1972; Pohlman & Beggs, 1974; Hoyt, 1973). Student ratings of instructors' behaviors are regarded as process criteria, i.e., assessments of what occurred during the program of instruction. Assessments of student learning are outcome criteria, i.e., assessments of the results of instruction. If both the process assessments and the outcome assessments are relevant, they should be related to each other.

This same distinction between criteria can also be derived from Lindbom and Osterberg's or from Kirkpatrick's levels of criteria, by substituting instructor for supervisor and student for employee. Thus, the behaviors of the instructor assessed by the process criteria should be related to the behaviors of the students assessed by the outcome criteria.

It should be noted that another type of rating is frequently used in faculty evaluation rating systems. Often students are asked to respond to general items concerning an instructor's overall success (e.g., "I would recommend this course to anyone"). Items such as this do not refer to outcomes of instruction, nor do they refer to specific instructor activities. However, for the following review, these items will be treated like ratings of instructor activities, that is, as process criteria.

Purpose of Evaluation

Relevance is not necessarily the only principle used in selecting the most appropriate criteria. Evaluation programs vary in their purpose with different types of criteria being more appropriate for different purposes. In general, there seem to be two primary purposes for any performance appraisal system (Carroll & Tosi, 1973; Doyle, 1973; Menges, 1973; Meyer, Kay, & French, 1965). These include appraisal to facilitate

the improvement of performance (either of individuals or of programs) and appraisal for administrative or personnel decisions.

Student learning, which may include attitude acquisition, is often judged to be the most relevant criterion for assessing instructional effectiveness (e.g., Hoyt, 1973; McKeachie, Lin, & Mann, 1971; Morse, Burgess, & Smith, 1956). For personnel decisions regarding promotion, tenure, etc., relevant outcome criteria such as student learning may be appropriate (Menges, 1974). However, for improving performance, outcome criteria are not very informative. They may indicate level of success but they are of limited diagnostic value in determining what might have contributed to that success. For performance improvement, process criteria, that is criteria concerning activities or behaviors of the instructor, may be the most appropriate, but it is essential that they also be relevant. The validation of a student rating of instructor behavior is an assessment of the relevance of the rating for assessing instructional effectiveness.

Two primary techniques for assessing the validity of the teaching effectiveness ratings have appeared in the literature. One technique examines ratings for relatedness to criteria which are regarded as more relevant to teaching effectiveness, i.e., outcome criteria. The other technique examines ratings for relationships to potential contaminating variables that are not considered to be part of the ultimate criterion. The next sections will review some of this literature.

Relationships Between Student Ratings of Instruction and Outcome Criteria

As stated earlier, a commonly accepted criterion of teaching effec-

tiveness is student learning. Based on that premise, a number of studies have used end-of-course performance as a relevant criterion for assessing the validity of behavior ratings. Because the feasibility of end-of-course measures as a criterion is limited, another approach has been to use students self-reports of the amount they learned in a course. End-of-course performance has also been criticized as deficient, and therefore another criterion has also been used--students' interest in further courses in the same subject area. Studies which illustrate the use of these criteria will be presented below.

Interest in Further Courses

McKeachie and Solomon (1958) and Tobias and Hanlon (1975) have used students' interest in further courses as a criterion for validating behavior ratings. McKeachie and Solomon express the opinion that perhaps "some of the most important outcomes of education are not measured by the final examinations used as criteria" (p. 379). They argue that "awakened interest" is also a relevant criterion. McKeachie and Solomon used the percentage of students taking additional courses as their indicator of an instructor's ability to awaken students' interest. Participating instructors were then ranked on (1) the percentage of students continuing to advanced courses, and (2) their rating on two general items, one item concerning the instructor and one concerning the course in general. Rank correlations were computed for five semesters. Correlations were significantly greater than zero for only two semesters.

Tobias and Hanlon (1975) argued that factors such as scheduling contingencies, convenience, etc., may contaminate actual number of students taking advanced courses as a criterion of student interest. Therefore, they argued for statements about behavioral intention as a more appropri-

ate criterion. Student ratings on six factors were used in a multiple regression analysis, with students' behavioral intentions as the criterion. Only ratings of instructor's Skill entered the multiple regression equation (r = .72). Little can be concluded from these studies other than the opinion that student interest may be a meaningful criterion and that ratings, at times, might be related to student interest.

Performance Measures as Criteria

A number of studies have used end-of-course performance measures as criteria for validating ratings of instructor effectiveness. Perhaps the most publicized study is that of Rodin and Rodin (1972) which reported a -.75 correlation across classes between mean student ratings on a single general item and mean student performance. This correlation of course indicates that the instructors (graduate teaching assistants in this case) who received the highest ratings taught the classes which showed the lowest performance. A number of authors have criticized Rodin and Rodin's methodology (Gessner, 1973; Frey, 1973; Bryson, 1974). The two major criticisms are that the ratings were made on graduate teaching assistants who had minimal responsibility for course design, and that only a single general rating item was used. However, Rodin, Frey, and Gessner (all 1975) agree that the Rodin and Rodin study does illustrate that high negative correlations can be found. Of course, the argument is over why negative correlations might be found.

Cohen and Berger (1970) expressed the belief:

> that student ratings of an instructor are not unidimensional and simply correlating a global measure of teaching effectiveness with an available, external educational criteria would mask any correspondence that did exist, and would fail to consider the particular components of teacher effectiveness that do relate to student achievement (p. 605).

This premise seems to be salient for most researchers as the majority of recent studies comparing student ratings to outcome criteria have used multidimensional rating systems.

Solomon, Rosenberg, and Bezdek's (1964) work is one of the earlier studies using end-of-course performance measures. They reported correlations across 24 American Government classes between two learning measures (factual gain, and comprehension gain) and student ratings. Significant correlations ($p < .05$) with factual gain (class mean) appeared for overall evaluation of instructor ($r = .46$), for students liking of the instructor as a friend ($r = .40$), and for a factor analytically derived scale (one of eight) titled Obscurity, Vagueness versus Clarity, Expressiveness ($r = -.58$). Comprehension gain score did not correlate significantly with any of the single item ratings, but it did correlate with two of the eight factors. These were Lethargy versus Energy ($r = -.44$) and Dryness versus Flamboyance ($r = -.42$).

Cohen and Berger (1970) reported correlations between five factors of the Michigan State University Student Instructional Rating Report and the course examination used by 25 sections of basic natural science. Significant correlations with mean section examination score were found for ratings of student interest ($r = .39$) and student-faculty interaction ($r = .37$). Instructor involvement, course demands, and course organization did not correlate significantly with mean end-of-course performance.

Gessner (1973) examined the relationship between ratings made by medical students and two criterion measures of their performance. Correlations were computed, not across classes teaching the same subject, but across 20 different areas of instruction. That is, ten instructors were responsible for 23 areas of instruction. Twenty of these areas

were represented on a national standardized examination. Students gave ratings for each of the 20 areas on two items (content/organization, and presentation). These ratings were then correlated with class performance across the 20 areas. Gessner's primary indicator of class performance was the difference between the national percent correct minus the department percent correct on each item averaged across all items in each of the 20 subject areas. Deviation scores were used instead of raw scores to reduce criterion bias that might occur due to difficulty of the subject area and student interest. Correlations with the deviation scores across the 20 subject areas were .77 with content and organization, and .69 with presentation. Correlations between the two ratings and mean raw scores on an examination constructed by the department were also computed. The two correlations in this case were not significant. Gessner (1975) reiterated his stand that the lack of correlation was due to differences between the subject areas in item difficulty which attenuated the validity of the department examination for between subject area comparisons.

Frey (1973) computed correlations between ratings on six factors and residual examination score (exam score minus score predicted by SAT) across eight Introduction to Calculus classes and across five Multidimensional Calculus classes. Frey concluded that teacher's presentation and organization-planning are important aspects of teaching calculus, and that there may be a trade-off relationship operating since the simple combination of teacher's presentation plus work-load correlated .98 and .92 with residualized exam scores for the two classes.

Bryson (1974) had students in 20 sections of college algebra rate instructors on 12 specific and 2 general items. Mean section ratings for the 14 items and for the average of the 12 specific items were correlated

with mean performance on a common final exam. All fifteen correlations were positive; however, only six were significant. Significant correlations were observed for "Demonstrates dynamism and enthusiasm for his subject" (r = .44), "Communicates a genuine desire to teach students" (r = .51), "Would you take a course from this instructor again?" (r = .48), mean of the 12 specific items (r = .48), and the two general items (.55 and .68).

Doyle and Whitely (1974) computed correlations between ratings and performance across 12 classes. Ratings were made on 49 specific items, which yielded 6 factors, and on 6 general items. Examination scores and residual examination scores were the criteria. For the general items, significant correlations were found between General teaching ability and the residual exam score (r = .51) and between Overall teaching effectiveness and residual exam score (r = .49). Correlation between these two items and raw exam score were .35 and .13, respectively, both not significant. To examine the relationship between the 6 factors and the performance criteria, a procedure was used to estimate factor loadings for the criteria. Highest estimated loadings for the residual exam scores were .36 on "motivation of students," and .31 on "exposition skills." Raw exam score loaded .35 on "exposition skill;" its next highest loading was .18 on "motivation of students." The highest loading for either criteria on "attitude toward students," "stimulation of thinking," or "generalization of content" was .10.

McKeachie, Lin, and Mann (1971) presented five studies relating ratings to various performance criteria. Student ratings were made on items from the Michigan questionnaire. In all studies, Skill, Structure, Feedback, and Rapport (Warmth) factors were examined. Overload (Difficulty) and Interaction were also examined in two of the studies. Performance

criteria included tests of knowledge and critical thinking, and essays in introductory psychology; tests of oral expression, grammar and reading in French; and thinking and attitude sophistication in economics. These studies also examined differential relationships with sex of the student as a moderator. Results tended to be very mixed. However, McKeachie et al. pull together the following conclusions (p. 444):

1. In four of the five studies teachers rated high on "Skill" tended to be effective with women students.

2. In all five studies teachers rated high in "Structure" tended to be more effective with women than with men. In fact, on the whole, the more structured instructors tend to be ineffective for male students.

3. Teachers who were high in "Rapport" ("Warmth") tended to be effective on measures of student thinking.

4. Teachers whom students rated as having an impact on beliefs were effective in changing attitudes.

In general, they conclude "that teaching effectiveness is not a unitary concept but one involving a number of complex interactions" (p. 444).

Sullivan and Skanes (1974) also illustrate the complexity of the relationship between ratings of instruction and student performance. Their research was conducted in ten multiple section courses (smallest number of section in a course was six and the largest was 40). Each course had a common exam. Students provided an overall evaluation of the instructor and evaluation of three specific instructor characteristics. The correlations between mean section performance and mean overall ratings were computed for each of the ten courses. Correlations ranged from -.28 to .57. Only one correlation was negative; it was not significantly different from zero. The average correlation was .39. A chi-squared test indicated a significant overall positive effect.

The introductory psychology course consisted of 40 sections. Twenty-

seven of the sections were taught by full-time faculty; the remaining thirteen sections were taught by graduate students as part-time instructors. Correlations between overall rating and student performance were computed separately for these two groups. Correlations were .53 ($p<.01$) for full-time faculty and .01 for graduate students. Following up the hypothesis that teaching experience might account for the observed difference, the sections taught by full-time psychology faculty were divided on the basis of their experience. There was a significant difference between these two groups. For faculty with more than one year experience the correlation between rating and performance was .68 ($p<01$). For faculty in their first year, the correlation was only .13. Sullivan and Skanes suggest that consistency in the style of presentation may be the underlying variable. They also speculate that inexperience of the instructors may have contributed to the negative relationship found by Rodin and Rodin (1972).

To examine the characteristics associated with effective instruction, Sullivan and Skanes (1974) divided instructors from biology, mathematics, and psychology (the courses with the largest number of sections) into four quadrants based on high versus low overall rating and high versus low student performance and examined mean ratings on the specific items. They drew the following conclusions: (1) Attitude toward students (e.g., friendly) showed little relationship to either student performance or overall rating; (2) Clarity of presentation was associated with overall rating but not necessarily performance; (3) Task orientation was related to performance but not overall rating; and (4) There were differences between courses in which items related to student performance. For example, pressure to work hard was related to performance in psychology and biology, whereas supportive orientation was related to performance in mathematics.

By comparing instructors for whom overall rating and student perfor-

mance were consistent (both high or both low) to instructors for whom ratings and performance were discrepant, Sullivan and Skanes were led to these conclusions: (1) The successful (in terms of student learning) and highly rated instructors seemed to combine task orientation with clarity of presentation and enthusiasm; (2) The unsuccessful and low rated instructors tended to exhibit none of these characteristics; (3) The dissonant group of instructors was either high in enthusiasm but not task orientation (high evaluation, low achievement) or low in enthusiasm but high on task orientation (low evaluation, high achievement).

At this point, one might conclude that instruction's task orientation is the attribute most relevant to the ultimate criterion of student learning. However, Sullivan and Skanes took the broader view that taking further courses and thereby attaining a higher level of achievement is also relevant to the ultimate criterion. Thus, instructor's ability to generate students' interest may also be relevant. To explore this possibility, Sullivan and Skanes examined data on psychology students. They found that students in the high task orientation classes were "at least as likely to take subsequent courses and are more likely to do well in those courses than ...if...emphasis had been on arousing interest and enthusiasm in the subject" (p. 589). However, they warned against overgeneralization, indicating that for some courses (e.g., mathematics) arousing interest may be more important. Recall also that Cohen and Berger (1970) found a significant correlation between student interest and achievement in a natural science course.

Any attempt to summarize the above findings is obviously limited by differences in performance measures, rating questionnaires, course settings and the apparent complexity of the relationships between instructor

characteristics as identified by student ratings and student performance. Precise, definitive statements would be tenuous at best. In lieu of such specificity, it may be appropriate to formulate statements concerning the likelihood, in general, of any instructional characteristic being related to student performance. Kulik and Kulik's (1974) four general factors seem an appropriate means of classifying the instructor characteristics represented in the above studies.

Table 1 presents the results of the above studies classified by Kulik and Kulik's (1974) factors. Factors or items representing each of the four general factors were related in some way to either student performance or student interest in one study or another. However, none of the four factors were related to student performance across all studies. The Skill factor was the characteristic most frequently related to student performance. Six of the eight studies reported a significant relationship between Skill related items or factors with performance. However, McKeachie et al. (1971) found that the relationship held only for female students, and that Skill might even be negatively related to performance for male students. Structure type items or factors appeared in five studies. For three, the relationship with performance was positive. But again, McKeachie et al. (1971) found that for males the observed relationship was just as likely to be negative. As for the Rapport and Difficulty dimensions, the studies provide little support for any relatedness to student performance. The Rapport dimension appeared in seven of the eight studies, but was only significantly related to student performance in the Cohen and Berger (1970) study and in one of the McKeachie et al. (1971) studies. Difficulty or work load appeared in only four of the studies. Sullivan and Skanes (1974) and Frey (1973) found this dimension to be related to performance.

Table 1

Relationships Between Student Ratings of Teacher Effectiveness and End-of-Course Performance

Reference	Effectiveness Dimension			
	Skill	Rapport	Structure	Difficulty
Solomon et al. (1964)	Obscurity, Vagueness vs. Clarity, Expressiveness			
Cohen & Berger (1970)		Student-Faculty Interaction		
Gessner (1973)	Presentation		Content/ Organization	
Frey (1973)	Presentation		Organization	Work load
Bryson (1974)				
Doyle & Whitely (1974)	Exposition Skill	Motivation of Students		
McKeachie et al. (1971)	Skill (females)	Warmth	Structure (females)	
Sullivan & Skanes	Clarity of Presentation			Task Orientation

Student Self-Reports of Learning as Criteria

Studies which examine the relationship between mean class ratings and mean class performance require a performance measure that is comparable across a number of classes. Gessner (1973) was able to standardize performance measures across a number of content areas by comparing performance to national norms. However, the rest of the studies reviewed were confined to lower level, multiple section courses using a common examination.

Hoyt (1973b) cites another drawback to the use of end-of-course performance criteria. He argues that student performance is influenced by a number of factors other than the teacher effectiveness. These include "scholastic atptitude, previous achievement in the discipline and supporting disciplines, personal interest in the subject, perceived relevance of the course for student goals, and academic motivation-persistence" (p. 369). While many of the studies reviewed above used residualized performance to control for aptitude (e.g., Doyle & Whitely, 1974) and some used pre-post gain as a control for previous achievement (e.g., Solomon et al., 1964), the remaining factors remain largely unconsidered. Doyle (1973) also asserts that "we can't judge a teacher on the basis of unqualified student performance. We need to attend to complex qualifiers like student ability and motivation" (p. 65). Similarity, Sockloff (1973) argues that the appropriate criterion is not just learning, but learning that is stimulated by the instructor.

An alternative to direct measures of performance is the use of student ratings of their own outcome attainment. A number of questions surround this type of criterion including: (1) Which outcomes are to be rated? (2) How are they to be rated and which of the confounding factors

mentioned in the previous paragraphs might be controlled? (3) Are they valid? and (4) How are they related to students' ratings of instructors?

Learning categories. Several authors have simply asked students to rate their learning progress in general. A number of authors, however, have attempted to make distinctions between outcomes and have had students make ratings on more than one dimension. Solomon, Rosenberg, and Bezdek (1964) had students rate two statements. One concerned the amount of factual information learned and the other asked about the amount of general understanding gained. However, it appears that students did not make this distinction as the correlation between responses to these two items was .89. In a study by Pohlman and Beggs (1974), students rated eleven items representing three categories which were derived from Bloom's taxonomy (Bloom et al., 1956). These were simple cognitive growth, complex cognitive growth, and affective growth. Hoyt (1973) with the assistance of a number of published taxonomies and the faculty at Kansas State University derived a list of eight learning outcomes. Intercorrelations of the eight items indicated some overlap between mean class responses to the eight items (mean correlation computed by this author .55). The eight item list was later revised to ten items (Hoyt, Owens, and Growling, 1973).

Weerts and Whitney (1975) reported the use of twenty-one items related to possible outcomes of instruction. These items had been previously rated by faculty at the University of Iowa as important outcomes which could be rated competently by students. Weerts and Whitney provided no indication of the dimensionality of their items, nor did they indicate how the original set of items had been generated. Many of the items seemed to refer to student interest and motivation, others referred to learning facts and principles, and others referred to problem solving, evaluating

and discovering relationships.

Hartley and Hogan (1972) developed a list of 26 outcome items. Students' responses to these items along with responses to 30 process criteria items were intercorrelated and factor analyzed. By using the five highest loading items on each factor, they identified three factors composed of outcome items--General Cognitive Development, Field-Specific Development, and Relevance. However, examination of the item loadings reveals that four of the five highest loading items on Field-Specific Development and three of the five highest loading items of Relevance loaded higher on General Cognitive Development. Obviously, it would be difficult to develop orthogonal rating scales for these factors. The basis for selecting the 26 items for study was not specified.

Hall (1970) derived outcome criterion scales from factor analysis of 26 items. The source of these items was not indicated. The factors were Global Course Satisfaction, Improved Thinking and Communication Skills, Increased Enthusiasm about Learning, Increased Personal Understanding, Change in Career Plans, and Cognitive Learning. The scales derived from these factors were not independent as the median between scales correlation was .52. Furthermore, two of the items appear on both the Global satisfaction scale and the Enthusiasm scale.

Finally, Frey (1972) reported a student accomplishment factor among his six factors. The specific items were not presented.

A number of questions emerge with regard to these outcome classifications. Are there common categories among the classifications? Do the categories defined by these classifications adequately reflect the set of possible learning outcomes? Can students discriminate among these categories and rate them accordingly?

There seem to be two problems with regard to interpreting the categories of learning outcomes represented in these studies. First, learning theorists are still debating the dimensionality of learning outcomes. Therefore it would be presumptuous to make any definitive statements concerning the adequacy of the classifications presented above. Second, the studies have developed their outcome criteria in different manners. Some (Pohlman & Beggs, 1974; and Hoyt 1970) developed their criteria from theoretical learning taxonomies. Others (Hall, 1970, and Hartley & Hogan, 1972) developed categories via factor analysis; the sources of their items, however, were not specified. Obviously, factor analytic dimensions are limited by the items that are put into the analysis. Apparently none of the studies submitted categories generated by theoretical taxonomies to factor analytic verification. However, the set of studies taken as a whole seem to suggest several broad general categories of learning. Furthermore, it might be the broad general categories that are more easily defined and therefore easier for students to understand and rate.

In all but the Solomon et al. (1964) study, items or subscales referring to affective orientation can be identified. For example, Hall (1970) identified a factor related to personal values, and one related to interest in learning. Many of the Weerts and Whitney items referred to interest and involvement in the course. One difficulty with affective orientation scales as criteria is partialling out the proportion of that orientation which can be attributed to students predispositional interest in the contents of the course from interest in the course which is generated by effective instruction. Recall that McKeachie and Solomon (1958) specified awakened interest as the relevant criterion.

The rest of the items appear to belong to the cognitive domain (Bloom

et al., 1956). Gagne (1970) has presented eight general types of learning. His work was not mentioned in any of the studies presented, and thus seems to provide an independent source for generating broad categories representative of possible learning outcomes. In general, it seems that three categories of cognitive development can be identified.

Representing what might be considered the most complex type of learning is a category that corresponds to Gagne's eighth type of learning, i.e., problem-solving. This category of learning does not refer to problem-solving in the commonly used sense of applying a given set of rules to a particular situation. Gagne's problem-solving refers to the generation of new rules, patterns, or principles by combining previously acquired principles, rules, theories, etc. Weerts and Whitney's (1975) items concerning developing generalizations and learning new ways to evaluate problems seem to be examples of problem solving criterion items. Items referring to creative thinking (e.g., Hoyt et al., 1973) might also represent problem-solving.

Another general category corresponds to Gagne's rule-using, which involves the application of principles or rules. For example, Hoyt et al. (1973) asked students to rate their progress in their ability to apply course material. Pohlman and Beggs' (1974) scale also contained items concerning application of knowledge. One of Hartley and Hogan's (1972) items asked students to rate their ability to use ideas and techniques gained from the course to solve problems.

The third category of development concerns the acquisition of facts and concepts. For example, Hall's (1970) cognitive learning factor contains items about the acquisition of facts and concepts. Hoyt et al. have an item concerning factual knowledge defined as terminology, class-

ification, methods, and trends. This third category is thus a combination of the lower levels of Gagne's hierarchy, which includes classifying, multiple discrimination, and verbal chaining.

For future use of self-reports of outcome attainment scales, it may be useful to develop the scales from such theoretical analysis followed by empirical verification of internal consistency and at least partial independence. It seems unreasonable to expect completely independent scales for two reasons. First, in practice multiple scales from a single questionnaire tend to be related as a result of common method variance. Secondly, from a theoretical standpoint, Hoyt (1973) has argued that progress in one area of learning may facilitate learning in another area. Gagne's (1970) hierarchy is also built on that assumption. Specifically, he suggests that learning at any level cannot occur unless relevant learning at lower levels is first accomplished.

It may also be noted that the proposed outcome categories (one affective and three cognitive) should not be considered as exhaustive of the set of possible outcomes. As Hoyt (1973) notes, outcomes for studio courses such as art, dance and music are not adequately represented. Physical education outcomes also seem neglected. Therefore, it may be worthwhile to spend time with educators in these fields to develop outcome categories for their courses.

Implicit in the above paragraph is the idea that outcome categories may not have equal relevance for all courses. Hoyt (1973) has addressed that possibility. In the Kansas State University rating system, each instructor specified, prior to receiving feedback from student ratings, the importance of each outcome for his/her course. To derive an overall outcome attainment score, outcomes were weighted by their rated importance.

Unimportant outcomes did not enter the assessment of course success.

Rating Scale. It was noted earlier that end-of-course performance measures may be contaminated by a number of factors. Hoyt (1973) suggests that the use of self-reported progress ratings may control a number of these factors. Specifically, he suggests the use of progress ratings made via an intraindividual comparison, that is, ratings by each student of his/her progress made in the referent course compared to his/her progress in other courses. Perhaps intraindividual ratings would control for scholastic aptitude, and academic motivation. However, intraindividual ratings still seem contaminated by previous achievement, personal interest and relevance of the course. It may be possible to phrase self-report items in alternative ways as an attempt to control for some of these additional factors. For example, personal interest and relevance seems to connote individual differences in student goals for a course. That is, students who perceive the material as interesting and relevant to their own work may desire a higher level of achievement. Therefore, asking students to rate the learning in a course compared to their goals at the beginning may provide relevant information. An alternative approach might be to obtain ratings of students' initial interest in taking a course and then, statistically partial initial interest out of intraindividual progress ratings.

Validity of Self-Reported Learning. Arguing that self-report ratings of achievement may control for various criterion contaminates is not the same as arguing that such ratings are in fact valid indicators of learning progress. Self-reports are susceptable to bias and distortion. Indeed, Hall (1970) chooses to regard his self-report criterion as more indicative of student satisfaction than student achievement. Only a few studies have directly addressed the issue of student achievement self-report validity.

Pohlman and Beggs (1974) examined the relationship between student self-ratings and objective test scores. However, they examined within-class correlations using students as the unit of analysis instead of across-class correlations using mean student performance as the unit of analysis. Partial correlations between ratings and post-test with pre-test partialled out were computed for simple cognitive and complex cognitive outcomes and for affective orientation. Correlations were generally low except for the affective domain. McGuigan (1974) also computed within class correlations between self ratings and several ways of computing performance scores (post-test, gain, G Statistic, grade). The only significant correlation was between rating and course grade (.33) for the sophomore course taught by traditional lecture. For freshman and senior Keller Plan courses, correlations were not significant. The meaning of these two studies is somewhat unclear. Specific rating instructions were not presented. If student ratings were made on intraindividual comparisons, the within class level of analysis was inappropriate. If rating instructions implied some absolute scale or required interindividual comparisons, the relevance of these studies as evidence of the validity of intraindividual ratings is uncertain. Also, for the Pohlman and Beggs study it is unclear why pre-test scores were partialled out of both post-test scores and ratings.

Solomon, Rosenberg, and Bezdek (1964) correlated mean student ratings of the amount of factual information learned, and increase in general understanding, with mean objective test gain scores for knowledge and comprehension across 24 classes. Since their results suggested that students did not discriminate between the two types of ratings, the two ratings showed similar correlations with the two parts of the objective test. With

knowledge gain, correlations for ratings of facts learned and understanding increase were .52 and .57, respectively. With comprehension gain, respective correlations were .11 and .21. Clearly neither rating was a valid indicator of mean comprehension gain. However, it is difficult to judge the validity of the ratings as indicators of factual gain. Apparently rating instructions called for some kind of absolute judgment concerning the amount learned. Since there is no commonly accepted absolute scale of knowledge, it is difficult to know what standard students used to rate their learning. Consequently, it is impossible to know which factors influenced self-ratings. A high correlation between self-rating and exam performance would be expected only if the same set of factors were reflected in both measures. If students responded with an intraindividual comparison such that a number of factors were controlled that were not controlled in the examination score, a high correlation may not be expected. Without further information about factors affecting the ratings and exam performance, the obtained correlations at best supply ambiguous information.

Hoyt (1973b) conducted an indirect test of the validity of student ratings of learning progress. He computed the intercorrelations between mean class student ratings on eight learning progress items and course instructor ratings of the importance of these same items. Positive correlations across courses between mean student ratings and instructor ratings for each of the eight items were hypothesized to occur if all of the following conditions were met: (1) teaching was generally effective, (2) instructors gave careful attention to the task of specifying the importance of the items as objectives, and (3) student ratings of learning progress were valid. Across 606 courses, correlations for the eight objectives ranged from .18 to .50 with a mean correlation of .32. In

contrast, the mean off diagonal correlation was .02.

Evidence favoring the validity of student ratings of their own learning outcomes is not strong. However, a number of conceptual and methodological problems have been identified with these studies. Furthermore, Pohlman and Beggs (1974), who found little support for the validity of self-reports with respect to objective measures of cognitive growth, do not suggest that self-reports be abandoned. They do suggest that objective tests be used along with ratings. Therefore, in light of the potential advantages of self-reports of learning as a criteria of mean classroom performance, perhaps self-reports should be further tested.

Self-Reports and Ratings of Instructors. The final question with regard to student ratings of learning concerns their relationship to ratings of instructor characteristics. Hartley and Hogan (1972) and Frey (1973) reported factor analysis results using process and outcome items. Hartley and Hogan's outcome factors showed clear separation from their process factors suggesting little, if any, relationship between the two types of ratings. Frey did not report the item loadings or data on the relationships among the scales developed from his factors. Therefore, the relationships between his outcome scales and his process scales are unknown.

Solomon, Rosenberg, and Bezdek (1962) examined the relationship between their two self-reported learning items with an overall evaluation item and eight rating factors. The correlations between the mean class outcome ratings and mean class overall evaluation were .79 and .85. With regard to the specific factors, mean class outcome ratings correlated significantly (-.82 and -.84) with mean class ratings on one dimension--Obscurity, vagueness versus Clarity, expressiveness. Factual gain score

(from the objective test) also correlated significantly with only that factor which is congruent with the hypothesis that both ratings of learning and the test were measuring a common characteristic.

Jiobu and Pollis (1971) also found that mean class rating of amount learned (intraindividual comparison scale) correlated highly (.84) with mean class overall evaluation rating. They then examined these two ratings as criterion variables in separate multiple regression analyses of ratings on process-type items. They concluded that "while some process variables are more important than others, no single variable has overwhelming importance" (p. 320). The highest zero order correlations with student perceived learning were with interest in the subject matter (.73), speaking ability (.73), ability to explain (.72) and encouragement of student thinking (.70). Although perceived learning and overall evaluation were highly related, they were not related in the same way to the different process items. For example Jiobu and Pollis stated that "apparently organization has little to do with perceived learning, but it figures moderately in overall evaluation" (p. 320).

Hoyt, Owens, and Growling (1973) computed the correlations between 36 process items and 10 outcome items. The correlations were computed separately for small classes (less than 30), medium classes (30-49), and large classes (over 50). For each class size, they found correlations over .60 for some process-outcome pairs, non-significant correlations for others. Furthermore, cursory examination of the relationships by this author suggests that the pattern of the correlations differed by class size. For small and medium size classes, the strongest correlation across all learning progress areas seemed to occur with items related to stimulation of students' interest. As class size increased, items refer-

ring to the instructor's attempts to stimulate students' involvement in classroom activities appeared to become more highly related to most aspects of perceived learning. For large classes, items referring to speaking style and personalism/consideration also showed strong relationships to many aspects of perceived learning. The variability of the correlations within and between class size categories is consistent with the conclusions that (a) students were discriminating between process items, (b) students were discriminating between learning items, (c) process items may differ in their relationships with different outcome items, and (d) relationships between process and outcome items may differ depending on the size of the class.

Doyle and Whitely (1974) also used students' ratings of how much they learned as a criterion for ratings of items on six factors and six general items. Mean class ratings of learning were not significantly different across their sample of 12 classes, therefore they did not compute correlations with mean class rating on the six general items. Correlations between ratings of amount learned and the six general items were computed across students. They were .52 with overall teaching effectiveness, .54 with how motivating the teacher was, .42 with liking for the instructor, .41 with overall course effectiveness, .39 with general teaching ability, and .34 with instructor's attitude about teaching. Extended factor loadings for ratings of how much was learned on the six factors were -.02 for attitude toward students, .16 for generalization of content, .23 for stimulation of thinking, .28 for expositional skills, and .35 for motivation of students.

Although the correlational studies used different measures of process type criteria, there appear to be some similarities in the results. Two aspects of instructor characteristics consistently related to student

ratings of learning. The instructor's ability to present ideas was related to learning in all four studies. This tendency is consistent with the results of many of the studies which used end-of-course performance measures as criteria. It is also consistent with the statement by Kulik and Kulik (1974) that the Skill dimension, which includes presentation, is the dominant characteristic on which instructors are evaluated by students. Obviously, presentation is an important aspect of effective instruction. The other general characteristic that was consistently related to students' self-perceived learning is motivation or stimulation of students' interest in or involvement with the subject. Of the four correlational studies, only Solomon et al. (1964) do not support this conclusion. However, of their eight factors, none seem to be congruent with involvement with the subject, although they did have two factors concerned with classroom participation. Thus, the study is not necessarily contradictory to the others with regard to involvement with the subject.

It is also interesting to note that, with reference to Kulik and Kulik's (1974) four general factors of instructional characteristics, only the Skill dimension is consistently supported as being related to student self-rated learning (see Table 2). This result is congruent with the consistency of the observed relationship between Skill and end-of-course examination performance. The Solomon et al. (1964) and Doyle and Whitely (1974) versions of the Rapport dimension did not relate to self-perceived learning. The Hoyt et al. (1973) version of Rapport related to perceived learning only in large classes. The tendency for there to be no relationship between Rapport and performance is also consistent with studies using examinations as criteria. As for the Structure dimension, Doyle and Whitely (1974)do not appear to have an analogous factor, and in the other studies

Table 2

Relationships Between Student Ratings of Teacher Effectiveness and Self-Reported Learning Progress

	Effectiveness Dimension			
Reference	Skill	Rapport	Structure	Difficulty
Solomon et al. (1964)	Obscurity, Vagueness vs. Clarity, Expressiveness			
Jiobu & Pollis (1971)	Ability to Explain			
Hoyt et al. (1973)	Speaking Ability	Personalism/ Consideration		
Doyle & Whitely (1974)	Exposition Skills			

structure of organization items were not strongly related to perceived learning. This is somewhat in contrast to studies using exam performance which in several cases found a significant relationship between structure ratings and performance. The Difficulty dimension was not represented in any of the four studies.

Other Factors Related to Student Ratings of Instructor Characteristics

Examining the relatedness of student ratings of instructor characteristics to various student outcome criteria provides one means for judging the relevance of students ratings of instructors as a criterion of teaching effectiveness. In addition, it is desirable to examine student ratings for contaminating factors that do not seem relevant to the ultimate criterion of teaching effectiveness. As mentioned at the beginning of this paper, contaminating factors such as size of the class, sex of the instructor, etc. have been reviewed extensively elsewhere. Recently questions about other types of factors that might influence the rating process have been examined.

Doyle and Whitely (1974) used student liking for the instructor as one of their seven general evaluative items, and Solomon et al. (1964) found students' ratings of "liked the instructor as a friend" correlated highly with overall evaluation. However, Grush and Costin (1975) took a different approach. They argued that students could distinguish between an instructor's attractiveness as a person and attractiveness as a teacher, and that while attractiveness as a teacher might be related to ratings of instructional skill, attractiveness as a person should not be related to ratings of skill. They found that attractiveness as a teacher did correlate highly with skill (.92 in one sample and .90 in another), but

attractiveness as a person also correlated significantly with skill (.56 and .71). However, the two attractiveness dimension themselves were related (.64 and .82), and the partial correlation between skill and attractiveness as a person with attractiveness as a teacher partialled out was quite low (-.06 and -.15). The correlation between skill and attractiveness as a teacher with attractiveness as a person partialled out remained high (.88 and .76).

Grush and Costin (1975) also examined skill ratings for bias due to students' personality traits. Students rated personality traits of themselves and their instructors. Of the eight traits assessed (four in one sample and four in the other), seven instructor traits correlated significantly with students' ratings of instructor's skill. On the other hand, none of the student traits correlated with skill, and the difference between the two correlations was significant for six of the traits.

Another issue that has recently received attention concerns the rating task required of the students. For example, Sockloff (1973) makes the distinction between attitude scales which ask for students' subjective reaction to the instructor versus scales which require an objective description. A rating instrument which requires a subjective reaction is often called an evaluative instrument, whereas a rating instrument which requires a description is called a descriptive instrument (e.g., Weerts and Whitney, 1975). The majority of the instruments discussed in this paper tend to be more evaluative than descriptive.

For most rating systems, five tasks can be identified--observation, recall, evaluation, rater's response and user's action. However, the allocation and sequencing of the tasks may vary. An evaluative rating

---

[1]This discussion was derived from Wherry, 1951.

instrument requires raters to observe something, recall what was observed, evaluate according to some standard what was recalled and then respond in terms of their evaluation. With the descriptive rating instrument, the evaluation task is removed from the responsibility of the rater. A descriptive rating instrument requires the rater to observe something, recall what was observed, and respond to the rating instrument with a description of what was recalled. The task of evaluating the data produced by the rating instrument is the responsibility of the user.

Levinthal, Lansky, and Andrews (1971) suggest that evaluative ratings can be interpreted as students' estimates of the discrepancy between their ideal instructor and how they see the instructor actually performing. This conceptualization of evaluation seems analogous to the conceptualization of job satisfaction (e.g., Wanous & Lawler, 1972; Locke, 1969). Thus students' evaluations of their instructors can be viewed as students' satisfaction with their instructors. Although the job satisfaction literature suggests that the simple discrepancy conceptualization may be an oversimplification, satisfaction does seem to stem from the interplay, in some manner, of a standard and the perceived characteristics of the referent person, object or event.

With an evaluative instrument, there is uncertainty about the standard of comparison used by students in the evaluative process. The standard may reflect a student's beliefs about what an effective teacher does, which may or may not be correct, or the standard may reflect what an entertaining instructor does. The uncertainty is compounded by the possibility that individual raters may not be using the same standard. Levinthal, Lansky, and Andrews (1971) suggest that if standards used by students rating the same instructor are different, their responses do not mean the

same thing, and therefore, they cannot be meaningfully combined. Even if all students used the same standard which in fact did reflect effective teaching, responses might remain uncertain since the ratings may not provide any indication as to why students were satisfied or dissatisfied with a particular characteristic. Dissatisfaction could occur because, in the opinion of the students, an instructor exhibited some characteristic too frequently, too infrequently, or at inappropriate occasions.

Descriptive rating scales have been offered as an alternative. However, a descriptive rating instrument does not solve the problem of validity. Responses may still be influenced by evaluative attitudes of the students, and evaluation of the descriptions still requires a rather unguided judgment, albeit the judgment of the user.

Descriptive information may be more flexible than evaluative information in that it allows instructors or other users the option of choosing their own standards. However, at this time, the state of the art provides few empirically verified guidelines for appropriately evaluating descriptive information. Relationships between the process, or evaluations of the process, of instruction and outcomes of instruction are not well established. Descriptive items, just as evaluative items, must refer to characteristics that are related to student learning or other acceptable outcome criteria. Evaluations made without knowledge of such relationships can obviously be in error. Validation of descriptive information is needed to have any hope of reducing errors of evaluation.

CHAPTER II

MODEL OF INSTRUCTIONAL EVALUATION

In the preceeding chapter, a number of variables relevant to student ratings of instruction were identified. This chapter proposes a model of hypothetical relationships between course characteristics, student characteristics, outcomes of instruction, and descriptions made by students on the Check-List of Instructional Characteristics (CLIC).

The CLIC is a factor analytically derived descriptive rating instrument containing 41 items which form six scales. Three scales refer directly to characteristics of the instructor--Knowledge and Skill, Consideration, and Critical Demands. One scale, Coordination, refers to the way readings, examinations and class presentations are used by the instructor. Two scales refer to student reactions. Items on one scale ask students to describe their Involvement in the course. The other scale asks for students' overall impression of the course and can be interpreted as a measure of students' Overall Satisfaction (Hoffman, 1975). The CLIC items grouped under their respective factors appear in Table 3.

Models, Path Analysis, and Multiple Regression: General Discussion

Before depicting a model of teacher evaluation ratings, it seems appropriate to discuss what is expected from such a model and how the model may be validated. Blalock (1964) and others (e.g., Boudon, 1968; Tukey, 1954; Turner & Stevens, 1959; Wright, 1960) have discussed path analysis as an approach to the study of variables related in complex networks of variables. Their discussions seem to provide some insight concerning the most useful way to approach a teacher evaluation model.

Table 3

CLIC Scales

Students' Perception of Instructor's Knowledge and Skill[a]

2. introduced many ideas during each class session.
3. used examples.
12. had an extensive knowledge of the subject.
13. presented an outline of the course.
16. knew the subject matter.
18. presented recent developments in the field.
21. explained how topics in the course were related to each other.

Critical Demands[a]

4. insisted that everything be done his/her way.
5. would "ride" the student who made a mistake.
9. demanded more than I could do.
14. insisted students follow a standard way of doing things in every detail.
19. expressed his/her displeasure with students who made mistakes.
22. was intolerant of students' mistakes.
23. criticized students in front of others.

Consideration[a]

1. expressed his/her appreciation when one of us did a good job.
6. invited criticism of the ideas he/she presented.
7. allowed students to express their problems related to the course.
8. asked students questions.
10. encouraged class discussion.
11. let me know when I had performed well.
15. put suggestions that were made by students into operation.

Coordination of class, readings and exams (Amount of Structure)

24. readings and class presentations were identical.
26. readings were understandable.
27. readings were helpful.
28. reading material could be substituted for the lectures.
31. examination questions could be anticipated.
32. examinations let you know your strengths.
33. examinations tested general ideas.

Student Involvement

34. I had sufficient background for the difficulty level of this course.
35. I was generally prepared for class.
36. I took an active part in class discussions.
38. I could answer other students' questions.

Table 3 (con't)

39. I asked questions in class.
40. I discussed the material outside of class.
41. I wanted to know more about the subject.
42. I was attentive in class.

Overall Satisfaction

17. The instructor was well organized.
20. The instructor was a very thorough lecturer.
29. The lectures were presented in an appealing way.
30. The lectures were easy to become interested in.
37. I enjoyed every minute of the course.

Note. Numbers indicate the order in which the items appear on the CLIC.

[a]These items refer to the instructor.

Path analysis requires the a priori specification of a causal model which describes the temporal order of the variables and the linkages among those variables. Path analysis is a set of procedures for examining the causal influence of a set of exogenous variables on a set of endogenous variables. The exogenous variables are variables which are assumed to be caused by factors which are not included in the model. There is no attempt to explain the variability of exogenous variables. The purpose of the model is to explain the variability of each of the endogenous variables. An endogenous variable is simply a variable whose variability is attributed to the exogenous variables and any endogenous variable that is brought into the model prior to the one being analyzed. In general, path analysis then uses regression techniques to test the adequacy of the specified model. The utility of the path analytic approach depends as much on the adequacy of the model's theoretical foundations as it does upon the strategies used to test the model. Thus, several competing theoretical models typically can be shown to fit a given set of data equally well.

In its most common form, path analysis requires that (1) all relevant variables have been included in the system, (2) the causal flow in the system is in only one direction and the temporal order of the variables is correctly specified, and (3) the relationships are all linear. When these requirements are met, path coefficients, which are regression weights, are interpreted as indicators of the direct effect one variable has on the other with the implication that conclusions about the importance of each variable can be reached via a comparison of these coefficients (e.g., Spaeth, 1975; however see Blalock, 1968; Tukey, 1954, for contrary opinions).

Path analysis is often approached diagrammatically with single headed arrows between variables representing hypothesized causal paths. The linkages between exogenous variables are generally curved lines with arrows at each end signifying that the variables may be related, but that no causal relationships are being hypothesized between those variables. A set of structural equations can be written which represents the diagram. When all of the possible paths in the specified causal direction are represented, the structural equations lead to several sets of normal equations with one set of normal equations for each endogenous variable regressed on all of the preceding variables. The utility of path analysis lies in the possibility of deleting some of the paths on theoretical grounds prior to data analysis. Deleting paths changes somewhat the structural equations although they still lead to normal equations. The hypothesis that the path between two variables is zero is equivalent to the hypothesis that the correlation between those variables can be completely accounted for by other variables (i.e., the partial correlation, with all of the other variables held constant, is zero). The acceptability of the model is determined by (1) solving for the path coefficients that do appear in the structural equations and (2) using these path coefficients to reproduce the correlation matrix. Unfortunately there is no commonly accepted test to determine how well the reproduced correlation matrix matches the original matrix.

Paths may also be deleted by computing path coefficients (i.e., beta weights) for each link and then (1) testing the significance of the coefficients or (2) accepting as meaningful only coefficients that exceed a specified value. This is a more pragmatic approach which has been dubbed "theory trimming" (see Heise, 1969; Kerlinger & Pedhazur, 1973; Land, 1969).

With the latter approach to deleting paths, the reproducibility of the correlation matrix is again the criterion of acceptance (Kerlinger & Pedhazur, 1973).

Returning now to the problem of specifying the relationships between the CLIC scales and instructional outcomes, it is noted that path analysis does not require a causal model (Wright, 1960). It can also be applied to systems of variables for which causal arguments are not necessary. With regard to the CLIC scales, an important question seems to be: "If scores on the six scales of the CLIC are known, can anything be inferred about (1) how interested students are in continuing their involvement with the subject or (2) how much the students learned about the subject?" Questions such as these can also be depicted diagrammatically and examined with a series of multiple regression statements in a manner analogous to the path analysis of causal models.

In the present study no arguments about causal relations will be made. The data, for the most part, consists of responses made by students at a single point in time. Although these responses are made about events or conditions that might logically be ordered temporally, and thus subjected to causal analysis, the responses occur after all of the events have occurred. Thus, there is the possibility of rating errors such as halo and self consistency (e.g., Wherry, 1951), which could lead to erroneous conclusions concerning causality. For example, ratings given to an instructor on the Consideration scale may be biased due to students' perception of how much they learned in the course. In addition to the possibility of rating errors, the causal order of the events being rated may not be as unambiguous as they appear. Thus, Green (1975) and Lowin and Craig (1968) found that low performance on the part of subordinates led

to lower consideration by their supervisors. A similar order may exist with regard to instructor/student relationship. Therefore, given a large positive relationship between Consideration and student learning, a causal conclusion derived from path analysis that student learning is highly influenced by the instructor's consideration behaviors, might be erroneous. However, given the same results, it would not be erroneous to suggest that an instructor's rating on Consideration can be used as an estimate of how much students learned.

Although causality will not be argued, it does seem useful to develop a model of teacher evaluation ratings as a means of investigating the validity of these ratings for estimating instructional effectiveness.

Because the literature concerning teacher evaluation is somewhat inconsistent, it seems premature to make any a priori assumptions that any of the possible direct relationships (i.e., partial correlations) are zero. Therefore, as suggested by Kerlinger and Pedhazur (1973) all relationships will be examined using regression analysis for theory trimming.

Relationships Among the CLIC Scales

The establishment of an overall model describing the hypothesized relationships among all of the variables which will be investigated in this study will begin with an examination of the relationships among the six CLIC scales. Not only will this examination provide some information about the CLIC itself, but it will serve to illustrate how the main analysis will be conducted. The data analyzed here are from 64 classes which used the CLIC in the Spring semester of 1975.

It was noted earlier that two of the CLIC scales can be regarded as descriptions of students' reaction to the course (i.e., Student Involvement and Overall Satisfaction), while four scales were more directly re-

lated to the instructor. Thus, Student Involvement and Overall Satisfaction may be considered outcomes of instruction. Therefore, Student Involvement was regressed on Knowledge and Skill, Critical Demands, Consideration, and Coordination. Overall Satisfaction has been interpreted as a more general scale and may be, in part, attributable to Student Involvement; Overall Satisfaction was regressed on all five other scales. In addition, size of the class and course format (lecture, lecture and discussion, discussion) were also used in the analysis.

Using stepwise multiple regression, Consideration (CON) and then Coordination (COOR) entered the equation for Student Involvement (SI) with significant ($p<.05$) beta weights. Size of the class was the next variable to enter. Its beta weight was also significant at the .05 level. The equation was:

$$Z_{SI} = .43Z_{CON} + .27Z_{COOR} - .21Z_{SZ}$$

The multiple correlation was .71 ($p<.01$) indicating that size of the class, Consideration and Coordination accounted for approximately 50 percent of the variance in Student Involvement.

For Overall Satisfaction, Knowledge and Skill followed by Student Involvement and then Consideration entered the regression equation with significant beta weights. However, the beta weight for Consideration was negative indicating that it was acting as a suppressor variable. Since there was only a one percent increase in the variance accounted for (R^2) when Consideration was added after Knowledge and Skill, and Student Involvement, Consideration was removed from the regression equation. The resulting multiple correlation was .92 ($p<.001$) indicating that Knowledge and Skill (KS) and Student Involvement (SI) accounted for 84 percent of the variance in Overall Satisfaction (OS). The equation was:

$$Z_{OS} = .70Z_{KS} + .36Z_{SI}$$

Since it could be argued that Student Involvement was a result of Satisfaction, regression equations were examined with Overall Satisfaction regressed on the four instruction scales and Student Involvement regressed on the four instructor scales plus Overall Satisfaction. Results were not as interpretable owing to a standardized beta weight for Overall Satisfaction greater than one and a large negative suppressor effect by Knowledge and Skill in the equation for Student Involvement.

From Figure 1, it seems that ratings on Knowledge and Skill have a direct relationship with student satisfaction as does ratings of Student Involvement. However, the relevance of Consideration and Coordination for students' satisfaction seems to depend on their relationship with Student Involvement. Critical Demands does not appear to have any implication for Student Involvement or Overall Satisfaction. This suggests that instructor's Knowledge and Skill and Student Involvement ratings might be a basis for making an inference about Students' Overall Satisfaction.

The Full Model

Figure 2 illustrates the hypothesized relationships among the variables chosen for this study. The literature reviewed in the previous chapter showed that student learning progress (assessed as end of course performance or by means of self-report) and interest at the end of the course have often been used as outcome criteria. Thus, the two boxes at the far right of Figure 2 represent these variables. As depicted in the model, the relationship between students' outcome interest and learning will be investigated with interest positioned as the antecedent variable. Thus, Learning Progress is hypothesized to be a function of Outcome Interest,

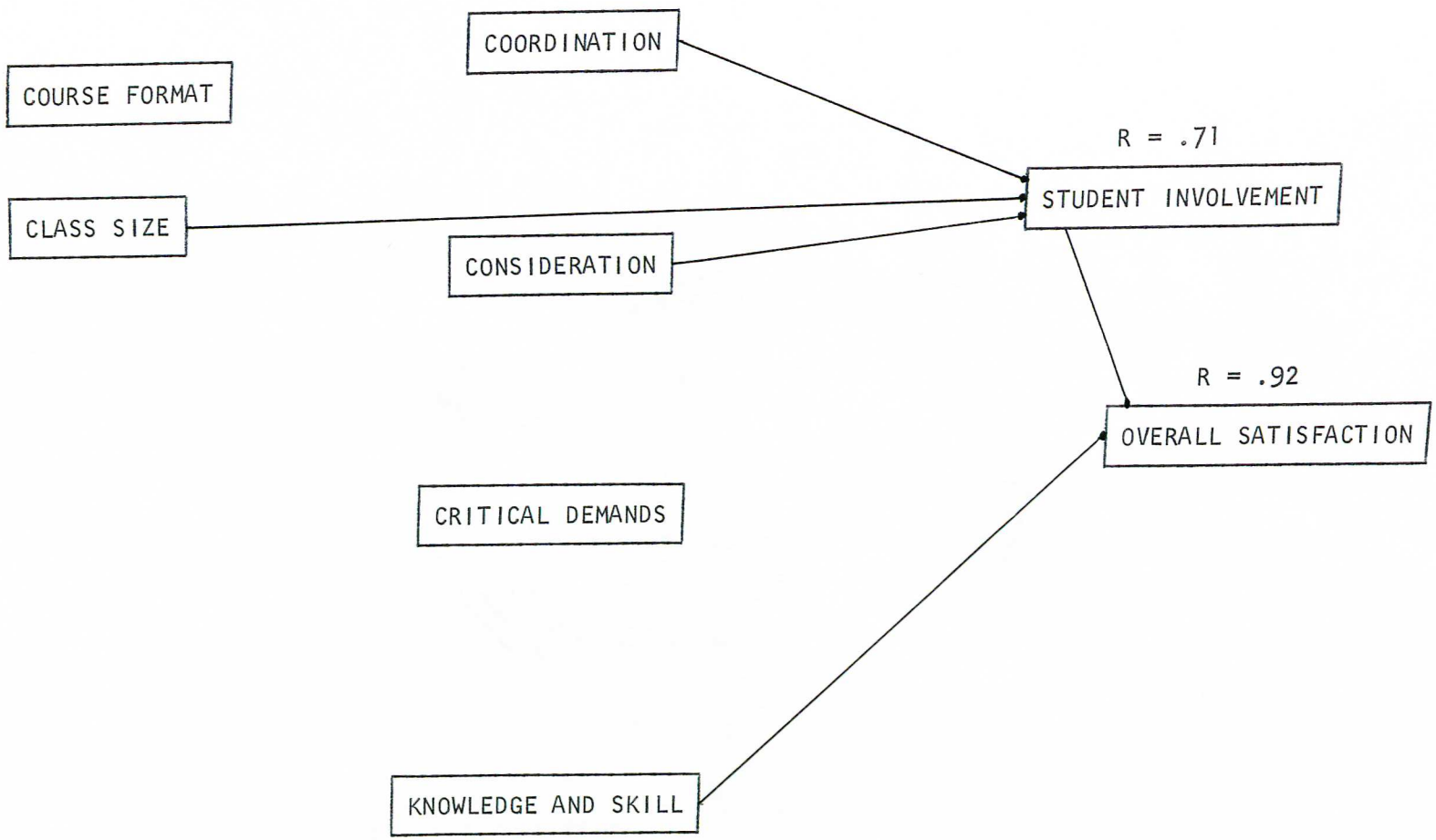

Figure 1. Model of relationships developed on 64 classes.

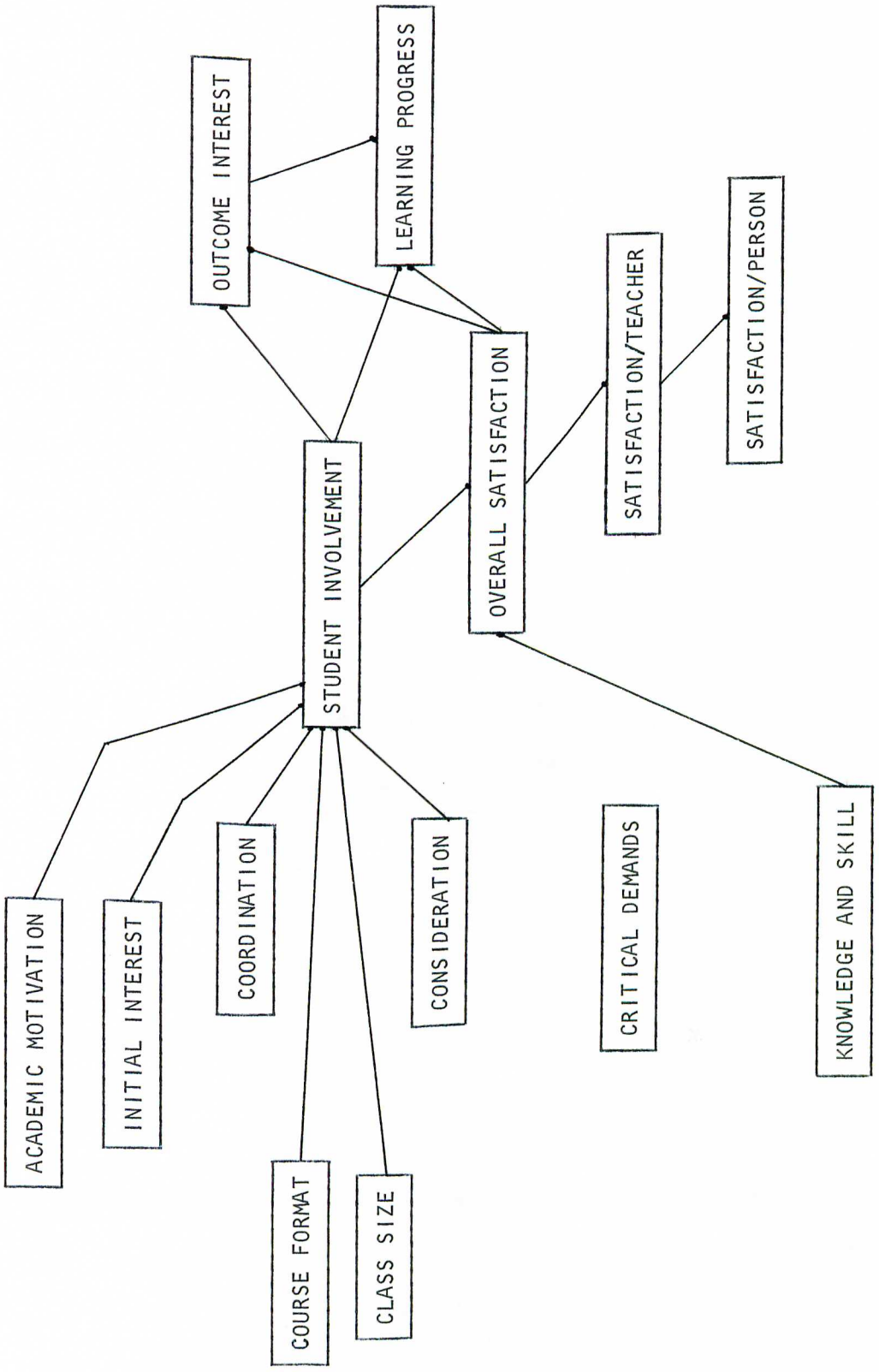

Figure 2. Hypothesized relationships among CLIC and supplementary scales.

Overall Satisfaction and Student Involvement. Note that this model is based on the assumption that students' ability will be methodologically or statistically removed from any estimate of Learning Progress. It is also hypothesized that Student Involvement and Overall Satisfaction will be indicators of Outcome Interest.

Satisfaction with the instructor as a teacher (Satisfaction/Teacher) and satisfaction with the instructor as a person (Satisfaction/Person) also appear in Figure 2. The Grush and Costin (1975) results suggest the hypothesis that Satisfaction/Teacher can account for any relationship between the CLIC scales and Satisfaction/Person, but that the reverse will not hold. That is, with Satisfaction/Person held constant, the CLIC will be related to Satisfaction/Teacher. This suggests that Satisfaction/Person be treated as an outcome of Satisfaction/Teacher.

As for the relationship between the CLIC satisfaction scale and Satisfaction/Teacher, there seems to be little theoretical ground regarding which variable to choose as the antecedent variable. The CLIC items do refer to the instructor so the title, Overall Satisfaction, is somewhat misleading. Both scales seem to refer to satisfaction with the instructor as a teacher. Since the purpose of this research is to shed some light on what inferences might be reached from CLIC ratings, Figure 2 indicates that Satisfaction/Teacher will be investigated as an outcome of Overall Satisfaction.

With regard to the six CLIC scales and size of the class, the relationships in Figure 2 are the same as in Figure 1. Class format has also been retained as a potential determinant of Student Involvement.

The last variables appear at the top of Figure 2. As suggested by Hoyt (1973), students' perceived relevance and interest in the course and

their general academic motivation seem to be important determinants of students' learning. Interest and relevance (labeled Initial Interest) are being combined in this context as a motivational construct, and as such, represents students' reasons for taking a particular course. Figure 2 illustrates the hypothesis that Initial Interest and Academic Motivation are related to Learning and Outcome Interest through their relationship with Student Involvement.

Analysis

Figure 2 indicates that many potential linkages are omitted. This might seem to suggest the use of the path analysis technique of calculating path coefficients for those paths that are represented and then reproducing the correlation matrix. However, traditional path analysis models are built (i.e., paths deleted) on theoretical assumptions about which variables should not be directly related. For traditional path analysis the absence of an arrow indicates the presence of a theory or rationale. In contrast, Figure 2 is built on assumptions about which variables should be related. The absence of an arrow indicates the absence of a theory. For example, Critical Demands is not connected to any other variable because there is no information available concerning its potential relationships with outcome criteria. This model was developed not as a rigid set of hypotheses, but as a heuristic tool for (1) determining the plausibility of the potential relationships between the variables and (2) determining which variables should be treated as antecedent variables for each successive criterion (endogenous) variable.

The analysis of the model (presented in Chapter 4) is an analysis of the amount of variance in each criterion variable that can be accounted for by the best set of antecedent variables. That analysis provides some

suggestion about what can be inferred from the CLIC scales about other aspects of instruction. The "best set" of variables was chosen by a modified stepwise multiple regression procedure: (1) at each step entering the variable which yielded the greatest increase in the multiple correlation, as long as that increase was statistically significant at the .05 level, and (2) at each step removing any variable which could be removed without reducing the multiple correlation significantly (again at $p < .05$). Interaction products were tested after the component variables were already entered as simple additive terms. Therefore, if the F level of an interaction term was sufficient for it to enter the equation and one or both of the additive terms were not in the equation, the additive terms were entered and the contribution of the interaction term was retested.

Sources of Contamination

It was mentioned in Chapter 1 that there are a number of potential sources of contamination and that they tend to be situationally specific. Figure 2 already indicates the expectation that class size and class format will be related to Student Involvement. These two scales along with expected course grade and grade point average are also examined for any significant relationships with the other CLIC factors.

CHAPTER III

METHOD

Participants

A total of 233 classes representing approximately 6000 students participated in this study. The participation of 115 of these classes was the result of the instructors' requests to use the CLIC for the evaluation of their classes. This group, which will be called the Regular classes, included 4084 students. The Regular classes represented fifteen academic departments. Physical Education, Philosophy, and Psychology classes were the most numerous. The remaining classes were predominantly from the Humanities and Behavioral Sciences areas. Due to the small number of courses, no attempt was made to do separate analyses by academic departments.

The participation of the remaining 118 classes was solicited after they were identified as sections from multiple section courses which used the same student performance criteria across all sections. Thus 75 sections of Introduction to Mathematics (1213 students), 33 discussion sections of Basic Principles of Speech (456 students) and ten sections of Elementary German I (108 students) participated. Questionnaires were not received from seven Speech sections and one German section which had been asked to participate. In addition, two speech instructors turned in questionnaires for more than one section in one package making it possible to identify only 28 individual sections.

Table 4 presents descriptive information about the Math, Speech and German samples.

Table 4

Description of Math, Speech and German Samples

Course Name	Number of Participating Sections	Number of Participating Instructors	Number of Instructors Teaching more than one section	Descriptions of Instructors	Course Schedule	Questionnaire Contents
Introduction to Mathematics	75	39	35	Graduate Teaching Assistants and Instructors	Three times a week, all with the same instructor	CLIC scales and SAT Math
Basic Principles of Speech	28	17	14	Graduate Teaching Assistants	Common lecture once a week. Discussion section twice a week	CLIC scales, supplementary scales, expected grade, GPA, SAT Math and SAT Verbal, Sex
Elementary German I	10	7	3	Instructors, and Professors	Four times a week, all with the same instructor	CLIC scales, and SAT Verbal

Student Performance Criteria

Math. There were five examinations, four hourly exams and a final, in the Math sections. For each of the seventy-five sections, the mean performance across students was calculated for each of the five exams. The section performance measure was the sum of these five means.

German. There were three examinations, two hourly exams and a final exam. Students who received a score of 90 percent or better on each of the first two examinations were excused from taking the final examination, and given an "A" for the course. Students who received a 90 percent or greater on the final examination were also given an "A" for the course. Therefore as the criterion of section performance, average final grade was selected.

Speech. Students in Speech received performance evaluations on speeches, papers and examinations. The course coordinator reported that the discussion section instructors should have had an impact on the performance of their students in each of these three areas. Ten percent of each student's grade was subjective points assigned by the instructor. Mean section performance was calculated by averaging the total points received by each student and then subtracting the average "instructor points."

For Math, German and Speech it was reported that standard criteria (e.g., scoring keys) were established for all grading procedures.

Instruments

Participating instructors for the Regular classes completed a short questionnaire describing their classes. Information on class size and course format was extracted from that questionnaire. Course format was coded dichotomously with the class identified as either (1) primarily

lecture or (2) primarily small group/discussion. Class size was coded 1 to 5 in the following manner:

1. less than 10 students
2. 10 - 30
3. 31 - 60
4. 61 - 120
5. over 120 students

Note that the variable Class size is not a linear transformation of actual class size.

The questionnaire completed by the students appears in the Appendix. The first page is the CLIC. For a more complete description see Hoffman, (1975). The second page contains the additional items used in this study.

Initial Interest. Items 44 through 48 were averaged as a measure of each student's Initial Interest in the course. Item 49 was not included, as preliminary analysis on a two-thirds subsample (2819 students) of the Speech and Regular students indicated that this item was not highly correlated with the other items and did not increase the internal consistency reliability of the scale.

Outcome Interest. Items 50 through 53 were averaged as a measure of students' interest in the content of the course at the end of the semester.

Learning Progress. Self-reported learning progress was assessed using the ten Kansas State University items (items 54-63) plus two additional items. On the two-thirds subsample of the Regular and Speech students, these twelve items were intercorrelated and factor analyzed. The first principal component accounted for 50 percent of the total variance. The second component increased the variance accounted for by only six percent. Table 5 presents factors loading for these two principal components and the (Varimax) rotated solution. Based on this evidence, it was concluded that these twelve items could be adequately represented

Table 5

Principal Components Factor Matrix and Rotated Factor Matrix for Twelve Learning Progress Items for the Two Factor Solution

	Principal Components		Rotated Factors	
Items	I	II	I	II
54	.66	.43	.18	.77
55	.69	.41	.22	.77
56	.72	.07	.48	.54
57	.74	.15	.43	.62
58	.62	.18	.32	.56
59	.74	-.18	.65	.38
60	.74	-.18	.66	.37
61	.64	-.27	.65	.24
62	.57	-.37	.67	.13
63	.74	-.25	.70	.32
64	.79	-.06	.62	.50
65	.80	.06	.54	.59

by a single scale. Therefore mean rating of these twelve items was used as a single estimate of students' Learning Progress.

Learning Progress, Initial Interest and Outcome Interest were all assessed on the same rating scale which asked students to describe the items in terms of other courses they had taken at the University of Maryland.

Satisfaction with the instructor. The mean of items 66-68 was used to assess Satisfaction/Teacher and the mean of item 69-71 was used to assess Satisfaction/Person. These items come from Grush and Costin (1975).

Expected grade, GPA, SAT Math, SAT Verbal, and Sex were assessed by items 72-78.

Academic Motivation. Following the suggestion that GPA is a function of motivation and ability (cf. Mitchell & Nebeker, 1973), Academic Motivation was defined as the residual of GPA (item 73) after the combined linear effects of SAT Math (item 74) and SAT Verbal (item 75) were substracted out. Beta weights were computed for the two SAT scores using the two-third subsample of Regular and Speech students. Unstandardized beta weights were .20 for SAT Math and .23 for SAT Verbal. The multiple R was .37. These weights were cross-validated on the remaining 410 students. The resulting R was .36. Beta weights were then computed on the total sample of students (.21 for each variable) and these weights used to calculate Academic Motivation.

CHAPTER IV

RESULTS

Means, Standard Deviations and Reliabilities

CLIC and supplementary scale scores were calculated for each student and these scores averaged across students within a class to provide the class variables used in this analysis. Table 6 presents the means and standard deviations of these class variables.

Interrater reliabilities of the questionnaire measures were estimated by subjecting students' responses within each class to an odd-even split. Split-half reliability estimates using the Spearman-Brown formula were calculated separately for the Regular, Math, Speech and German classes. For Speech and for German the reliability estimates of the performance criteria were calculated from by an odd-even split of students' performance scores. For Math, an internal consistency reliability was estimated using the average correlation between the five tests. Table 7 presents these reliability estimates.

Speech, Math and German classes had fewer raters than Regular classes which would suggest that reliabilities for these classes would be lower than for Regular classes. However for some of the scales the drop was much more than would have been expected due to the reduction in the number of raters alone (e.g., Student Involvement, Coordination, Overall Satisfaction).

Expected grade, grade point average (GPA), the two SAT scores, and Academic Motivation were also treated as class variables although they seem to have less association with the class per se than do the other

Table 6

Means and Standard Deviations of CLIC and Supplementary Scales

Scales	Classes							
	Regular (N=115)		Speech (N=28)		Math (N=75)		German (N=10)	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
Knowledge and Skill	4.31	.34	3.83	.33	3.88	.33	4.07	.18
Consideration	3.68	.56	4.06	.35	3.54	.48	4.13	.20
Critical Demands	1.97	.41	2.17	.42	1.96	.32	1.82	.27
Student Involvement	3.56	.35	3.64	.16	3.49	.18	3.69	.18
Coordination	3.23	.40	2.85	.23	3.15	.22	3.53	.19
Overall Satisfaction	3.64	.51	3.29	.31	3.25	.46	3.81	.22
Initial Interest	3.65	.43	3.07	.28				
Outcome Interest	3.67	.49	2.98	.27				
Learning Progress	3.46	.38	3.29	.22				
Satisfaction/Teacher	3.74	.71	3.95	.64				
Satisfaction/Person	3.84	.55	4.04	.50				
Expected Grade	4.11	.33	4.21	.22				
GPA	3.83	.43	3.64	.30				
SAT Math	3.58	.41	3.52	.33	3.14	.23		
SAT Verbal	3.79	.43	3.54	.23			3.87	.37
Academic Motivation	6.62	.34	6.51	.28				

Table 7

Split-Half Interrater Reliability Estimates for CLIC and Supplementary Scales

	Classes			
	Regular	Speech	Math	German
Average number of raters per class	32	15	18	11
Number of classes	115	28	75	10
Knowledge and Skill	.85	.66	.77	.72
Consideration	.90	.75	.90	.71
Critical Demands	.86	.86	.84	.88
Student Involvement	.80	.05	.15	--[a]
Coordination	.79	.51	.19	--[a]
Overall Satisfaction	.85	.57	.79	.15
Initial Interest	.76	.47		
Outcome Interest	.76	.22		
Learning Progress	.82	.36		
Satisfaction/Teacher	.88	.85		
Satisfaction/Person	.85	.73		
Expected Grade	.53	.46		
GPA	.58	.18		
SAT Math	.69	.35		
SAT Verbal	.70	--[a]		
Academic Motivation	.21	--[a]		
Performance		.51	.88[b]	.43

[a] Indicates negative split-half correlations.

[b] Internal consistency reliability (Spearman-Brown).

variables. Therefore the lower reliabilities for these variables seemed understandable. In fact, reliabilities for SAT Math and SAT Verbal in the Regular classes seemed unexpectedly high. On the other hand, the low reliability of Academic Motivation led to an examination of the correlations between this variable and the eleven scale scores. The highest correlation was .07 (with both Student Involvement and Learning Progress) and thus Academic Motivation was eliminated from further consideration.

In the classes other than the Regular group, the low split-half reliabilities that occurred for Student Involvement, Coordination, Overall Satisfaction, the Interest scales, and Learning Progress prompted an investigation of the internal consistency reliability of the scales at the student level of analysis. Table 8 shows the Spearman-Brown internal consistency reliability estimates calculated from the average interitem correlation for each scale, Interitem correlations were calculated by pooling all students within each of the four types of classes. In general, estimates appeared acceptable and consistent across the four groups of classes.

As noted in the previous chapter, there were fourteen Speech Instructors who taught at least two classes, and 35 Math instructors who taught at least two classes.[2] For these instructors it was possible to examine the stability of ratings of an instructor made by different sections. Table 9 presents correlations between ratings made by one section and ratings made by another section on the same instructor, computed across instructors. Consideration, Critical Demands and the Satisfaction scales

---

[2]There was one instructor in Math and one in Speech who taught three sections. For this analysis one section, selected at random, was removed for each of those instructors.

Table 8

Internal Consistency Reliability Estimates

Scales	Classes			
	Regular	Speech	Math	German
Knowledge and Skill (7)[a]	.82	.85	.86	.69
Consideration (7)	.86	.83	.82	.77
Critical Demands (7)	.82	.80	.72	.69
Student Involvement (8)	.72	.70	.74	.82
Coordination (7)	.68	.71	.74	.68
Overall Satisfaction (5)	.86	.75	.84	.80
Initial Interest (5)	.83	.86		
Outcome Interest (4)	.82	.70		
Learning Progress (12)	.92	.93		
Satisfaction/Teacher (3)	.94	.90		
Satisfaction/Person (3)	.90	.89		

[a]Numbers in parentheses indicate the number of items on each scale.

Table 9

Same Instructor/Different Section Correlations for

CLIC and Supplementary Scales

Scales	Classes: Speech (N=14)	Classes: Math (N=35)
Knowledge and Skill	.52	.38*
Consideration	.62*	.72**
Critical Demands	.87**	.63**
Student Involvement	.10	.14
Coordination	-.31	.20
Overall Satisfaction	.63*	.75**
Initial Interest	-.12	
Outcome Interest	-.13	
Learning Progress	.25	
Satisfaction/Teacher	.81**	
Satisfaction/Person	.72**	
Expected Grade	-.09	
GPA	-.24	
SAT Math	-.02	
SAT Verbal	.08	
Academic Motivation	-.08	
Performance	-.44	.25

*$p < .05$

**$p < .01$

seemed to be the only stable characteristics.[3] Note that neither for Speech nor for Math were the correlations for the performance criteria significant.

Sources of Contamination

Table 10 presents the correlations between expected grade, grade point average (GPA), SAT Math and SAT Verbal and the six CLIC scales using only the Regular classes. Correlations between these four sources of contamination and the five supplementary scales are also presented. GPA, SAT Math and SAT Verbal showed no significant correlation with any of the eleven scales. Hoyt's (1973) assertion, that ratings of Learning Progress made on an intraindividual scale control for individual differences due to ability (SAT scores) and motivation (a component of GPA), was supported.

In contrast to GPA and the SAT scores, only Knowledge and Skill and Coordination did not have significant ($p<.05$) correlation with expected grade. Overall Satisfaction was originally constructed to have a low correlation with expected grade (Wetrogen, 1970). Causal assumptions would have to be made in order to decide whether the significant correlations between the scales and expected grade constitute support for unwanted contamination or support for the validity of the scales as indicators of instructional effectiveness.

---

[3] A similar analysis could have been conducted on the Regular instructors who taught more than one class. However, as Kulik and Kulik (1974) suggest, ratings are a function of the course and the instructor. Thus correlations in Table 6 reflected the relationship between same course/same instructor unit pairs. For the most part, Regular instructors taught different courses and thus same course/same instruction unit pairs occurred very infrequently.

Table 10

Correlations Between Potential Sources of Bias and

CLIC and Supplementary Scale for 115 Regular Classes

	Expected Grade	GPA	SAT Math	SAT Verbal
Expected Grade		.21	.09	.17
GPA			.58	.72
SAT Math				.68
Knowledge and Skill	.02	.06	.13	.13
Consideration	.34**	.05	.09	.10
Critical Demands	-.28**	-.09	-.08	-.19
Student Involvement	.53**	.03	-.06	-.06
Coordination	.14	-.03	.01	.05
Overall Satisfaction	.25**	-.02	.03	-.01
Initial Interest	.23*	.06	.01	.08
Outcome Interest	.43**	-.04	-.12	-.03
Learning Progress	.28**	.00	.02	.09
Satisfaction/Teacher	.32**	-.02	.06	.04
Satisfaction/Person	.33**	-.06	.07	.05

---

*$p < .05$

**$p < .01$

Multiple Regression Analyses

Regular Classes

Table 11 presents the correlations between the 13 variables used in the following analyses. As indicated in Chapter 2, regression equations were computed for each of the following variables: Learning Progress, Outcome Interest, Satisfaction/Person, Satisfaction/Instructor, Overall Satisfaction and Student Involvement. A stepwise multiple regression solution was computed for each of these variables using all of the variables to the left of the focal variable as displayed in Figure 2.

In addition to the variables that appear in Figure 2, a number of interaction terms were also examined for any contribution after simple additive effects of the interacting variables were accounted for. These included Size X Critical Demands, Size X Consideration, Size X Coordination, Format X Critical Demands, Format X Consideration and Format X Coordination. In general these variables did not enter the stepwise analysis. To simplify the following discussion, these terms will be mentioned only if they made a significant contribution.

The final model for the Regular classes appears as Figure 3. An arrow indicates that the corresponding antecedent variable entered and remained in the regression equation with a significant ($p<.05$) beta weight.

To help interpret the model, Knowledge and Skill (KS), Critical Demand (CD), Consideration (CON), and Coordination (COOR) were regressed on Class Size (SZ), Course Format (FM) and students' Initial Interest (II). These results are also depicted in Figure 3. Forty-four percent of the variance in Consideration ($R = .66$, $F = 24.73$, $p<.001$) was accounted for by the joint linear effects of Size, Format and Initial Interest.

Table 11

Intercorrelations Among CLIC and Supplementary Scales for 115 Regular Classes

Scales	2	3	4	5	6	7	8	9	10	11	12	13
1. Knowledge and Skill	.58	-.28	.19	.29	.70	.32	.28	.59	.74	.62	-.13	.10
2. Consideration		-.46	.54	.36	.58	.46	.48	.68	.72	.79	-.47	.43
3. Critical Demands			-.27	-.16	-.32	-.37	-.41	-.38	-.46	-.57	.30	-.09
4. Student Involvement				.33	.54	.42	.72	.56	.54	.46	-.40	.43
5. Coordination					.32	.07	.18	.34	.40	.38	-.18	.33
6. Overall Satisfaction						.26	.47	.59	.86	.59	-.20	.19
7. Initial Interest							.78	.70	.36	.43	-.19	.13
8. Outcome Interest								.69	.48	.45	-.24	.25
9. Learning Progress									.70	.65	-.32	.32
10. Satisfaction/Teacher										.84	-.18	.24
11. Satisfaction/Person											-.26	.32
12. Class Size												-.32
13. Course Format												

Note: Correlations greater than .19, $p<.05$; correlations greater than .25, $p<.01$.

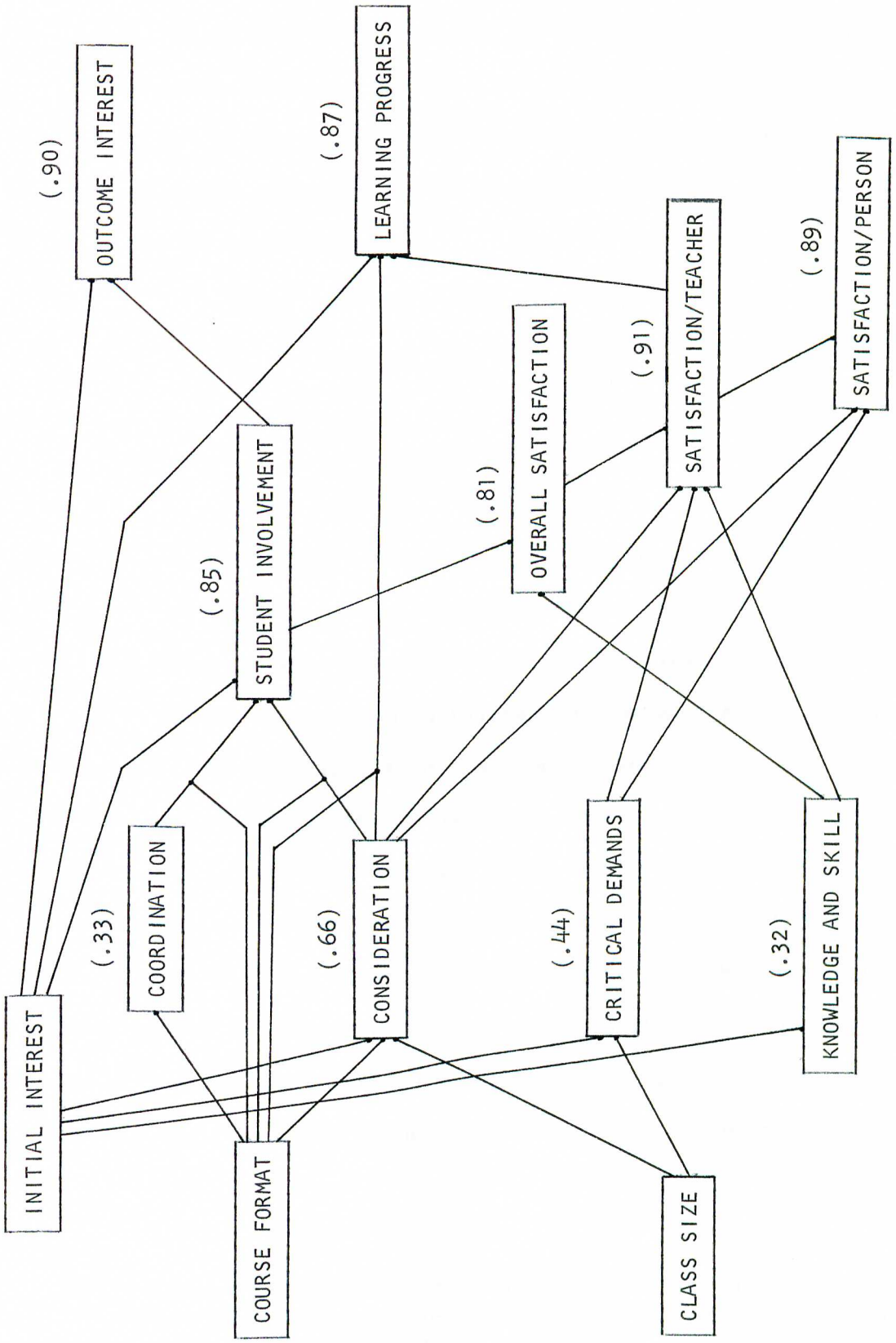

Figure 3. Relationships for Regular classes. Numbers in parenthesis are the multiple correlations.

Twenty percent of the variance in Critical Demands was accounted for by Size and Intitial Interest ($\underline{R}$ = .44, $\underline{F}$ = 11.53, $\underline{p}$<.001). Ten percent of the variance in Knowledge and Skill was accounted for by Initial Interest ($\underline{r}$ = .32, $\underline{F}$ = 11.30, $\underline{p}$<.001). Finally, eleven percent of the variance in Coordination was accounted for by Format ($\underline{r}$ = .33, $\underline{F}$ = 11.97, $\underline{p}$<.001). The regression equations were:

$$Z_{CON} = .29Z_{FM} - .31Z_{SZ} + .36Z_{II}$$

$$Z_{CD} = .24Z_{SZ} - .32Z_{II}$$

$$Z_{KS} = .32Z_{II}$$

$$Z_{COOR} = .33Z_{FM}$$

The negative weights were not indicative of suppressor variables since the correlation between Initial Interest and Critical Demands and the correlation between Size and Consideration were both negative.[4] Initial Interest appeared to be related to Knowledge and Skill, Critical Demands and Consideration, such that for classes rated as more interesting and relevant at the beginning of the semester the instructors were rated as having more Knowledge and Skill, being more Considerate and making fewer Critical Demands. Course format showed significant direct relationships with Coordination and Consideration. Thus, courses described by the instructor as seminar or group discussion courses were rated by the students as being more Coordinated (more structured) and the instructor was rated higher on Consideration. Size of the class was

[4]Throughout this chapter, Size and Critical Demands appear with negative beta weights. These negative weights do not indicate suppressor effects in any of the equations.

directly related to Consideration and Critical Demands. In larger classes, instructors received lower ratings on Consideration and higher ratings on Critical Demands.

Student Involvement. As suggested in Figure 2, Knowledge and Skill (KS), Critical Demands (CD), Consideration (CON), Coordination (COOR), Initial Interest (II), Size (SZ), Format (FM), and the interaction terms entered the regression analysis for Student Involvement (SI). The following variables were selected by the stepwise procedure: Consideration, Format, Initial Interest, Coordination, the interaction of Format and Coordination, and finally the interaction of Format and Consideration. The multiple correlation was .85 ($F = 38.76$, $p < .001$), indicating that these variables combined to account for 72 percent of the variance in Student Involvement. After the additive terms were in the equation, the addition of the Format X Coordination interaction term increased the variance accounted for by 25 percent; then the addition of the Format X Consideration interaction term increased the variance accounted by another seven percent. The interaction terms were interpreted as indicating that the relationships of Consideration and Coordination with Student Involvement were dependent on Format. For interpretation, two regression equations were derived from the multiple regression result by substituting the appropriate values for Format.

The unstandardized equation derived from the multiple regression procedure was:

$$SI = .3911 - .83CON - 1.87COOR - 9.65FM + .87(FM)(CON) + 1.87\ (FM)(COOR) + 11.55$$

Setting FM = 1 for lecture classes:

$$SI = .3911 + .03CON + .00COOR + 1.905$$

This equation was then standardized to Z-scores based on the means and

standard deviations of the total sample (i.e., both lecture and discussion classes). Thus

$$Z_{SI}(.35) + 3.56 = .39[Z_{II}(.43) + 3.65] + .03[Z_{CON}(.56) + 3.68] + .00[Z_{COOR}(.40) + 3.24] + 1.90$$

$$Z_{SI} = .48Z_{II} + .05Z_{CON} + .00Z_{COOR} - .28 \qquad \text{(Lecture)}$$

An equation for discussion classes was derived in the same manner. For discussion classes, FM = 2; therefore:

$$SI = .39II + .90CON + 1.87COOR - 7.74$$

Standardizing using means and standard deviations of the total sample:

$$Z_{SI} = .48Z_{II} + 1.44Z_{CON} + 2.14Z_{COOR} - 1.43 \qquad \text{(Discussion)}$$

These equations illustrate that the relationship of Coordination and Consideration with Student Involvement were dependent on the Course Format. For lecture classes, neither Consideration nor Coordination were highly associated with Student Involvement. For discussion classes, Consideration and Coordination were directly related to Student Involvement such that large Consideration and large Coordination scores were associated with more Student Involvement. On the other hand the effects of Initial Interest were constant across Format. High Initial Interest was associated with high Student Involvement. The constants in the equations, which indicate the residual effects of Format, suggest that Student Involvement is greater in discussion classes.

Overall Satisfaction. Knowledge and Skill, followed by Student Involvement and then Initial Interest were the only variables accepted in the multiple regression equation according to the stepwise criteria. The multiple correlation was .82 ($F = 78.99$, $p < .001$). However, Initial Interest entered the equation as a suppressor variable. Noting that Initial Interest increased the multiple R^2 by only two percent led to the

conclusion that Initial Interest could be parsimoniously removed from the equation with little, if any, effect on the overall interpretation. Knowledge and Skill, and Student Involvement accounted for 66 percent of the variance ($\underline{R}$ = .81, $\underline{F}$ = 11.79, $\underline{p}$<.001) in Overall Satisfaction (OS). The equation was:

$$Z_{OS} = .62Z_{KS} + .42Z_{SI}$$

Thus, in general, high Knowledge and Skill and high Student Involvement were associated with high Overall Satisfaction.

Satisfaction with the Instructor as a Teacher. Satisfaction/Teacher was examined in two ways. First, the partial correlations between Satisfaction/Teacher and the CLIC scales, with the effects of Satisfaction/Person removed, were computed. Three of the six partial correlations were significant at .001. These were .50 ($\underline{F}$ = 32.19) for Knowledge and Skill, .33 ($\underline{F}$ = 11.43) for Student Involvement, and .82 ($\underline{F}$ = 197.04) for Overall Satisfaction. The zero order correlations (Table 11) with Satisfaction/Teacher were .74 for Knowledge and Skill, .54 for Student Involvement, and .86 for Overall Satisfaction.

Second, Satisfaction/Teacher (ST) was regressed on all six CLIC scales, plus Size Format, students' Initial Interest, and the interaction terms. Overall Satisfaction, Consideration, Size, Critical Demands, Knowledge and Skill, and Student Involvement entered and remained in the regression equation with beta weights significant at .05 or less. Size, however, entered as a suppressor variable accounting for only two percent of the variance. Therefore, the regression analysis was repeated excluding Size. The following variables entered the equation: Overall Satisfaction, Consideration, Knowledge and Skill, and Critical Demands. The multiple correlation was .91 ($\underline{R}^2$ = .83, $\underline{F}$ = 138.92, $\underline{p}$<.001). The

equation was:

$$Z_{ST} = .56Z_{OS} + .25Z_{CON} + .16Z_{KS} - .12Z_{CD}$$

Low scores on Critical Demands and high scores on Overall Satisfaction, Consideration, and Knowledge and Skill were associated with high Satisfaction/Teacher.

Satisfaction with the Instructor as a Person. The partial correlations between Satisfaction/Person and the CLIC scales, with the effects of Satisfaction/Teacher removed, were computed. The partial correlations for Consideration and for Critical Demands, .50 and -.37 were significant at .001 (F = 35.72 and 18.57, respectively). The partial correlation for Overall Satisfaction was also significant at .001 (F = 26.07), however the sign was reversed from the zero order correlation between Overall Satisfaction and Satisfaction/Person. The zero order correlations (Table 11) with Satisfaction/Person were .79 for Consideration, -.57 for Critical Demands, and .59 for Overall Satisfaction.

Satisfaction/Teacher had the highest correlation with Satisfaction/Person and entered the equation first. Consideration, Overall Satisfaction, and Critical Demands also entered and remained in the equation. Overall Satisfaction entered the equation as a suppressor variable, and therefore the direct effect of Overall Satisfaction on Satisfaction/Person, after the effects of Satisfaction/Teacher, Consideration and Critical Demands were accounted for was assumed to be zero. The resulting multiple correlation was .89 (F = 148.00, $p<.001$). The equation was:

$$Z_{SP} = .50Z_{ST} - .18Z_{CD} + .35Z_{CON}$$

Low Critical Demands, high Satisfaction/Teacher and high Consideration were associated with high Satisfaction/Person.

Outcome Interest. All of the variables to the left of Outcome

Interest (OI) in Figure 2, Initial Interest was the first variable to enter the stepwise regression equation. Considering the methodology, it did not seem unexpected that Initial Interest correlated .78 (F = 182.40, $p<.001$) with Outcome Interest, accounting for 62 percent of the variance. At this point in the procedure, four of the six CLIC scales had partial correlations with Outcome Interest, holding Initial Interest constant, which were significant at .05 or less. These were .21 (F = 4.63) for Consideration, -.21 (F = 4.61) for Critical Demands, .70 (F = 93.98) for Student Involvement, and .44 (F = 23.41) for Overall Satisfaction. Having the largest partial, Student Involvement was selected as the second variable in the equation, and it increased the variance accounted for by 19 percent. Holding Initial Interest and Student Involvement constant, none of the partial correlations were significant, therefore no other variable entered the equation.

Initial Interest and Student Involvement accounted for 81 percent of the variance in Outcome Interest (R = .90, F = 232.66, $p<.001$). The equation was:

$$Z_{OI} = .58Z_{II} + .48Z_{SI}$$

Thus, in general, high Initial Interest and high Student Involvement were associated with high Outcome Interest.

Learning Progress. The final variable in the rating model was Learning Progress (LP). Initial Interest was the first variable to enter the equation, accounting for 48 percent of the variance in Learning Progress. With Initial Interest held constant, five of the six partial correlations between the CLIC scales and Learning Progress were significant at .05 or less. They were .54 (F = 39.48) for Knowledge and Skill, and .57 (F = 45.87) for Consideration, .41 (F = 19.84) for Student Involvement, .40

($\underline{F}$ = 18.45) for Coordination, and .59 ($\underline{F}$ = 51.24) for Overall Satisfaction. The largest partial however, was for Satisfaction/Teacher, .66 ($\underline{F}$ = 76.06). Thus, Satisfaction/Teacher was the second variable to enter the equation. Format, Consideration, and the interaction of Format and Consideration were also selected for the regression equation. The addition of Satisfaction/Teacher, Consideration and Format increased the variance accounted for by 26 percent. The addition of the interaction term ($\underline{F}$ = 7.92, $\underline{p}<.001$ for the increase) increased the $\underline{R}^2$ by two percent ($\underline{R}$ = .87, $\underline{F}$ = 58.04, $\underline{p}<.001$). The interaction term was interpreted as signifying that the effect of Consideration was dependent on course Format. Thus two equations were derived, one for lecture courses and one for discussion courses, using the same procedure described for Student Involvement. The resulting equations were:

Lecture: $Z_{LP} = .43Z_{II} + .36Z_{ST} + .11Z_{CON} - .06$

Discussion: $Z_{LP} = .43Z_{II} + .36Z_{ST} + .75Z_{CON} - .24$

The relationship between ratings on Consideration and ratings on Learning Progress appeared to be stronger for discussion classes than for lecture classes. On the other hand, the effects of Initial Interest, and Student Involvement were constant across Format. High Initial Interest and high Student Involvement were associated with high Learning Progress. The constants indicate the residual effects of Format.

A regression equation was also selected for Learning Progress with Satisfaction/Teacher and Satisfaction/Person removed from the analysis. The following variables entered and remained in the regression equation: Initial Interest, accounting for 48 percent of the variance, then Consideration, Knowledge and Skill, Student Involvement and Format accounting for an additional 25 percent of the variance in Learning Progress,

and then the interaction of Format X Consideration accounting for another three percent of the variance. Although Overall Satisfaction was the first scale to enter after Initial Interest, it did not remain after Knowledge and Skill and Student Involvement were entered. The multiple correlation was .87 ($F = 49.83$), $p<.001$). Again two equations were derived:

$$\text{Lecture:} \quad Z_{LP} = .36Z_{II} + .21Z_{SI} + .28Z_{KS} + .11Z_{CON} - .01$$

$$\text{Discussion:} \quad Z_{LP} = .36Z_{II} + .21Z_{SI} + .28Z_{KS} + .92Z_{CON} - .37$$

Thus, in general, over and above the effects of Initial Interest, high Knowledge and Skill, high Student Involvement, and high Consideration were associated with high Learning Progress in discussion classes. In lecture classes, Consideration was less strongly related to Learning Progress.

Figure 4 illustrates the relationships when Satisfaction/Teacher and Satisfaction/Person were not considered as predictors of Outcome Interest or Learning Progress.

Speech Classes

As noted in Chapter 3, data was available for a total of 33 sections of Speech. However, it was possible to separate only twenty-eight individual sections. Seventeen instructors participated, fourteen of them teaching more than one course. Thus the analysis was conducted (1) using the 17 instructors as the unit of analysis and (2) using the 28 sections as the unit of analysis. Only the analysis for the 28 sections is presented because, as indicated in Table 9, a number of variables were not stable within instructors teaching more than one section, and because the results of the analysis for the seventeen instructors were readily interpretable from the results obtained on the twenty-eight sections.

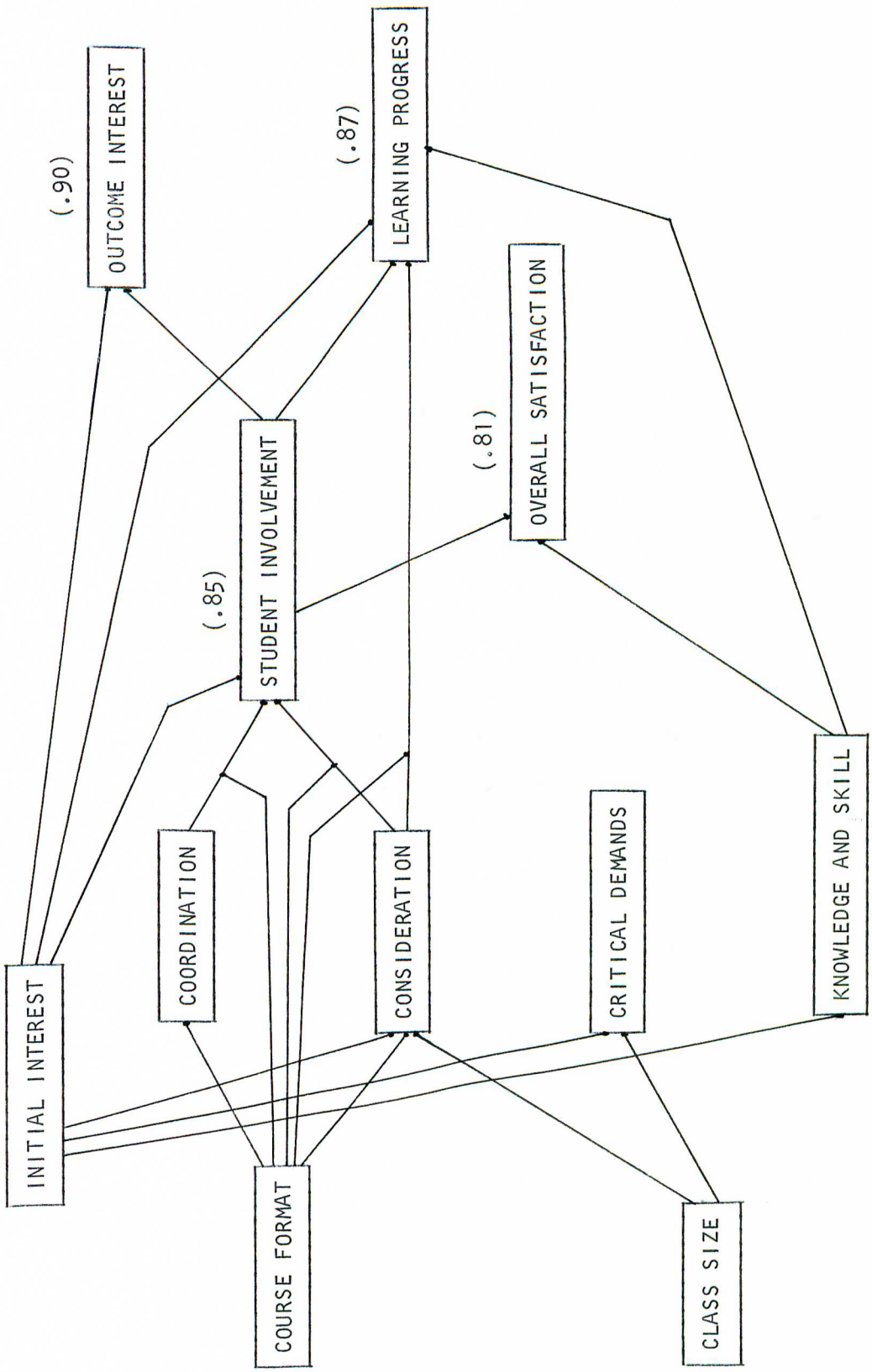

Figure 4. Relationships for Regular with Satisfaction/Teacher and Satisfaction/Person removed.

Table 12 presents the intercorrelations of the variables used in the analysis of the Speech sample. Class format and its interaction terms were not included in the analysis since the format was the same for all sections. Class size was defined as the number of students who received grades for each section. The mean class size was 18.50, standard deviation 3.07. No significant correlations were obtained between class size and the six CLIC scales, nor did size contribute to any of the regression analyses. The interaction terms with size were not analyzed.

The mean Student Performance score, as defined in Chapter 3, was 80.76 when corrected to a 100 point scale. The standard deviation was 3.45. The means and standard deviations of the other variables appear in Table 6. Figure 5 depicts the results of the following analyses.

Student Involvement. Initial Interest was the only antecedent variable significantly correlated with Student Involvement (r = .42, $p<.05$). Therefore, it was the first variable to enter the stepwise procedure, accounting for 18 percent of the variance in Student Involvement. No other variable significantly increased the variance accounted for. High Initial Interest was associated with high Student Involvement.

Overall Satisfaction. Critical Demands entered the regression equation first, followed by Knowledge and Skill, and Coordination. No other variable entered the equation. The resulting multiple correlation was .85 (R^2 = .73, F = 21.42, $p<.001$). The equation was:

$$Z_{OS} = .45Z_{KS} - .44Z_{CD} + .29Z_{COOR}$$

Low Critical Demands, high Knowledge and Skill and high Coordination were associated with high Overall Satisfaction.

Satisfaction with the Instructor as a Teacher. With the effects of Satisfaction/Person removed, the partial correlations between Satisfac-

Table 12

Intercorrelations Among CLIC and Supplementary Scales for 28 Speech Classes

Scales	2	3	4	5	6	7	8	9	10	11	12	13
1. Knowledge and Skill	.81	-.61	.04	-.04	.71	.05	.26	.44	.73	.55	-.02	-.02
2. Consideration		-.74	-.09	-.03	.61	-.12	.11	.22	.85	.67	.19	-.20
3. Critical Demands			.10	-.05	-.73	.05	-.14	-.16	-.85	-.80	-.36	.22
4. Student Involvement				.36	.19	.42	.59	.49	-.11	-.22	-.10	.08
5. Coordination					.30	.30	.11	.20	.10	.15	-.29	-.30
6. Overall Satisfaction						.09	.33	.48	.77	.65	.05	-.11
7. Initial Interest							.79	.55	-.05	-.01	.22	.23
8. Outcome Interest								.65	.14	.10	.22	.33
9. Learning Progress									.22	.15	.01	.13
10. Satisfaction/Teacher										.91	.20	-.26
11. Satisfaction/Person											.27	-.15
12. Student Performance												.25
13. Class Size												

Note: Correlations greater than .37, $p<.05$; correlations greater than .48, $p<.01$.

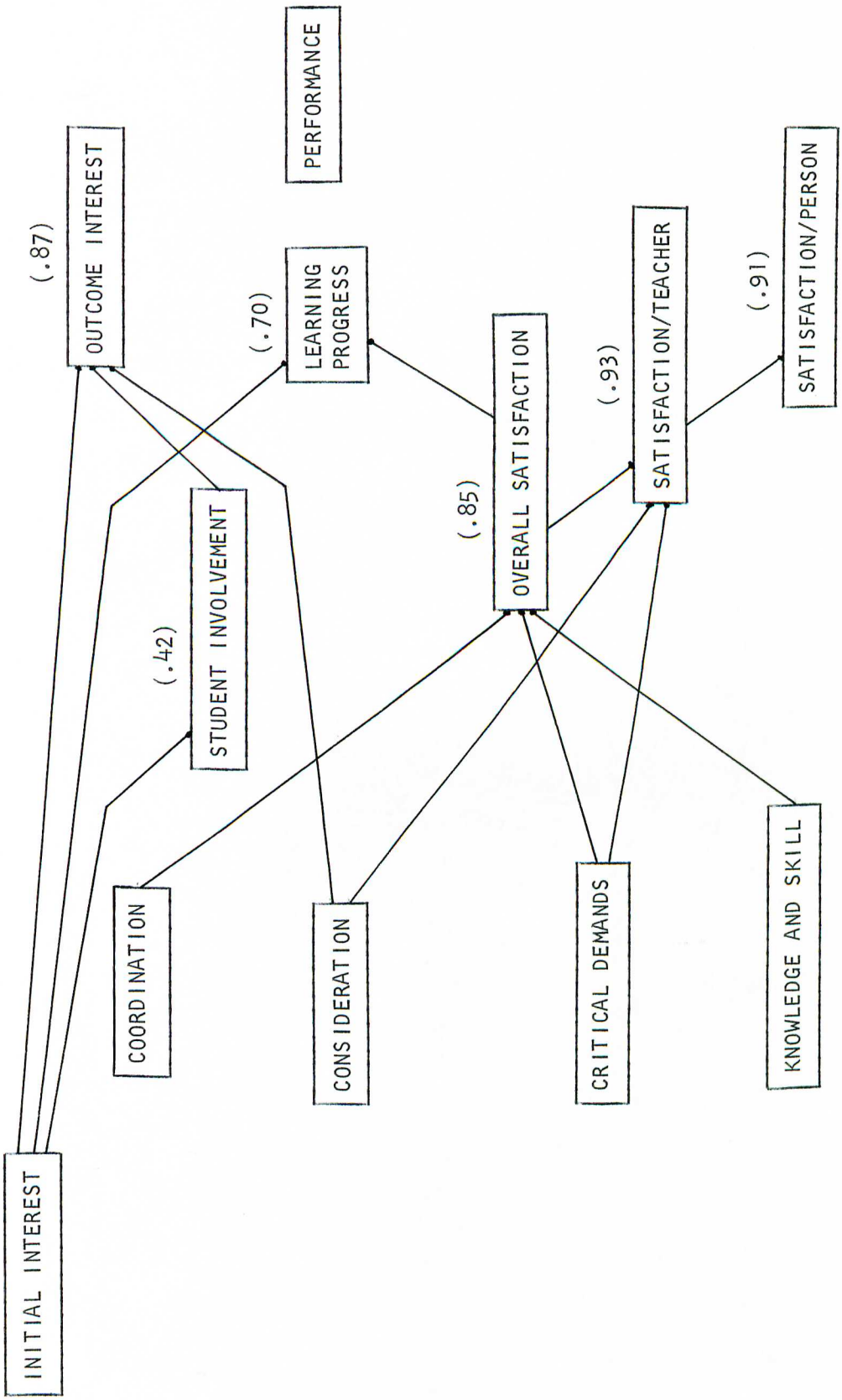

Figure 5. Relationships for Speech classes.

tion/Teacher and four CLIC scales were significant at .001. These were .66 (F = 19.20) for Knowledge and Skill, .78 (F = 40.00) for Consideration, -.49 (F = 8.01) for Critical Demands, and .58 (F = 12.61) for Overall Satisfaction. The zero order correlations for these scales (Table 12) were .73, .85, -.85, and .77, respectively.

A regression equation was selected with Satisfaction/Person not included in the analysis. Consideration, Critical Demands, and Overall Satisfaction entered and remained in the equation. The multiple R was .93 (F = 51.52, $p<.001$), indicating that .86 of the variance in Satisfaction/Teacher was accounted for by these three variables. The regression equation was:

$$Z_{ST} = .45Z_{CON} - .33Z_{CD} + .26Z_{OS}$$

In general, high Consideration, low Critical Demands, and high Overall Satisfaction were associated with high Satisfaction/Teacher.

Satisfaction with the Instructor as a Person. For Satisfaction/Person, Satisfaction/Teacher was the first variable to enter the equation. Either Knowledge and Skill, or Consideration could have entered with a significant beta weight, indicating that each of them had significant partial correlations with Satisfaction/Person, with Satisfaction/Teacher held constant. However, if either one were accepted in the equation, it would have entered as a suppressor variable, accounting for only a small increase in the variance accounted for in Satisfaction/Person. The correlation between Satisfaction/Teacher and Satisfaction/Person was .91 (F = 128.27, $p<.001$), indicating that 83 percent of the variance in Satisfaction/Person could be accounted for by Satisfaction/Teacher.

Outcome Interest. Initial Interest had a zero order correlation of .79 (F = 44.74, $p<.001$) with Outcome Interest and entered the regression

equation first. With Initial Interest held constant, the partial correlations of Student Involvement and Overall Satisfaction with Outcome Interest were each significant at the .05 level. The partial correlations were .46 for Student Involvement, and .42 for Overall Satisfaction. After Student Involvement was entered into the equation, the second order partial correlations for Knowledge and Skill (.40), Consideration (.41), Critical Demands (-.39), and Coordination (-.41) were all significant. Consideration was entered into the equation. Coordination could have entered after Consideration, however it would have entered as a suppressor variable accounting for only a small proportion of the variance in Outcome Interest. Initial Interest, Student Involvement, and Consideration accounted for 76 percent of the variance in Outcome Interest with a multiple correlation of .87 ($F = 25.21$, $p < .001$). The equation was:

$$Z_{OI} = .69Z_{II} + .32Z_{SI} + .22Z_{CON}$$

High Initial Interest, high Student Involvement, and high Consideration were associated with high Outcome Interest.

Learning Progress. When the six CLIC scales along with the supplementary scales were regressed on Learning Progress according to the stepwise procedure, only Outcome Interest entered the equation. The correlation between Outcome Interest and Learning Progress was .65 ($F = 19.31$, $p < .001$, $r^2 = .43$). High Outcome Interest was associated with high Learning Progress.

For comparison with the Regular results, the data was reanalyzed treating Initial Interest as a covariate and forcing it into the equation first. At this point, Overall Satisfaction and Knowledge and Skill both had significant partial correlations with Learning Progress. These were .52 for Overall Satisfaction and .50 for Knowledge and Skill. Overall Satisfaction entered the equation first. The second order partial for Know-

ledge and Skill was not significant. Initial Interest accounted for 31 percent of the variance in Learning Progress; Overall Satisfaction then accounted for an additional 18 percent of the variance. The multiple correlation was .70 (F = 12.15, $p<.001$). The equation was:

$$Z_{LP} = .51Z_{II} + .43Z_{OS}$$

Student Performance. As observed in Table 12, mean section performance did not correlate significantly with any of the variables included in the model. The only variable with which Performance was significantly correlated was mean section rating of expected grade (r = .46, $p<.05$). Note that expected grade was not significantly correlated with Learning Progress (r = -.06). The only variables, beside Performance, which were significantly correlated with expected grade, were Critical Demands (r = -.38, $p<.05$) and GPA (r = .68, $p<.05$).

Math Classes

As for the Speech sample, regression analyses were conducted on the 75 sections and on the 35 instructors who taught more than one course. Again, similar to the Speech sample, the results from each analysis were essentially the same. Only the results from the analysis on the 75 sections will be presented.

Table 13 presents the intercorrelations of the variables included in the analysis of the Math sections. Format and its interactions were excluded. While computing mean test performance, it was observed that a large number of students did not complete the course. Therefore the following indices related to class size were included in the analysis: number of students in each section who took the first test, number of students who took the final test, and the ratio of the number who took the final to the number who took the first test, indicative of the com-

Table 13

Intercorrelations Among CLIC Scales and Supplementary Items for Math Classes

Scales	2	3	4	5	6	7	8	9	10	11
1. Knowledge and Skill	.75	-.34	.40	.25	.85	.02	.01	.04	-.10	.27
2. Consideration		-.34	.54	.17	.83	-.07	.01	.01	-.01	.33
3. Critical Demands			-.02	.08	-.31	.00	-.21	-.15	-.18	-.15
4. Student Involvement				.32	.58	-.16	-.08	-.03	-.13	.22
5. Coordination					.32	.14	.08	.00	-.18	.29
6. Overall Satisfaction						-.03	.12	.15	-.03	.43
7. SAT Math							.09	-.02	.26	.23
8. Number for Final Test								.90	.62	.51
9. Number for First Test									.24	.40
10. Percent Completing Course										.42
11. Student Performance										

Note: Correlations greater than .22, $p<.05$; correlations greater than .29, $p<.01$.

pletion rate for each section. Table 14 presents the means and standard deviations of these indices and the performance criterion. Regression analyses were conducted on Student Involvement, Overall Satisfaction, and Performance. Figure 6 depicts the following results.

Student Involvement. Using the stepwise regression procedure, Consideration and Coordination were selected for the regression equation. The multiple correlation was .59 (F = 18.99, $p<.001$) with 34 percent of the variance in Student Involvement accounted for by Consideration and Coordination. The equation was:

$$Z_{SI} = .50Z_{CON} + .24Z_{COOR}$$

In general, high Consideration and high Coordination were associated with high Student Involvement.

Overall Satisfaction. Knowledge and Skill followed by Consideration, then Student Involvement entered and remained in the stepwise equation. The multiple correlation was .91 (F = 122.13, $p<.001$) with 84 percent of the variance in Overall Satisfaction accounted for. The equation was:

$$Z_{OS} = .53Z_{KS} + .33Z_{CON} + .19Z_{SI}$$

High Knowledge and Skill, high Consideration and high Student Involvement were associated with high Overall Satisfaction.

Student Performance. The analysis for the performance criterion was modified somewhat by treating SAT Math as a covariate and forcing it into the equation first. The effects of any other variables entering the equation could then be interpreted independently of SAT Math. The correlation between SAT Math and Performance was .23, indicating that 5 percent of the variance in (mean class) Performance was accounted for by (mean class) SAT Math. Number of students who took the final then entered the equation, with an additional 25 percent of the variance accounted for.

Table 14

Means and Standard Deviations of Math

Supplementary Items

	Mean	Standard Deviation
Number for 1st Test	24.72	5.34
Number for Final Test	21.87	5.83
Percent Completing Course	88	11
Student Performance	363.27	26.40

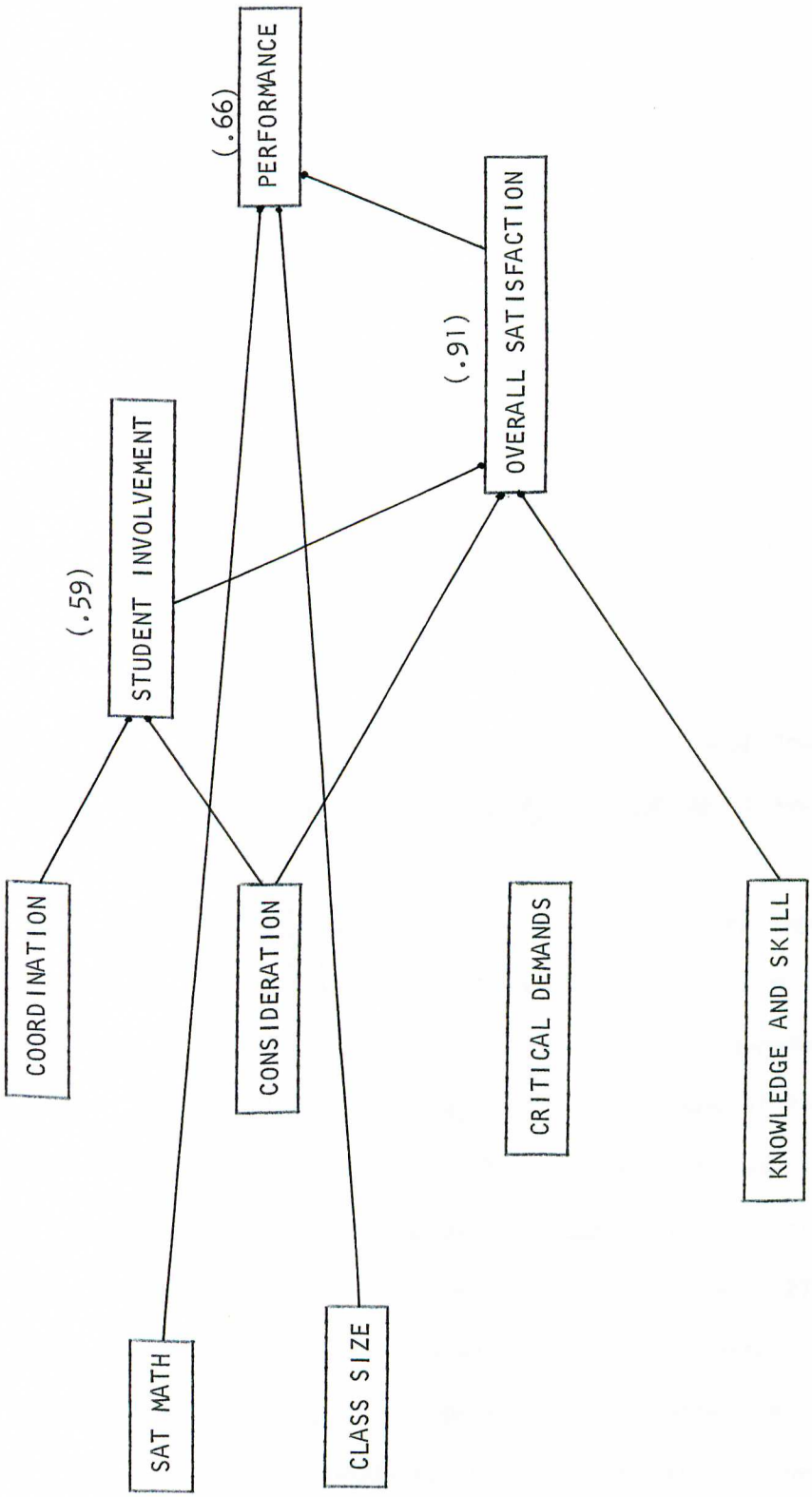

Figure 6. Relationships for Math classes.

At this point, Critical Demands was the only CLIC scale which did not have a significant (at .05 or less) second order partial correlation with Performance. The partial correlations were .33 for Knowledge and Skill, .40 for Consideration, .36 for Student Involvement, .27 for Coordination, and .45 for Overall Satisfaction. Overall Satisfaction was the third variable to enter the equation. Following this step, no other scale contributed enough unique variance to enter the equation. SAT, number of students who took the final, and Overall Satisfaction accounted for 44 percent of the variance in Performance ($\underline{R} = .66$, $\underline{F} = 18.53$, $\underline{p} < .001$). Overall Satisfaction, after SAT Math and number of students who took the final (SZ), accounted for 14 percent of the variance in Performance ($\underline{F} = 18.03$, $\underline{p} < .001$). The equation was:

$$Z_{PERF} = .20Z_{SAT} + .45Z_{SZ} + .38Z_{OS}$$

This equation may be interpreted as indicating that independent of class SAT and class size, high Overall Satisfaction was associated with high class Performance.

It could be argued that number of students who complete the course could itself be regarded as an outcome criterion, and therefore it should not be included as a predictor of Performance. Therefore, the regression analysis for Performance was repeated without any of the size indices. After SAT Math was forced into the equation, the same five CLIC scales again had significant (at .05 or less) partial correlations. These were .29 for Knowledge and Skill, .35 for Consideration, .27 for Student Involvement, .26 for Coordination, and .45 for Overall Satisfaction. Overall Satisfaction was the second and only other variable to enter the equation. Overall Satisfaction added 19 percent to the variance accounted for in Performance. SAT Math and Overall Satisfaction together accounted

for 24 percent of the variance in Performance (R = .49, F = 11.55, p<.001). The equation was:

$$Z_{PERF} = .24Z_{SAT} + .44Z_{OS}$$

German Classes

Table 15 presents the intercorrelations of the variables analyzed for the ten German classes. Means and standard deviations for the CLIC scales and SAT Verbal appear in Table 6. For the German performance criterion, the mean for the ten sections was 3.07 (on a four point scale) with a standard deviation of .24. Number of grades given was also examined ($\bar{X}$ = 14.0, S.D. = 3.20).

Student Involvement. As shown in Table 15, no variables correlated significantly with Student Involvement.

Overall Satisfaction. Consideration had the highest correlation with Overall Satisfaction and entered the equation first. Student Involvement was the variable which increased the multiple correlation the most, however the increase was not significant. The program continued to the next step and Knowledge and Skill entered the equation. At this point the increase due to Consideration, after the effects of Student Involvement, and Knowledge and Skill were accounted for, was no longer significant, however the unique contributions of Student Involvement and Knowledge and Skill were significant. Therefore Consideration was removed from the equation. The multiple correlation of Overall Satisfaction with Knowledge and Skill and Student Involvement was .90 (F = 14.94, p<.005). The equation was:

$$Z_{OS} = .68Z_{KS} + .54Z_{SI}$$

In general high Knowledge and Skill and high Student Involvement were associated with high Overall Satisfaction.

Table 15

Intercorrelations of CLIC Scale and Supplementary Items for German Classes

	2	3	4	5	6	7	8	9
1. Knowledge and Skill	.75*	-.11	.08	.43	.72*	-.15	.12	.22
2. Consideration		-.26	.37	.62	.74*	-.26	.08	.68*
3. Critical Demands			.07	-.54	-.21	.11	-.52	-.47
4. Student Involvement				-.25	.59	-.05	-.39	.20
5. Coordination					.32	-.19	.57	.49
6. Overall Satisfaction						.06	.24	.23
7. SAT Verbal							.14	-.25
8. Number of Grades Given								.01
9. Student Performance								

*$p < .05$

Performance. Table 15 shows that each of the correlations between the CLIC scales and Performance was in the expected direction. However, only one scale, Consideration, correlated significantly with Performance (r = .68, p<.05). When SAT Verbal was forced into the regression equation first as a covariate, the multiple correlation between Performance with SAT Verbal and Consideration, the next highest contributor, was not significant (R = .68, F = 3.08). The multiple correlation required for significance at the .05 level with 2 and 7 degrees of freedom is .76.

Combined Sample

In addition to the analyses reported above, an analysis was conducted on Speech, Math and German combined using the six CLIC scales, student performance, and class size indices and SAT reports. The student performance criteria were equated across samples by using the within sample Z-score for class mean performance. That is, for each class, the mean student performance scores were converted to a Z-score using the mean and standard deviation of the sample to which the class belonged. Thus for each sample the mean class criterion was zero and the standard deviation was one.

Since two-thirds of this combined sample were Math classes, and since the correlations for the ten German classes were in the same direction, the results for this sample were very similar to those reported for the Math Sample above. Therefore, these results will not be presented.

CHAPTER V

DISCUSSION: THE VALIDITY OF THE CLIC

The purpose of this research was to assess the validity of ratings made by students about various aspects of instruction described by the Check-List of Instructional Characteristics (CLIC). That goal is the same as asking about the relevance of the CLIC as a criterion of instructional effectiveness. Two essential types of evidence are needed in order to make that assessment: (1) evidence demonstrating relationships between CLIC ratings and other criteria judged to be relevant indicators of effective instruction, and (2) evidence demonstrating that CLIC ratings are not contaminated by factors that are not considered relevant attributes of effective instruction. This chapter will review that evidence and suggest some conclusions concerning the validity of the CLIC. The evidence is summarized in light of results from past research.

In addition to the discussion of the validity evidence per se, there are a number of issues concerning the process by which the evidence was collected, analyzed, and interpreted which can be considered. These issues are addressed first.

Methodological Considerations

Obviously, the manner in which data are gathered and analyzed places limits on the interpretations that can be drawn from the results. Therefore it seems appropriate to consider what constraints the methodology of this study might place on the interpretation of the CLIC validity evidence.

Statistical Considerations

Reliability estimates. A fundamental rule in the area of measurement and statistics is that data must be reliable in order to demonstrate any statistical relationships. Table 7 shows that the interrater reliability estimates for a number of the CLIC and supplementary validation scales were quite low and in some cases non-existent. For example, the interrater reliabilities of Student Involvement and Coordination were of sufficient magnitude to be considered acceptable in only the Regular sample. Nevertheless, a number of significant correlations were obtained for these scales.

One might suggest the class level reliabilities may have been low because the reliabilities of the individual student scores were low. However, the internal consistency reliability estimates computed at the individual level (Table 8) dispell that argument.

The standard deviations of the class scale scores offer another possible reason for the low reliability estimates. Table 6 shows that the standard deviations for Student Involvement and for Coordination for Speech, Math and German are all approximately one-half the magnitude of the respective standard deviations for Student Involvement and Coordination in the Regular sample. Each of these differences were significant at .05 or less.[5] A formula presented by Ghiselli (1964) indicates that a fifty percent reduction in the standard deviation of a measure with an original reliability of .80, approximately the reliability of Student Involvement and Coordination in the Regular sample, will result in a

[5]The significance level was determined by comparing the ratio of the two variances being compared to the F-distribution.

reliability of .20. Thus, the restriction in the variance across classes may account for the low reliability estimates in the Speech, Math, and German samples.

Multiple regression weights and partial correlations. An arrow in any of the figures which depict the regression results identifies variables that entered and remained in the corresponding regression equations with significant beta weights. In path analysis terms these arrows would be called direct effects. For this discussion the term direct relationship will be adopted and is defined as a relationship between variables that cannot be attributed to another set of variables. That is, a direct relationship exists when the partial correlation between two variables is significantly different from zero when the other antecedent variables are held constant. That, of course, is equivalent to saying that the beta weight is significantly different from zero, or that the variable entered and remained in the stepwise regression procedure. Obviously, the existence of a direct relationship is dependent on the variables included in the analysis.

In Chapter 4, beta weights for each of the direct relationships were presented in each of the equations. The actual weights will not be interpreted in this discussion since these weights carry little unambiguous explanatory meaning (e.g., Kerlinger & Pedhazur, 1973). Furthermore, the purpose of the research is not to provide point predictions about instructional performance, a strategy which might require the use of beta weights. If this were the purpose, the utility of the beta weights would need to be verified by cross-validation. Even then, unit weights (i.e., all predictors included in the equation weighted by a one) might prove to be at least as useful as the actual weights (Wainer, 1976). The weights were

presented only for the interest of the reader and for guidance in possible future cross-validation research.

Even the use of unit weights does not solve the potential problem of over interpreting the regression equations. In many cases, the difference between a variable that entered the equation and one that did not was small. For example, in the Speech analysis the part correlation between Knowledge and Skill and Learning Progress, with the effects of Initial Interest removed from Knowledge and Skill, was .50. The analogous part correlation between Overall Satisfaction and Learning Progress was .52. Therefore, Overall Satisfaction was selected for the regression equation based on a difference of .02. Knowledge and Skill did not enter the equation. The difference could have resulted from chance error variance.

On the other hand, two potential predictors can have similar partial correlations with the criterion and, in contrast to the above example, both of them eventually enter the regression equation. At any step in the stepwise regression procedure, the part correlations tell how much unique variance each scale can account for in the criterion, but they do not tell how much variance the scales share, how many of the scales will enter the equation in the final analysis, or how much of the total criterion variance will be accounted for. The part (and partial) correlations and the regression equations present complementary, but not redundant information.

Halo and Common Method Variance

A prevalent concern of the questionnaire methodology used in this research is the possibility of finding spurious relationships between variables as a result of halo and common method variance (Campbell & Fiske, 1959; Remmers, 1934; Wherry, 1951). In the current research, that concern

would seem to be most pronounced with the regression results for the Outcome Interest and Learning Progress scales in the Regular and Speech samples. However, the use of multiple regression provides somewhat of a safeguard against accepting such spurious relationships.

Assume for the moment that each of three scale ratings are influenced to a large extent by a halo or common method factor. Also assume that the scales would be uncorrelated except for this common factor. Therefore, the correlation between any two of the three scales would equal the product of the correlations between each scale and the common factor. The more this common factor contaminates the ratings, the more highly correlated the scales will be. However, covarying on one of the scales when examining the relationships between the two remaining scales, to a certain extent, provides a control for the contaminating effects of the common error factor. The more highly contaminated the scores are, the more effective this technique becomes. For example, if 64 percent of the variance in each scale results from the common factor, the correlation between any two scales is .64. However each of the part correlations, which indicates the increase in the multiple correlation after the covariate has been accounted for, would be .29. Thus, the part correlations, while not accurate, provide estimates that are closer to the true relationships of .00.

This principle applies to the results for Outcome Interest and for Learning Progress. In each case, the effects of Initial Interest were partialled out of the criterion. The inclusion of Initial Interest in the regression equations reduced the chances of finding high correlations between CLIC scales and either criterion which were spurious.

Rating Scale Characteristics

Intraindividual ratings scales. In the introductory chapter the potential consequences of using intraindividual comparison rating scales for students' self-reports were discussed. Results of this study seem to support the argument that the intraindividual scale can control for the contaminating influence of student ability and previous performance. Neither SAT nor GPA were correlated with Initial Interest, Outcome Interest or Learning Progress, the three scales rated on the intraindividual basis. It was also suggested in Chapter 1 that the intraindividual scale might not control for students' interest or motivation to take the course, another contaminating factor when assessing instructional effectiveness via student achievement. This study also supports that contention. Students' Initial Interest and Learning Progress were significantly correlated. In addition, Outcome Interest was also highly correlated with Initial Interest.

Descriptive versus evaluative rating scales. Chapter 1 also discussed the theoretical distinctions between the use and interpretation of descriptive versus evaluation ratings scales. It was suggested that descriptive scales would be less susceptible to idiosyncrasies of students and therefore less ambiguous than evaluative scales. The results of this research cast some doubt on that argument. The CLIC and supplementary questionnaire contained both descriptive and evaluative rating scales. There was no discernable difference between the interrater reliabilities of these scales, suggesting that the students were in as much agreement on their evaluations as they were on their descriptions. Furthermore, the scales most highly related to the learning criteria (Satisfaction/Teacher in Regular and Overall Satisfaction in Speech, and Math) were evaluative, not descriptive scales.

These results are contrary to expectations and to previous research regarding the distinctions between individual perceptions and evaluations (Schneider & Snyder, 1975). However, the discussion of Chapter I offers an explanation. Recall that a descriptive scale requires students to simply record their perceptions. An evaluative scale requires students to judge their perceptions in terms of some standard or criterion and then record their evaluations. It is expected that students within a class should provide similar descriptions since their perceptions should be similar. On the other hand, to the extent that they judge their perceptions using different standards, their evaluations will vary. Note however, if students use the same standard, their evaluations will be in just as much agreement as their perceptions. It is only when raters' value standards vary that one should expect descriptions and evaluations to behave differently. The results of this study suggest that students were in agreement concerning the standards they used to judge an instructor.

Invariance of student value standards can not explain why the evaluation scales tended to dominate relationships with outcome criteria. The relatively stronger relationships between the evaluation scales and the outcome criteria perhaps can be explained best by the inclusiveness of the different scales. That is, ratings on Overall Satisfaction and Satisfaction/Teacher may have been influenced by a larger number of instructional factors than the more descriptive scales, and therefore accounted for more of the factors relevant to instructional effectiveness.

The suggestion that the evaluative scales may have more overall relevance, i.e., less deficiency, than the descriptive scales does not negate the usefulness of the descriptive scales. Their greater specificity makes them more useful than the overall evaluative scales for assessing the strengths and weaknesses of particular aspects of instruction.

Learning Progress Criterion

The results of this study and the results of Hartley and Hogan (1972), Hoyt (1973), and Solomon et al. (1964) all show relatively high correlations among learning progress ratings. The strength of these relationships among theoretically distinguishable learning categories presents no conceptual difficulty. As explained in Chapter 1, Gagne's (1970) eight types of learning are based on the assumption that the categories are correlated. The problem with correlated learning criteria emerges when one tries to interpret the relationships between process criteria, such as the CLIC scales, and learning outcome criteria. Statistically, there is no unique way to disentangle the correlations among the outcome criteria in such a way as to make unambiguous, unqualified statements about which process criteria are related to which outcome criteria.

One solution would be to make some assumptions about the order, in a path analysis sense, among various learning categories and then examine a series of multiple regression equations following the technique used in this study. Gagne's learning categories suggest one possible order and could be used to guide such an investigation. For this study, the items selected did not appear to present themselves in a way amenable to categorization. The items themselves suggested no unambiguous ordering. For example, it is not clear whether "Learning how professionals in this field go about the process of gaining new knowledge" would come before or after "Developing skill in expressing myself orally or in writing." The interitem correlations suggested no clear clusters of items. Therefore, in this investigation, it was assumed that the learning outcomes occurred simultaneously and that they were all equally important for all classes. A single Learning Progress score, defined as the mean of the twelve learning items, was used as the self-reported learning criterion.

Causal Relationships

In Chapter 2, the methodological hazards of trying to support relationships between the actual characteristics referred to by the CLIC scales and the outcome criteria were discussed and the idea of attempting to support causal relationships was rejected. That decision was based on (1) a lack of empirical information concerning the temporal order of the events referred to by the CLIC scales and (2) the possible contamination of the CLIC ratings by students' perception of their achievement in the course. These are methodological constraints that should not be construed to mean that causal hypotheses are not appropriate. In fact, causal hypotheses about relationships between the activities described by the CLIC scales and instructional outcomes are necessary if CLIC feedback is to be used to assess performance in order to determine areas of instruction that might be improved.

It may be noted that much of the literature on student evaluation of faculty suggests that feedback from questionnaires like the CLIC be used to guide instructional improvement (e.g., Hoyt, Owens, & Growling, 1973; and Menges, 1973). While, in the most rigorous sense, causal conclusions are not appropriate, the results of the research into the validity of the CLIC suggest that using the CLIC as a diagnostic tool for improving performance may be a much more appropriate alternative than making changes without using any empirically derived guidelines.

Contamination in the CLIC Scales

The potential contamination of CLIC ratings is an important consideration regarding the relevance of the CLIC scales for evaluating instructional effectiveness.

As displayed in Table 7, GPA, SAT Math, and SAT Verbal were not signi-

ficantly correlated with any of the CLIC scales and thus do not appear to be sources of bias in the CLIC ratings. Expected grade, on the other hand, was correlated with Consideration, Critical Demands, Student Involvement, and Overall Satisfaction. It was suggested in Chapter 4 that whether these correlations constitute evidence of unwanted bias depends on the interpretation given to the variables. Thus, if the CLIC scales and Expected Grade are each valid estimates of instructional effectiveness (i.e., how much students learned), the correlations between Expected Grade and the CLIC scales may represent this common validity. If the interpretation is justified and if Learning Progress reflects how much students learned, then the partial correlations between Expected Grade with Consideration, Critical Demands, Student Involvement and Overall Satisfaction, holding Learning Progress constant, should be closer to zero than the zero order correlations. These partial correlations for Overall Satisfaction, Critical Demands, Consideration and Student Involvement were .11 (n.s.), -.19 ($p<.05$), .21 ($p<.05$), and .47 ($p<.01$) as compared to the zero order correlations of .25, -.28, .34, and .53. An adaptation of a test developed by O. J. Dunn, J. Neill, and L. Bundick cited in Darlington (1975) showed that the partial correlations were significantly closer to zero for Overall Satisfaction ($p<.01$), Critical Demands ($p<.01$), and Consideration ($p<.02$).[6] For Student Involvement, the difference was not significant at the traditional .05 level ($p<.08$). These results tend to support the hypothesis that the correlations between Expected Grade and the CLIC may be somewhat attributable to Learning Progress. It may also be noted that for the Speech classes, Expected Grade correlated signifi-

[6]The test and adaptation were suggested by Dr. Phillip Bobko, personal communication, March, 1976.

cantly with only one CLIC scale (Critical Demands); its correlation with Learning Progress was not significant. Thus, for the Speech classes, it could be argued that Expected Grade was not correlated with the other CLIC scales because it did not reflect students' estimates of how much they learned.

Size of the class correlated significantly with four CLIC scales, Consideration, Critical Demands, Student Involvement and Overall Satisfaction. Class Format correlated significantly with Consideration, Student Involvement, Coordination and Overall Satisfaction. This suggests that for future uses of the CLIC, norm tables might be constructed partitioning classes according to Size and Format. From the zero order correlations, it is suggested that norms for Critical Demands should be developed separately for large and small class, and that norms for Coordination should be developed separately for different formats. Consideration, Student Involvement, and Overall Satisfaction were each correlated significantly with both Format and Size. Using regression analysis, the significance of the partial correlations for Size independent of Format, and the partial correlations for Format independent of Size with Consideration, Student Involvement, and Overall Satisfaction were each tested. For Consideration and for Student Involvement, the partial correlations for both Size and Format were significant. This suggests that norms might be developed for each of the ten cells which result when classes are categorized by Size (five levels) and Format (two levels). The usefulness of this strategy should be verified on future samples. For Overall Satisfaction, neither partial correlation was significant. This suggests that norms need to be based on only one of these variables. It might be suggested that norms be based on Format since that variable has fewer

categories and therefore might provide more stable estimates with fewer total classes. Again, the utility of these norm categories and the need for new ones should be checked on a continuing basis.

Satisfaction/Person was also investigated as a source of bias in the CLIC ratings. In the Speech sample and in the Regular sample the results of Grush and Costin (1975) were replicated. That is, the partial correlations between the CLIC Knowledge and Skill scale (analogous to their Skill scale) and Satisfaction/Person, independent of the effects of Satisfaction/Teacher, were not significant. For Speech, none of the analogous partial correlations for any of the CLIC scales was significant. For the Regular sample however, the partial correlations for Consideration and for Critical Demands were significant. This is, independent of Satisfaction/Teacher, high Consideration and low Critical Demands scores were associated with high Satisfaction/Person.

On the other hand, the partial correlations between Knowledge and Skill, and Satisfaction/Teacher, with the effects of Satisfaction/Person removed, were significant in both the Regular and Speech samples. These results also replicate the findings of Grush and Costin (1975). In addition, for the Speech sample, the analogous partial correlations for Consideration, Critical Demands, and Overall Satisfaction were significant, and in the Regular classes, the partial correlations for Student Involvement and Overall Satisfaction were significant.

These results suggest that the relationships between Knowledge and Skill, Student Involvement, Coordination, and Overall Satisfaction with Satisfaction/Person are an artifact of their relationships with Satisfaction/Teacher, a situation which Bingham (1939) might term "valid" halo.

For Consideration and Critical Demands, particularly in the Regular classes, the results are not as clear. The significant partial correla-

tions for these two scales and Satisfaction/Person after the effects of Satisfaction/Teacher are removed might be construed as evidence of unwanted contamination in the Consideration and Critical Demands ratings. This, in turn, could have led to the hypothesis that the observed relationship between Consideration and Learning Progress was the result of common variance due to Satisfaction/Person. That hypothesis, however, is not supported by the data. After the effects of Initial Interest and Satisfaction/Teacher were accounted for, the partial correlation of Satisfaction/Person with Learning Progress was .00. Satisfaction/Person had no residual variance in common with Learning Progress, therefore the direct relationship between Consideration and Satisfaction/Person could not have affected the direct relationship between Consideration and Learning Progress. The correlations between Satisfaction/Person and the CLIC scales do not seem to provide evidence of invalidity for the CLIC scales.

Students' Initial Interest may also be a source of contamination in the CLIC scales. For the Regular sample, Initial Interest was significantly correlated with five of the six CLIC scales. Only Coordination was not significantly correlated with Initial Interest. These correlations suggest that CLIC feedback should be normed by students' Initial Interest. This suggestion is in agreement with the Kansas State University procedure (Hoyt, Owens, & Growling, 1973) which uses mean class level of students' desire to take the course as a norming variable.

Currently, the Initial Interest scale is not a regular part of the CLIC rating system. Information is collected from the instructors about the proportion of students who are majors versus non-majors in the course subject and the proportion of the students who are taking the course as

an elective versus a requirement. Under the hypothesis that these two items might be substituted for students' Initial Interest, the relationships between these items and Initial Interest was examined. The correlations were .75 ($p<.05$) between major/non-major and elective/requirement, .18 (n.s.) between major/non-major and Initial Interest, and .05 (n.s.) between elective/requirement and Initial Interest. Neither of these items would make appropriate substitutes for students' Initial Interest. For future uses of the CLIC, alternative methods for assessing students' Initial Interest should be considered, including the possibility of adding the Initial Interest scale to the CLIC. Residual scores could then be reported with the effects of Initial Interest removed from the ratings.

Validity of the Check-List of Instructional Characteristics

Evidence concerning the validity of the CLIC is presented in two ways. First, the discussion will focus on the outcome criteria, exploring the relationships of the CLIC scales to each of the criteria, and the similarities and differences in the results from each of the four research samples. Second, the discussion will focus on the CLIC scales, highlighting the validity evidence of each of them and comparing the results of this study to the generalizations made about the research reviewed in Chapter 1.

Table 16 presents a summary of the significant relationships between each of the CLIC scales and the outcome criteria for Regular, Speech, and Math classes. The asterisks indicate that the criterion being represented is the uncontaminated residual criterion as presented in the previous chapter. The partial correlations appear in parentheses. The underlining indicates which scale(s) entered the regression equation for each of the

Table 16

Validity Relationships for the CLIC Scales

Sample	Knowledge and Skill	Consideration	Critical Demands	Coordination	Student Involvement	Overall Satisfaction
Regular		OI* (.21)	OI* (-.21)		OI* (.70)	OI* (.44)
	LP* (.54)	LP* (.57)		LP* (.40)	LP* (.41)	LP* (.59)
Speech	OI** (.40)	OI** (.41)	OI** (-.39)		OI* (.46)	OI* (.42)
	LP* (.50)					LP* (.52)
Math	PERF* (.33)	PERF* (.40)		PERF* (.27)	PERF* (.36)	PERF* (.45)

Note: Asterisks indicate significant relationship independent of respective covariates. Partial correlations are in parentheses. Underlining indicates scales that entered the regression equations. LP = Learning Progress; OI = Outcome Interest; PERF = Performance.

criteria indicated.

Self-Reported Learning Progress

Self-reported Learning Progress was used as an outcome criterion in two samples, the Regular classes and the Speech discussion classes. It has been suggested (Hoyt, 1973; Doyle, 1973) that students' interest in a course may bias learning criteria as assessments of instructional effectiveness. This suggestion was supported by the relationship between Initial Interest and Learning Progress in both the Regular and Speech classes. As a result of the multiple regression procedure, the bias in Learning Progress attributable to Initial Interest was removed from the Learning Progress scale and the residual became the criterion.

Regular Classes. The significant direct relationships of Satisfaction/Teacher indicates that this variable accounted for variance in the residual of Learning Progress after the bias of Initial Interest was removed. In discussion classes, Consideration was also directly related to Learning Progress. If this had been the only regression analysis computed, there would be no basis for asserting that any of the CLIC scales, except Consideration, were of any utility for evaluating students' Learning Progress. However, four of the six CLIC scales were directly related to Satisfaction/Teacher. (See Figure 3, p. 66.) This might suggest that these four scales, Knowledge and Skill, Critical Demands, Consideration, and Overall Satisfaction, may have been indirectly related to Learning Progress. That is, their relationships with Learning Progress may have been mediated by Satisfaction/Teacher.

This contention is partially supported by the significant partial correlations between three of these four CLIC scales and Learning Progress, with Initial Interest held constant. Knowledge and Skill, Consideration,

and Overall Satisfaction were all significantly correlated with Learning Progress, independent of Initial Interest. Critical Demands was the only one of the four scales directly related to Satisfaction/Teacher that was not related to Learning Progress when Satisfaction/Teacher was not considered.

In addition, the analogous partial correlations for the remaining two CLIC scales, Student Involvement and Coordination, were also significant. That is, five of the six CLIC scales were significantly related to Learning Progress after the effects of Initial Interest were removed.

Because the CLIC scales are correlated, one would expect that not all of the scales were providing unique information concerning students' self-perceived learning progress. Figure 4, p. 75 illustrates the regression equation for Learning Progress when Satisfaction/Teacher and Satisfaction/Person were excluded from the analysis. Initial Interest and the interaction of Format and Consideration again showed direct relationships with Learning Progress. In addition, Student Involvement and Knowledge and Skill were directly related to Learning Progress. The total variance accounted for by these two CLIC scales was the same as Satisfaction/Teacher. Even though Initial Interest accounted for 48 percent of the variance on Learning Progress, the CLIC scales, along with Format, accounted for an additional 28 percent of the variance. That 28 percent of variance represents 54 percent of the variance in Learning Progress not accounted for by Initial Interest.

Speech Classes. In Speech, Initial Interest also accounted for a significant proportion of the variance in Learning Progress (31 percent). Overall Satisfaction was also directly related to Learning Progress, accounting for 26 percent of the residual variance in Learning Progress left unaccounted for by Initial Interest.

Knowledge and Skill, Critical Demands, and Consideration were directly related to Overall Satisfaction, which suggests that these three scales were related to Learning Progress through their relationships with Overall Satisfaction. However, if this were so, each of the partial correlations between the three scales and Learning Progress, with Initial Interest held constant, would have been significant. As reported, the partial correlation between Knowledge and Skill and Learning Progress, with Initial Interest held constant, was the only significant partial correlation other than Overall Satisfaction.

Student Involvement, which was related to Learning Progress in the Regular sample and which had a significant correlation with Learning Progress in this sample, was not related to Learning Progress after the effects of Initial Interest were accounted for. None of the other CLIC scales were significantly correlated with Student Involvement, so that for the Speech sample the zero order relationship between Student Involvement and learning appears to be a function of students' Initial Interest. It was noted earlier that the reduced variance for Student Involvement may have reduced the strength of its observable relationships with the other variables.

In general, Knowledge and Skill, and Overall Satisfaction received the strongest support for their relevance to student learning. However Consideration and Student Involvement, and to some extent Coordination, may also be related to student learning.

Student Performance

Student performance measures were used as criteria in three settings. Each setting was composed of sections of the same course, taught by different instructors, but using identical student performance assessments.

Speech. None of the CLIC scales were significantly related to the performance criterion in the Speech classes. If this measure is accepted as the most relevant criterion for assessing the instructional effectiveness of the Speech discussion instructors, it would seem that the CLIC has little validity. However, that conclusion may not be warranted. Only one variable, Expected Grade, was significantly correlated with the performance criterion. Expected Grade was not significantly correlated with Learning Progress; indeed, it correlated significantly with only two other measures, Critical Demands and GPA. This set of circumstances implies that while students assessed, with some degree of accuracy, their performance in terms of the course grading standards (i.e., they predicted the grades they would receive), they did not equate the grades with how much they learned. Perhaps the highly rated instructors in this setting were effective in teaching the students something that was not a part of the course performance criterion. A similar argument has been made by McKeachie, Lin, and Mann (1971). Also, in the Speech sections, the instructors were not completely responsible for the course performance of their students. The common lecture material may have been predominant in the performance criterion.

German. The German sample was quite small. Nevertheless, one of the CLIC scales, Consideration, was significantly correlated with the performance criterion. However, Consideration failed to maintain its significance when SAT Verbal was partialled out of the relationships between Consideration and Performance. The sample size of ten sections tends to restrict the conclusions that can be drawn from the results of this setting.

Math. The Math setting may represent the most significant test of the validity of the CLIC since instructors had full responsibility for

presenting course material, and the sample size was large enough to detect meaningful relationships.

Figure 6, p. 85) illustrates that SAT Math, Class size at the Final, and Overall Satisfaction were directly related to students' performance in Math. SAT Math was entered in the analysis as a covariate since it was considered as a source of variance in performance that was likely to bias the results. Size at the Final, on the other hand, was not expected to be a significant variable. Since it did enter the equation, it can also be interpreted as a covariate, i.e., a potential source of bias in the criterion. The results indicated that Overall Satisfaction and Performance were related, independent of any linear relationships with SAT Math or Class size at the Final. Furthermore, that same statement could be made about all of the other CLIC scales except Critical Demands. Each of the respective partial correlations with performance, with SAT Math and Class size at the Final held constant, were significant. However, these scales were providing redundant information concerning students' performance. Knowledge and Skill, Consideration, and Student Involvement combined to account for 83 percent of the variance in Overall Satisfaction. After Overall Satisfaction was included in the regression equation for performance, no other CLIC scale added significantly to the regression equation.

SAT Math and Size at the Final explained 30 percent of the variance in Performance. Overall Satisfaction accounted for another 14 percent which is 20 percent of the residual variance in Performance left unexplained by SAT Math and Size.

The results from the Regular and the Speech samples suggest that more of the variance in Performance might have been accounted for if the Initial Interest scale had been available for the Math sample. These results also

suggest that if the effects of Initial Interest could have been removed from Performance, the residual relationship between Overall Satisfaction and Performance may have been even larger.

It was indicated above that the significant relationship between Performance and Size of the class at the Final was unexpected. The relationship was indicated that performance was higher in larger classes. One explanation would be that the "effective instructors." i.e., those with the high performing classes, were also the instructors who retained their students in the class. Less effective instructors, on the other hand, had more students drop the course. One test of this hypothesis concerned the ratio of the number of students who took the final examination compared to the number who took the first test. This term was significantly correlated with Performance, although it did not enter the regression equation. The hypothesis could also be tested by examining the partial correlation between Performance and number of students who took the final, with number of students who took the first test held constant. That partial correlation (.37) was significant ($p<.001$). The significant partial correlation and the significant correlation for the ratio term both suggest that ability to retain students might be a characteristic of instructional effectiveness. None of the CLIC scales correlated significantly with any of the size indices and therefore provided no guidance for explaining the hypothesized ability to retain students.

An alternative explanation can also be offered. Perhaps SAT Math was not removing all of the bias in the criterion attributable to class ability differences. Unaccounted for class ability differences could have affected both retention rate and performance. Thus, Size of the Class at the Final and Performance may have been spuriously related due to their respective relationships with class ability.

Whichever explanation of the effect of Size at the Final is chosen, the interpretation of the CLIC scales remains unchanged. Each of the CLIC scales, except Critical Demands, was significantly related to performance after the effects of Size at the Final and SAT Math were accounted for. Furthermore, if Size of the Class at the Final is considered an alternative criterion and not a source of contamination in Performance, and its effects on Performance not removed, the same five CLIC retain their significant relationships with Performance.

It may also be noted that the CLIC scales related to the Math Performance criterion were the same scales related to the Learning Progress criterion in the Regular sample, thus, increasing the strength of the evidence for the validity of Knowledge and Skill, Consideration, Coordination, Student Involvement, and Overall Satisfaction as indicators of student learning.

Outcome Interest

Outcome Interest, available as a criterion in the Regular and the Speech classes, was also correlated with students' Initial Interest, and therefore contaminated as a criterion of instructional effectiveness. Again, the use of partial correlations and multiple regression effectively removed the contamination of Initial Interest. With Initial Interest partialled out, several of the CLIC scales were significantly related to Outcome Interest.

For the Regular classes, four scales, Consideration, Critical Demands, Student Involvement, and Overall Satisfaction, were related to the interest criterion. Although Initial Interest accounted for 62 percent of the variance in Outcome Interest, Student Involvement was able to account for an additional 19 percent, representing 50 percent of the residual variance

in Outcome Interest unaccounted for by Initial Interest.

In the Speech setting, five CLIC scales had either a significant first order partial correlation (Initial Interest held constant) or a significant second order partial correlation (Initial Interest and Student Involvement held constant) in the expected direction. With Initial Interest held constant, Student Involvement and Overall Satisfaction were significantly related to Outcome Interest. With Initial Interest and Student Involvement held constant, Knowledge and Skill, Consideration, and Critical Demands were significantly related to Outcome Interest. Holding Student Involvement constant seems to reduce the variance in Outcome Interest but has little effect on the variance of the other CLIC scales, thus increasing the proportion of the (residual) variance in Outcome Interest accounted for by the CLIC scales.

Consideration, Student Involvement and Initial Interest were the scales directly related to Outcome Interest in the Speech classes. Initial Interest accounted for 62 percent of the variance in Outcome Interest. Consideration and Student Involvement accounted for 37 percent of the residual variance left unexplained by Initial Interest.

These results summarized in Table 16, suggest that Student Involvement, Overall Satisfaction, and Consideration may be the scales most closely associated with Outcome Interest, and that Knowledge and Skill, and Critical Demands may also have some relevance to this criterion.

Student Involvement and Overall Satisfaction as Internal Outcome Criteria

Chapter 2 suggested that Student Involvement and Overall Satisfaction may themselves be considered as outcomes of instruction.

Student Involvement. The results for Student Involvement presented in Chapter 2 were replicated in the Math sample. That is, Consideration

and Coordination were directly related to Student Involvement. The results of the Regular classes suggest a modification to these findings. Initial Interest, a variable not included in the original analysis or in the Math analysis, was directly related to Student Involvement. Consideration and Coordination were directly related to Student Involvement in discussion courses and not in lecture courses.

In the Speech discussion classes, none of the CLIC scales were related to Student Involvement, possibly a result of the restricted range of Student Involvement in the Speech classes.

Overall Satisfaction. In the German, Math, and Regular classes, Knowledge and Skill, and Student Involvement were directly related to Overall Satisfaction, confirming the previous findings. In the Math sample, Consideration was also directly related to Overall Satisfaction. That relationship seems to concur with Sullivan and Skanes' (1974) hypothesis that consideration type behavior may be important in courses such as Math.

In the Speech setting, Knowledge and Skill again showed a direct relationship to Overall Satisfaction. Student Involvement and Overall Satisfaction were not correlated allowing Coordination and Critical Demands to show direct relationships with Overall Satisfaction.

In general, Knowledge and Skill, and Student Involvement seem to be the scales most highly associated with students' Overall Satisfaction.

The CLIC Scales, Validity and Previous Research

Although Knowledge and Skill has been interpreted as students' description of the instructor, and Overall Satisfaction interpreted as students' evaluation of the instructor, the two scales were highly correlated in all of the settings and appear to be supplying similar information about students' achievement. In all of the settings except German, each scale was

related to a learning criterion, and one scale could be traded for the other in the regression equation with little loss of information.

Knowledge and Skill, and Overall Satisfaction seem to belong to Kulik and Kulik's (1974) Skill category of instructor characteristics. The Skill category refers to overall teaching performance, emphasizing items concerned with stimulating and interesting presentations and clarity of explanations. Overall Satisfaction contains items referring to appealing and interesting presentations. Knowledge and Skill contains items relating clarity of presentation. The results of this study are congruent with past research presented in Chapter 1. Scales similar to Knowledge and Skill and Overall Satisfaction have been consistently related to the learning performance of students.

In contrast to the results for student learning, Knowledge and Skill, and Overall Satisfaction were not alike in the way they related to Outcome Interest. Overall Satisfaction was related to Outcome Interest in both the Speech and the Regular samples. Knowledge and Skill, however, was related to Outcome Interest only in the Speech sample and then only after both Initial Interest and Student Involvement were held constant. The results for Knowledge and Skill are somewhat in contrast to the results of Tobias and Hanlon (1975) who found a Skill dimension to be related to their interest criterion. Perhaps the difference in the results occurred because Knowledge and Skill does not contain items about how stimulating or interesting the class presentations were. Overall Satisfaction does contain such items and was consistently related to Outcome Interest.

Consideration, which might be classified in Kulik and Kulik's (1974) Rapport category, also shows consistency in its relationships to outcome criteria. In previous research, ratings of Rapport have not been shown to be consistently related to outcome criteria. Kulik and Kulik describe

the Rapport category as emphasizing "the instructor's empathy, concern for and interaction with students" (p. 52). Examination of scales classified by Kulik and Kulik as belonging to the Rapport category suggests that Rapport items are not necessarily directly related to the task of teaching students. For example, Hildebrand and Wilson's (1972) Instructor-Individual Student Interaction scale contains items such as "has a genuine interest in students," and "is valued for advice not directly related to the course." The CLIC Consideration items "let me know when I had performed well," and "allowed students to express their problems related to the course," also connote something about the instructor's interest in students and willingness to give advice. However, the CLIC items are more directly related to the task of instruction. Therefore, the Consideration scale may not be precisely congruent with the Rapport dimension, with the critical difference being in the task relevance of the items. That difference may be why Consideration appears to be a valid index of instructional effectiveness, related to both student learning and student interest.

Coordination would seem to be the CLIC scale most like Kulik and Kulik's Structure dimension. The Coordination scale is, however, more inclusive than the Structure dimension. That is, Coordination items consider the structure of the course as a whole, including readings, examinations, and class presentations. Past research has provided mixed support for the validity of the Structure dimension. On the other hand, Table 16 shows that the Coordination scale is likely to be a valid indicator of student learning. Only in the Speech setting was Coordination not related to a learning criterion, and the instructors in that setting were not responsible for course structure, which may account for the non-significant relationships for Coordination.

Critical Demands appears to have no relevance for evaluating the

criterion of student learning. With Initial Interest partialled out, Critical Demands was not related to Learning Progress in either the Speech or the Regular samples, nor was it related to the performance criterion in the Math setting.

On the other hand, Critical Demands may have some relevance for assessing students' outcome interest in the course contents. With Initial Interest held constant, Critical Demands and Outcome Interest were significantly correlated in the Regular sample. In the Speech sample, Critical Demands was significantly correlated with Outcome Interest when Student Involvement and Initial Interest were held constant. Although Critical Demands did not enter the regression equation in either sample, in the Speech sample the second order partial for Critical Demands was only slightly smaller than the second order partial for Consideration which did enter the equation.

Even though Critical Demands contains the item "demanded more than I could do," the overall tone of the scale is not congruent with the possible interpretation that Critical Demands belongs to Kulik and Kulik's Difficulty or work load category. Difficulty per se is not represented by the CLIC.

The Student Involvement scale is somewhat unique in that it does not refer directly to characteristics of the instructor. Since the items refer to student activities, Student Involvement is not represented among Kulik and Kulik's (1974) four instructor dimensions. However, from this study it does appear to be a useful scale. It was related to outcome interest in both the Speech and Regular settings, and it was correlated with the learning criteria in Regular and in Math.

Because Student Involvement does not refer to instructor activities, it is not susceptible to one of the criticisms that has been levied

against traditional ratings of the instructor. Traditional evaluative ratings of instructors implicitly make the assumption that there is one optimal strategy for the most effective instruction (Hoyt, 1973b). That is, an evaluative rating scale sets up a model of what the effective instructor does or should do. However, Hoyt (1973b), Menges (1973) and others have suggested that there may not be just one way to effectively teach students, and that the innovative instructor may be penalized by such student ratings for his/her non-traditional approach. Descriptive rating scales may reduce this problem to a certain extent because recipients of the ratings may evaluate them in terms of any special instructional techniques they might be using. Alternatively, interpretation of Student Involvement ratings do not necessarily depend on the style of instruction. That is, no matter what instructional approach is used to facilitate student involvement, ratings on Student Involvement could be used to evaluate an instructor's effectiveness with regard to both students' outcome interest and learning progress.

Note that the above criticism of the model approach for evaluative or descriptive ratings does not invalidate the use of such ratings. The results of this research, as well as much of that reviewed in Chapter 1, indicates that a model can be developed that does represent the approach that is typically the most effective. Otherwise, correlations between ratings of instructors and outcome criteria would tend to be zero. The criticism does suggest that ratings be interpreted carefully, particularly when non-traditional teaching methods are being evaluated.

Conclusions

The evidence presented by this study suggests that the Check-List of Instructional Characteristics, a student response questionnaire providing

descriptions of various aspects of instruction, provides a valid assessment of instructional effectiveness. That is, it is a relevant criterion for evaluating teaching and it can be used as a guide for improving instruction. Each of the CLIC scales was related to at least one of the outcome criteria in at least two of the samples. Overall Satisfaction, Knowledge and Skill, Consideration, and Student Involvement were highly consistent in their relationships with the various outcome criteria. Suggestions were made for controlling bias in the CLIC ratings.

In general, the validity of the CLIC is supported by the conclusions of Menges (1973), Costin et al. (1971) and others that student ratings can be a useful index of instructional effectiveness. Furthermore, the variety of academic departments sampled by this study (e.g., Psychology, Physical Education and Math) suggests that the validity of the CLIC may be quite general.

These positive results should not be interpreted to mean that CLIC ratings be used as the only assessment of instruction. Other types of ratings, such as the operational use of items similar to those included in the Learning Progress scale, and other methods, such as classroom visitation and video tape, can provide additional information. No assessment system is without error, and the use of several methods may reduce the possibility of making incorrect evaluations.

A number of questions about the CLIC were not answered by this research. Two of the most important questions seem to be: (1) Can diagnostic use of CLIC feedback improve instructional effectiveness? and (2) Can CLIC feedback be used to assess non-traditional instructional approaches? With regard to the first question, past research suggests that such feedback may be useful, but only to the extent that faculty actually apply the

feedback in a diagnostic manner (Centra, 1973; Pambookian, 1974). The answer to the second question is entirely open.

APPENDIX

CHECK-LIST OF INSTRUCTIONAL CHARACTERISTICS AND SUPPLEMENTARY SCALES

CHECK-LIST OF INSTRUCTIONAL CHARACTERISTICS

PLEASE NOTE: Your responses are anonymous and confidential. Your instructor will not receive individual responses; he/she will receive feedback in summary form only.

Your responses to this questionnaire will provide your instructor with descriptive information about your reactions to this course. Please mark your responses on the answer sheet provided. Use a number 2 pencil.

0 = A
1 = B
2 = C
3 = D
4 = E
5 = F
6 = G
7 = H
8 = I
9 = J

In the block labeled "your last name", write in the code for this course (e.g. ENGL 100). Blacken the letter boxes which match the letters. For the course number, blacken the letter boxes according to the code in the box to the right. For example, for 100, blacken B, A, A.

Each of the following phrases should be marked from a to e according to how accurately it describes the course. If an item is not applicable (e.g. you did not have any exams), do not respond to that item.

a	b	c	d	e
//	//	//	//	//
completely accurate description	generally accurate description	somewhat accurate description	generally inaccurate description	completely inaccurate description

THE INSTRUCTOR...

1. expressed his/her appreciation when one of us did a good job.
2. introduced many ideas during each class session.
3. used examples.
4. insisted that everything be done his/her way.
5. would "ride" the student who made a mistake.
6. invited criticism of the ideas he/she presented.
7. allowed students to express their problems related to the course.
8. asked students questions.
9. demanded more than I could do.
10. encouraged class discussion.
11. let me know when I had performed well.
12. had an extensive knowledge of the subject.
13. presented an outline of the course.
14. insisted students follow a standard way of doing things in every detail.
15. put suggestions that were made by students into operation.
16. knew the subject matter.
17. was well organized.
18. presented recent developments in the field.
19. expressed his/her displeasure with students who made mistakes.
20. was a very thorough lecturer.
21. explained how the topics in the course were related to each other.
22. was intolerant of students' mistakes.
23. criticized students in front of others.

THE COURSE

24. Readings and class presentations were identical.
25. Readings were necessary to understand class presentations.
26. Readings were understandable.
27. Readings were helpful.
28. Reading material could be substituted for the lectures.
29. Lectures were presented in an appealing way.
30. Lectures were easy to become interested in.
31. Examination questions could be anticipated.
32. Examinations let you know your strengths.
33. Examinations tested general ideas.

SELF-DESCRIPTION

34. I had sufficient background for the difficulty level of this course.
35. I was generally prepared for class.
36. I took an active part in class discussions.
37. I enjoyed every minute of the course.
38. I could answer other students' questions.
39. I asked questions in class.
40. I discussed the material outside of class.
41. I wanted to know more about the subject.
42. I was attentive in class.
43. I would recommend this course to anyone.

For the following items, describe this course in terms of other courses you have taken at this university using the following scale:

a	b	c	d	e
More than most other courses	More than many other courses	About the same as other courses	Less than many other courses	Less than most other courses

At the beginning of the semester...

44. how relevant did you expect this course to be to your educational objectives?
45. how relevant did you expect this course to be to your career aspirations?
46. how relevant did you expect this course to be to your personal growth?
47. how much did you want to take this course?
48. how interesting did you expect this course to be?
49. to what extent did you take this course only to fill a curriculum requirement?

* * * * *

50. If you were given the opportunity, how much would you like to take an advanced course in this subject?
51. If you could avoid taking further courses in this subject, how much would you like to avoid them?
52. How likely is it for you to read an available book or article related to the content of this course?
53. How much do you intend to use what you have learned in this course (e.g. in further courses, career, personal growth, recreation)?

Describe how much progress you have made in the following areas:

54. Gaining factual knowledge (terminology, classifications, methods, trends).
55. Learning fundamental principles, generalizations, or theories.
56. Learning to apply course material to improve rational thinking, problem-solving and decision making.
57. Developing specific skills, competencies and points of view needed by professionals in the field most closely related to this course.
58. Learning how professionals in this field go about the process of gaining new knowledge.
59. Developing creative capacities.
60. Developing a sense of personal responsibility (self-reliance, self-discipline).
61. Gaining a broader understanding and appreciation of intellectual-cultural activity (music, science, literature, etc.).
62. Developing skill in expressing myself orally or in writing.
63. Discovering the implications of the course material for understanding myself (interests, talents, values, etc.).
64. Developing specific skills, competencies and points of view that I can use later in life.
65. My overall learning progress.

Use the following scale to describe your instructor in terms of other instructors you have had at this university:

a	b	c	d	e
More than most other instructors	More than many other instructors	About the same as other instructors	Less than many other instructors	Less than most other instructors

66. How much do you like your instructor as a teacher?
67. How much would you like to be in another course taught by your instructor?
68. How much would you like to recommend this instructor to a friend of yours who had to take the same course?
69. How much do you like your instructor as a person?
70. How much would you like to work with your instructor on a project of yours not related to educational activities?
71. How much would you like to have your instructor as a friend?

* * * * * * * * *

72. What grade do you expect to receive in this course?
    a. A b. B c. C d. D e. F
73. What is your cumulative GPA? (Leave blank if you do not have a GPA.)
    a. 3.40-4.00 b. 2.80-3.39 c. 2.20-2.79 d. 1.60-2.19 e. below 1.60
74. What is your SAT Mathematics score?
    a. 650 or over b. 550-649 c. 450-549 d. 350-449 e. below 350
75. What is your SAT Verbal score?
    a. 600 or over b. 500-599 c. 400-499 d. 300-399 e. below 300
76. What is your sex?
    a. Female b. Male

BIBLIOGRAPHY

Bingham, W. V. Halo, invalid and valid. Journal of Applied Psychology, 1939, 23, 221-228.

Blalock, H. M., Jr. Causal inference in nonexperimental research. Chapel Hill, N.C.: University of North Carolina Press, 1964.

Blalock, H. M., Jr. Theory building and causal inferences. In H. M. Blalock, Jr., & A. B. Blalock (Eds.), Methodology in social research. New York: McGraw-Hill, 1968.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. Taxonomy of educational objectives. Handbook I: Cognitive Domain. New York: McKay, 1956.

Boudon, R. A new look at correlation analysis. In H. M. Blalock, Jr., & A. B. Blalock (Eds.), Methodology in social research. New York: McGraw-Hill, 1968.

Bryson, R. Teacher evaluations and student learning: A reexamination. The Journal of Educational Research, 1974, 68, 12-14.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Carroll, S. J., Jr., & Tosi, H. L. Management by objectives: Applications and research. New York: MacMillan, 1973.

Centra, J. A. Effectiveness of student feedback in modifying college instruction. Journal of Educational Psychology, 1973, 65, 395-401.

Cohen, S. H., & Berger, W. G. Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. Proceedings, 78th Annual Convention, APA, 1970, 605-606.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-535.

Cronbach, L. J. Beyond two disciplines of scientific psychology. American Psychologist, 1975, 30, 116-127.

DeWolf, V. A. Student ratings of instruction in postsecondary institutions: A comprehensive annotated bibliography of research reported since 1968. Volume I. University of Washington, Educational Assessment Center, 1974.

Doyle, K. O. Faculty evaluation: Some considerations and a model. In A. L. Sockloff (Ed.), Proceedings: The First Invitational Conference on faculty effectiveness as evaluated by students. Philadelphia: Measurement and Research Center, Temple University, 1973.

Doyle, K. O., & Whitely, S. E. Student ratings as criteria for effecting teaching. American Educational Research Journal, 1974, 11, 259-274.

Dunnette, M. D. A note on the criterion. Journal of Applied Psychology, 1963, 47, 251-254.

Feldhusen, J. F., & Starks, D. D. Bias in college students' ratings of instructors. College Student Survey, 1970, 4, 6-9.

Fleishman, E. A., Harris, E. F., & Burtt, H. E. Leadership and supervision in industry. Bureau of Educational Research, Report No. 33, The Ohio State University, 1955.

Frey, P. W. Comparative judgment scaling of student course ratings. American Educational Research Journal, 1973, 10, 149-154.

Frey, P. W. Student evaluation. Science, 1975, 187, 557-558.

Gagne, R. M. The conditions of learning. New York: Holt, Rinehart & Winston, 1970.

Gessner, P. K. Evaluation of instruction. Science, 1973, 180, 566-570.

Gessner, P. K. Student evaluation. Science, 1975, 187, 558-559.

Ghiselli, E. D. Dimensional problems of criteria. Journal of Applied Psychology, 1956, 40, 1-4.

Ghiselli, E. D. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Goldstein, I. L. Training: Program development and evaluation. Monterey, Calif.: Brooks/Cole, 1974.

Green, C. N. The reciprocal nature of influence between leader and subordinate. Journal of Applied Psychology, 1975, 60, 187-193.

Grush, J. E., & Costin, F. The student as consumer of the teaching process. American Educational Research Journal, 1975, 12, 55-66.

Guba, E. G. The failure of educational evaluation. Educational Technology, 1969, 9, 29-38.

Guion, R. M. Criterion measurement and personnel judgments. Personnel Psychology, 1961, 14, 141-149.

Hall, D. T. The effect of teacher-student congruence upon student learning in college classes. Journal of Educational Psychology, 1970, 61, 205-213.

Hanke, J. E., & Houston, S. R. Teacher and student perceptions as predictors of college teaching effectiveness. College Student Journal, 1972, 6 (1), 45-46.

Hartley, E. L., & Hogan, T. P. Some additional factors in student evaluation of courses. American Educational Research Journal, 1972, 9, 241-250.

Heise, D. R. Problems in path analysis and causal inference. In E. F. Borgatta (Ed.), Sociological methodology: 1969. San Francisco: Jossey-Bass, 1969.

Hildebrand, M., & Wilson, R. C. From effective university teaching and its evaluation. In K. E. Eble (Ed.), The recognition and evaluation of teaching. Washington, D.C.: The American Association of University Professors, 1971.

Hoffman, R. G. Development of the check-list of instructional characteristics: Progress report I. Unpublished manuscript, Department of Psychology, University of Maryland, 1975.

Hoyt, D. P. The Kansas State University program for assessing and improving instructional effectiveness. In A. L. Sockloff (Ed.), Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia: Measurement and Research Center, Temple University, 1973. (a)

Hoyt, D. P. Measurement of instructional effectiveness. Research in Higher Education, 1973, 1, 367-378. (b)

Hoyt, D. P., Owens, R. E., & Growling, T. Interpreting ratings in "Student reactions to instruction and courses--short form". Manhattan, Kansas: Office of Education Resources, Kansas State University, 1973.

Jiobu, R. M., & Pollis, C. A. Student evaluations of courses and instructors. The American Sociologist, 1971, 6, 317-321.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston, 1973.

Kirkpatrick, D. L. Techniques for evaluating training programs. Journal of the American Society of Training Directors, 1959, 13, 3-26.

Kulik, J. A., & Kulik, C. C. Student ratings of instruction. Teaching of Psychology, 1974, 1, 51-57.

Land, K. C. Principles of path analysis. In E. F. Borgotta (Ed.), Sociological methodology: 1969. San Francisco: Jossey-Bass, 1969.

Levinthal, C. F., Lansky, L. M., & Andrews, O. E. Student evaluations of teacher behaviors as estimations of real-ideal discrepancies: A critique of teacher rating methods. Journal of Educational Psychology, 1971, 62, 104-109.

Lindbom, T. R., & Osterberg, W. Evaluating the results of supervisory training. Personnel, 1954, 31, 224-228.

Locke, E. A. What is job satisfaction? Organizational Behavior and Human Performance, 1969, 4, 309-336.

Lowin, A., & Craig, J. R. The influence of level of performance on managerial style: An experimental object lesson in the ambiguity of correlational data. Organizational Behavior and Human Performance, 1968, 3, 441-458.

Mitchell, T. R., & Nebeker, D. M. Expectancy theory predictions of academic effort and performance. Journal of Applied Psychology, 1973, 57, 61-67.

McGuigan, F. J. Amount learned: An empirical basis for grading teachers and students. Teaching of Psychology, 1974, 1, 10-15.

McKeachie, W. J. Correlates of student ratings. In A. L. Sockloff (Ed.), Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia: Measurement and Research Center, Temple University, 1973.

McKeachie, W. J., Lin, Y., & Mann, W. Student ratings of teacher effectiveness: Validity studies. American Educational Research Journal, 1971, 8, 435-445.

McKeachie, W. J., & Solomon, D. Student ratings of instructors: A validity study. Journal of Educational Research, 1958, 51, 379-382.

Menges, R. J. Evaluating learning and teaching. New Directions for Higher Education, 1973, 4, 59-75.

Meyer, H. H., Kay, E., & French, J. R. P., Jr. Split roles in performance appraisal. Harvard Business Review, 1965, 43, 123-129.

Morsh, J. E., Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. The Journal of Educational Psychology, 1956, 47, 79-88.

Pambookian, H. S. Initial level of student evaluation of instruction as a source of influence on instructor change after feedback. Journal of Educational Psychology, 1974, 66, 52-56.

Pohlman, J. T. A description of teaching effectiveness as measured by students. Journal of Educational Measurement, 1975, 12 (1), 49-54.

Pohlman, J. T., & Beggs, D. L. A study of the validity of self-reported measures of academic growth. Journal of Educational Measurement, 1974, 11, 115-119.

Remmers, H. H. Reliability and halo effect of high school and college students' judgments of their teachers. Journal of Applied Psychology, 1934, 18, 619-630.

Rodin, M. Student evaluation. Science, 1975, 187, 555-557.

Rodin, M., & Rodin, B. Student evaluations of teachers. Science, 1975, 177, 1164-1166.

Seashore, S., & Yuchtman, E. Factorial analysis of organizational performance. Administrative Science Quarterly, 1967, 12, 377-395.

Schneider, B., & Snyder, R. A. Some relationships between job satisfaction and organizational climate. Journal of Applied Psychology, 1975, 60 (3), 318-328.

Sharon, A. T., & Bartlett, C. J. Effect of instructional conditions in producing leniency on two types of rating scales. Personnel Psychology, 1969, 22, 251-263.

Smock, H. R. Progress and problems in evaluating college courses and instruction. Presented at the Annual Meeting of the American Psychological Association, New Orleans, September 1974.

Sockloff, A. L. Instruments for student evaluation of faculty: Ideal and actual. In A. L. Sockloff (Ed.), Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia: Measurement and Research Center, Temple University, 1973.

Solomon, D., Rosenberg, L., & Bezdek, W. E. Teacher behavior and student learning. Journal of Educational Psychology, 1964, 55, 23-30.

Spaeth, J. L. Path analysis. In D. J. Amick, & H. J. Walberg (Eds.), Introductory multivariate analysis. Berkeley, Calif.: McCutchan Publishing Corporation, 1975.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Tobias, S., & Hanlon, R. Attitudes toward instructors, social desirability, and behavioral intentions. Journal of Educational Psychology, 1975, 67 (3), 405-408.

Tukey, J. W. Causation, regression, and path analysis. In O. Kempthorne, T. A. Bancroft, J. W. Gowen, & J. L. Lush, Statistics and mathematics in biology. New York: Hafner Publishing Co., 1954.

Turner, M. E., & Stevens, C. D. The regression analysis of causal paths. Biometrics, 1959, 15, 236-258.

Wainer, H. W. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Wanous, J. P., & Lawler, E. E. Measurement and meaning of job satisfaction. Journal of Applied Psychology, 1972, 56 (2), 95-105.

Weerts, R. R., & Whitney, D. R. The effect of student, course, and instructor characteristics on types of items used in student evaluation of instruction. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Weiss, R. S., & Rein, M. The evaluation of broad-aim programs: Experimental design, its difficulties and an alternative. Administrative Science Quarterly, 1970, 15, 97-109.

Wetrogen, L. I. Development of the course guide: A course evaluation questionnaire. Unpublished master's thesis, University of Maryland, 1970.

Wherry, R. J. The control of bias in rating VII: A theory of rating. PRB Report No. 922, Personnel Research Section, Department of Army, 1952.

Wright, S. Path coefficients and path regressions: Alternative or complementary concepts? Biometrics, 1960, 16, 189-202.

CURRICULUM VITAE

Name: Roger Gene Hoffman.

Permanent address: 9915 Good Luck Road, Apt. T-2, Seabrook, Maryland, 20801.

Degree and date to be conferred: Ph.D., 1976.

Date of birth: March 2, 1948.

Place of birth: Wichita, Kansas.

Secondary education: Wichita High School East, Wichita, Kansas, 1966.

Collegiate institutions attended	Dates	Degree	Date of Degree
Kansas State University	1966-1970	B.S.	1970
University of Maryland	1970-1971	M.A.	1974
University of Maryland	1973-1974	Ph.D.	1976
University of Maryland	1975-1976		

Major: Industrial/Organizational Psychology

Professional positions held:

Visiting Assistant Professor
Department of Psychology
University of Maryland
College Park, Maryland
August, 1976 -

Research Assistant
Office of the Dean for Undergraduate Studies
University of Maryland
College Park, Maryland
June, 1974 - August, 1976

Research Assistant
Department of Psychology
University of Maryland
College Park, Maryland
August, 1973 - June, 1974

Teaching Assistant
Department of Psychology
University of Maryland
College Park, Maryland
August, 1970 - June, 1971