4-1979

The Relationship between Intellectual Development and Ratings of Teaching Effectiveness

Carol Ann Norman

Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Teacher Education and Professional Development Commons

Recommended Citation


https://scholarworks.wmich.edu/dissertations/2711

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
THE RELATIONSHIP BETWEEN INTELLECTUAL DEVELOPMENT AND RATINGS OF TEACHING EFFECTIVENESS

by

Carol Ann Norman

A Dissertation Submitted to the Faculty of The Graduate College in partial fulfillment of the Degree of Doctor of Education

Western Michigan University Kalamazoo, Michigan April 1979
ACKNOWLEDGEMENTS

More than I believed could be possible, this study was for me a deeply satisfying experience. It gave me cause to ponder all over again questions I had asked months and years earlier concerning intellectual development, student ratings, and effective teaching. And it gave me an invaluable opportunity to meet and to think along with many individuals who questioned, who reasoned, who encouraged, and who helped me immeasurably at those critical moments along the way. I would like to acknowledge their help and support and in particular to express my gratitude to:

Dr. Mary Anne Bunda, teacher, dissertation advisor, and friend with a dynamite mind and a stunning ability to slice through trivia in search of essentials.

Dr. Dorothy McCuskey, teacher, academic advisor, and friend who knew when to nudge me along and when to let me linger awhile.

Drs. John Nangle and Leo Stine whose sense of the possible and of the practical helped me keep in mind the theoretical and empirical questions to be considered.

Dr. Carole Widick, a long-time friend whose work with the Perry model initially sparked my interest in the model and whose questions continually challenged my own research efforts.

Nancy Woods, friend and superb rater with a wonderful combination of intelligence, enthusiasm, insight, and humor.

Teachers in the College of General Studies who agreed to participate in the study and who helped me understand with greater compassion the personal and professional concerns of academic instruction.

Pam Fox who typed this manuscript with more than patient understanding.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Pat Henry and Barb Ruhs, personal friends who shared my joys and my sorrows.

My parents who long ago helped me learn that dreams translated into goals can become a reality.

To these people in particular I would like to say thank you.

Carol Ann Norman
INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in "sectioning" the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University Microfilms International
300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4EJ, ENGLAND

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
NORMAN, CAROL ANN
THE RELATIONSHIP BETWEEN INTELLECTUAL DEVELOPMENT AND RATINGS OF TEACHING EFFECTIVENESS.

WESTERN MICHIGAN UNIVERSITY, ED.D., 1979

© Copyright by
Carol Ann Norman
1979
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>STATEMENT OF THE PROBLEM</td>
</tr>
<tr>
<td>II</td>
<td>A LITERATURE REVIEW</td>
</tr>
<tr>
<td></td>
<td>Dimensions of Student Ratings</td>
</tr>
<tr>
<td></td>
<td>Reliability, Validity, and Usefulness of Student Ratings</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
</tr>
<tr>
<td></td>
<td>Validity</td>
</tr>
<tr>
<td></td>
<td>Instructional Usefulness of Student Ratings</td>
</tr>
<tr>
<td></td>
<td>Cognitive Developmental Models</td>
</tr>
<tr>
<td></td>
<td>An Elaboration of the Perry Model</td>
</tr>
<tr>
<td></td>
<td>Research on the Perry Model</td>
</tr>
<tr>
<td></td>
<td>Structural Properties</td>
</tr>
<tr>
<td></td>
<td>Instrumentation</td>
</tr>
<tr>
<td></td>
<td>Research Foci</td>
</tr>
<tr>
<td></td>
<td>Cross-Validation and Student Rating Research</td>
</tr>
<tr>
<td></td>
<td>Conclusion</td>
</tr>
<tr>
<td>III</td>
<td>DESIGN AND METHODOLOGY</td>
</tr>
<tr>
<td></td>
<td>Population and Sample Characteristics</td>
</tr>
<tr>
<td></td>
<td>The Intellectual Development Measure</td>
</tr>
<tr>
<td></td>
<td>Field Test</td>
</tr>
<tr>
<td></td>
<td>Knewi Administration</td>
</tr>
<tr>
<td></td>
<td>Knewi Training Materials</td>
</tr>
<tr>
<td></td>
<td>Rater Training Procedures</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary Characteristics of Four Cognitive Developmental Models</td>
</tr>
<tr>
<td>2</td>
<td>Perry Model of Intellectual Development: Positions and Representative Thought Structures</td>
</tr>
<tr>
<td>3</td>
<td>Summary of Analyses for Stability for Prediction</td>
</tr>
<tr>
<td>4</td>
<td>Summary of Student Responses to Data Collection Procedures</td>
</tr>
<tr>
<td>5</td>
<td>Sample Characteristics: Age, College Level, and Field of Study by Sex</td>
</tr>
<tr>
<td>6</td>
<td>Intercorrelations for Teaching Effectiveness Subscales</td>
</tr>
<tr>
<td>7</td>
<td>Mean Effectiveness Ratings by Scale and Groups</td>
</tr>
<tr>
<td>8</td>
<td>Coefficients Showing the Relationship Between Intellectual Development and Ratings of Teaching Effectiveness for Each Student Group</td>
</tr>
<tr>
<td>9</td>
<td>Double Cross-Validation Coefficients for Predictive Stability of Student Response Data Within Teachers</td>
</tr>
<tr>
<td>10</td>
<td>Correlation Coefficients for Predictive Stability of Student Response Data Across Teachers</td>
</tr>
<tr>
<td>11</td>
<td>Coefficients for Predictive Stability of Student Response Data Across Students</td>
</tr>
<tr>
<td>12</td>
<td>Multiple Correlation Coefficients for Predictor Variables</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74</td>
</tr>
</tbody>
</table>

Grouping of Students for Cross-Validation Analyses

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
CHAPTER I

STATEMENT OF THE PROBLEM

In higher education student ratings of teaching effectiveness have been systematically collected for over fifty years (Kulik, 1976; Miller, 1974). During that time student ratings have been linked with a range of interests, chief among them being the interest in improving teaching and ultimately in facilitating the learning process (Dressel, 1976; Eble, 1971; Genova et al., 1976).

The desire to use student ratings for instructional development purposes has prompted an extensive search for relationships that would increase knowledge of the teaching and learning process. Initially the search focused on the definition of effective teaching and on the traits comprising it. Later the search was enlarged to encompass student characteristics, particularly those characteristics believed to have practical and appropriate instructional application.

This study, which employs a cognitive stage model of learner development, attempts to examine the possibility of a relationship between student intellectual development and student ratings of teaching effectiveness in higher education. As such, the study continues the search for student characteristics that might be related to student ratings and that might some day help provide a complementary link between teaching and learning.

Assessing Effectiveness With Student Ratings

The assessment of effective teaching presumes that we know
what good teaching is and that we can identify the particulars that comprise it. Despite the vast collective experience teachers can draw upon and the existence of richly documented research efforts, the question of whether or not good teaching can be reduced to a set of essentials is still asked.

An array of arguments seems to suggest that good teaching defies precise or even clear description. The individuality of teachers and students, unique characteristics of course content and instructional methodology, multiple and sometimes disparate educational goals, the mysteries still shrouding the teaching-learning process, and the varying instructional, institutional, cultural, and historical contexts of teaching and learning—all have at one time or another been identified as impediments to the acquisition of stable and generalizable knowledge about effective teaching (Benezet, 1973; Cohen et al., 1973; Lucas, 1971; Turner, 1973).

The discussion of effective teaching, however, and the search for specifications continue (Eble, 1970, 1972; Higham, 1974; Hightet, 1950; McKeachie, 1965). Motivated by professional interest and often by academic necessity (Eble, 1970; Genova et al., 1976), individuals acting alone and in concert with others have researched a range of characteristics describing teaching performance and supportive activities. A major aim has been and continues to be the definition and description of effective teaching to underscore faculty developmental efforts in higher education (Dressel, 1976; Eble, 1973; Hildebrand et al., 1971).

The persistent question, "Can effective teaching be identified?" appears to have been answered in a fairly positive fashion. The
results of dozens of studies and the judgments of countless students and teachers have identified recurrent themes and characteristics of effective teaching. Whether informally generated or based on carefully designed studies and factor analytic research, dominant and similar factors have emerged. The factors include the teacher's command of subject and presentation skill, ability to structure or organize the class, ability to establish rapport with students as a group or as individuals, and ability to arouse and sustain student interest and motivation (Doyle, 1975; Eble, 1973; Kulik and McKeachie, 1975).

The dimensions of instruction noted above are those dimensions that students and teachers commonly identify as significant characteristics of effective teaching. Because of the recurrence of these characteristics, they are often incorporated into standardized student rating forms. The forms are usually comprised of between ten and forty statements, each describing a specific activity and each related to a more general effectiveness characteristic. Students respond to each statement by indicating the degree to which the statement applies to a given teacher. Student responses are then analyzed, and the results are typically reported to the teacher sometime after final grades are submitted.

The Uses of Student Ratings

The student rating form has been increasingly used as a source of information about teaching effectiveness (Gustad, 1961; Astin and Lee, 1966; Bejar, 1974). Easy to administer, quick to score,
and a relatively inexpensive way to gather student opinion systematically, the rating form can enable the teacher to gather much information about many facets of the instructional process. Ideally the teacher can use the rating results to identify instructional strengths and weaknesses and then can use this information in the development of more effective instructional strategies (Kulik, 1976).

The distance between the ideal and the real, however, has prompted many educational analysts to express discontent with the actual usefulness of student rating data. Some dissatisfaction involves practical difficulties with interpreting rating data and then applying the results in the classroom (Doye, 1975; Kulik, 1976; Kulik and McKeachie, 1975; Marsh et al., 1975). Others have argued that the greatest limitation is the lack of sufficient models to guide interpretation and to suggest both general and specific changes leading to improved teaching (Kerlinger, 1963; Mitzel, 1960; Ryans, 1960).

The problem with existing models stems in part from the conceptualization of the teaching-learning process. Three conceptual limitations are particularly noteworthy. First, student ratings typically focus on the teacher and on the degree to which certain personal qualities and instructional characteristics associated with effectiveness are present or absent. Such a focus ignores the idiosyncratic characteristics of teacher and students (Follman, 1975), the great variation in instructional objectives and methodologies (McNeil and Popham, 1973), and the contextual variations of teaching environments (Turner, 1973). Second,
the study of student ratings frequently ignores the great diversity among college students, despite the fact that student diversity is frequently observed and well documented (Feldman and Newcomb, 1969). For example, students are often studied as if they were homogeneous and responded to teaching in a uniform manner (Glaser, 1973; Warren, 1973). Third, when the question of learner differences is approached, the defined differences are frequently of a demographic nature and are tangentially useful to the teacher who wishes to understand why student so variously perceive teaching performance and how these perceptual differences can suggest new directions for instructional practices (Doyle and Whitely, 1974; Kulik and McKeachie, 1975).

Within this context of discontent, psychological theory has been viewed with interest. Cronbach (1957), Kerlinger (1963), and Messick (1970) have advocated the use of psychological developmental models as a potentially fruitful way to approach the study of effectiveness ratings. Of special interest to Cronbach (1957), Messick (1970), and Glaser (1973) were cognitive developmental models that defined individual differences in terms of learning theory and human development. They believed such models could more directly confront the complexity of the teaching-learning process and were potentially more amenable to our efforts to improve teaching. It is appropriate, therefore, to examine the basic components of cognitive developmental models before discussing the model used in this study.

Cognitive Developmental Theory

The field of cognitive development concerns intellectual growth
and development. Models classified as cognitive development do not present a single comprehensive profile of the individual. Rather, the models selectively focus on one aspect of development and contribute an important albeit a partial view of individual development.

Cognitive developmental models take an information-processing view of the individual. From this viewpoint the individual is seen as an interactionist with the world outside the self. The individual selectively sifts through stimuli, translates the stimuli into meaningful patterns, and uses these patterns as guides for behavior and problem-solving activity (Suedfeld, 1971).

Of major interest to cognitive developmental theorists is the way an individual processes information, that is, on the cognitive structures used to mediate, transform, and pattern input stimuli in some characteristic manner (Bieri, 1971). Bieri also noted that terms other than cognitive structures have been used to describe these internal structures: conceptual systems, control, personal constructs, plans, programs, schema, style, and system. Whatever semantic predilection a theorist might have, the cognitive developmental models assume the existence of a mediating structure and that this structure determines the way an individual translates and then interacts with external reality (Flavell, 1963; Harvey et al., 1961).

Cognitive theorists view development as a progression of forms through which an individual interprets experience. Progression consists of upward movement through sequentially related stages. Beginning with relatively simple thought structures, the models project a number of stages, each representing a qualitatively different and more complex structure; that is, each stage represents
a structure more differentiated and integrated than the structure of the preceding stage. Together the stages describe coherent movement toward the "highest stage" believed to represent the most advanced form through which to process information and to interpret experience. Most theorists presume irreversible development. In other words, structural changes evincing greater complexity cannot be dissolved. The underlying assumption of the models, however, does not presume that all individuals will progress through the stages at the same rate nor necessarily that all individuals will attain the highest stage.

The Perry Model

Beginning with Piaget, developmental models have focused primarily on developmental stages of infancy, childhood, adolescence, and young adulthood. Only recently have existing models been extended or new models been generated for the period of adulthood encompassing the college years. One such model was that of Perry (1968, 1970) who postulated a scheme of intellectual and ethical development during the college years. The model, an hypothesized continuum with nine positions, describes the stages through which students typically pass as they change from viewing the world in terms of right/wrong and black/white modes of thinking, which Perry describes as a dualistic orientation, to a relativistic mode where knowledge and truth are viewed contextually, where diversity is tolerated and appreciated, and where one's roles are colored by perceived options and perspectives.
Each of the nine positions in the Perry model represents a qualitatively different mode of thinking. Consequently, students at different stages of development are hypothesized to conceive of knowledge, learning, the teacher's role, and their own roles as learners in different ways.

The Perry Model and Student Ratings of Teaching Effectiveness

Perry's model of intellectual and ethical development suggests a way to study student ratings of teaching effectiveness. For example, in his studies Perry (1968, 1970) found that dualists and relativists responded differently to the same educational experiences. He also found that students in a transition from one stage to another stage experienced particular difficulty in their attempts to understand what teachers wanted, how students' performances were graded, and how to cope with their perceptions of increased diversity within the instructional environment. These transition stages were especially difficult instructional experiences for the student and presumably for the teacher as well.

The Perry studies suggest a rationale for differentiated student ratings of the same teacher. If students at different stages of development perceive "truth," the nature of knowledge, their roles as learners, and the roles of teachers in different ways, then do they also rate teaching effectiveness in different ways? Perry (1970) suggests that they do.

As it stands, the scheme may be of immediate solace to a teacher in that it explains on impersonal grounds how he can be so differently perceived by various students in the class. This solace can be of no mean value, in that it can free his thinking for a more differential address to students "where they are". (p. 210)
Formally, however, neither Perry nor researchers who followed him have explored the possibility of a relationship between student intellectual development according to the Perry model and formal student ratings of various dimensions of teaching effectiveness. To date, research based on the Perry model has addressed questions concerned primarily with the validity of the model and its applicability in different instructional and institutional contexts. Whether or not knowledge of developmental level can increase our understanding of student ratings in instructionally useful ways, however, remains a question largely unexplored. This study consequently represents an initial step in the direction of such an exploration.

Statement of the Problem

This study attempts to explore the possibility of a relationship between student intellectual development as described by Perry and student ratings of teaching effectiveness. The plan of the study entails gathering data on six variables. The first variable, a measure of intellectual development, will be treated as the criterion variable. The remaining five variables, each regarded as a general characteristic of teaching effectiveness, will be treated as predictor variables. The statistical procedures will involve multiple regression analysis and the use of double cross-validity analyses. Both will be used in an attempt to answer two research questions:
1. Does knowledge of the way a student rates teaching effectiveness enable one to predict the student's stage of intellectual development according to the Perry model, and

2. If predictive power exists, how stable is that power?
CHAPTER II
A LITERATURE REVIEW

This chapter presents a review of literature related to the study of student ratings of teaching effectiveness and student cognitive development during the college years. The review begins with a discussion of dimensions of effective teaching frequently incorporated into student rating forms and is followed by a discussion of student ratings in relationship to technical and practical criteria identified as reliability, validity, and usefulness. Cognitive developmental models applicable to the study of college students are described. The Perry model used in this study and research related to the model are reviewed. The chapter concludes with a brief discussion of cross-validation.

Dimensions of Student Ratings

Interest in teaching effectiveness is no recent development. The subject has been discussed for many years, and materials allegedly describing effective teaching are numerous. The unfamiliar twist in the emergent interest in the teaching function, especially with respect to student ratings and standardized procedures for gathering such data, is the search not for characteristics representing judgments of individuals and committees but rather for objective characterization of effective teaching.

The search for objective characteristics has prompted an array of questions. What teaching qualities are students capable of
discerning? Are teachers best rated along one dimension or several, or are there many? Can student and faculty perceptions of teaching quality significantly converge so as to render ratings meaningful? Such questions have stimulated a great number of careful and scholarly investigations and are often referred to the court of factor analysis for judgment.

Factor analytic research has played an important role in the generation of effectiveness characteristics (Genova et al., 1976; Hildebrand, 1971, Kulik and McKeachie, 1975). Factor analytic research on student rating items is a way of attributing meaning to those items found to be statistically related. Because factor analytic research has been so important to teaching effectiveness research, this review will survey only one area—effectiveness of the teaching process itself—and will be restricted to those studies in which factor analytic techniques were used to isolate dimensions appropriate for student ratings of teaching effectiveness.

One of the first factor analyses of student ratings was conducted by Smalzreid and Remmers (1943). Using the ten-item Purdue Rating Scale, the researchers found two factors and labeled them Empathy and Professional Maturity. Later studies by Craeger (1950) and Bendig (1954) confirmed these factors but assigned them somewhat different lables—for example, Bendig preferred the label Instructional Competencies rather than Professional Maturity and chose the more specific label of Instructor Empathy for the first factor.

One of the most influential factor analytic studies was conducted by Isaacson et al. (1964). Using a 145-item instrument, they factor
analyzed the items, reduced the number to 46 and ultimately identified six stable dimensions appearing across different students and teachers over different semesters. The dimensions appeared with a high degree of consistency and with high factor similarity coefficients (.30 or more with few item exceptions). They labeled these dimensions Skill, Rapport, Structure, Overload, Feedback, and Interaction. The first two factors correspond to those identified by Smalzreid and Remmers (1943). The remaining factors were consistent with those identified by Gibb (1955), who used a similarly large pool of items, and by Cosgrove (1959), Turner et al. (1969), and Frey (1973).

Later factor analytic studies investigated dimensions of effectiveness appropriate for student ratings and also examined the importance of dimensions as perceived by students and faculty. Using such an approach, Hildebrand, Wilson, and Dienst (1971) based their study on 91 items rated as important by faculty and students and later factor analyzed into five dimensions. The researchers labeled the dimensions Analytic/Synthetic Approach, Organization/Clarity, Instructor-Group Interaction, Instructor-Individual Student Interaction, and Dynamism/Enthusiasm.

Using an approach similar to that employed by Hildebrand et al. (1971), Doyle and Whitely (1974) identified dimensions closely corresponding with the Hildebrand study. They labeled the dimensions Expositional Skill, Generalization of Course Content, Attitudes Towards Students, Motivation of Interest in Subject, and Stimulation of Ideas and Interest.

Factor analytic studies have made several noteworthy contributions to the study of effective teaching. First, the studies support the
argument that teaching effectiveness is neither so abstract nor so mysteriously complex as to defy knowledge of the subject. Indeed, whether one looks for general qualities or specific characteristics, the studies provide reasonably consistent answers to the question of what constitutes effective teaching (Eble, 1971). Second, for all the diversity attributed to teaching effectiveness, factor analytic studies using a variety of instruments with different students and faculty at different institutions and at different historical moments consistently reveal a fairly small and stable set of dimensions which Kulik and McKeachie (1975) have summarized under the labels of Skill, Rapport, Structure, Group Interaction, and Difficulty. Third, even when studies consider separately student and faculty perceptions of effective teaching, there is still agreement upon the attributes that comprise effective teaching (Doyle and Whitely, 1974; Hildebrand et al., 1971).

One instrument that incorporates the stable dimensions identified by Kulik and McKeachie (1975) is the Teacher Description Questionnaire, a student rating instrument developed by Hildebrand et al. (1971) and the instrument used in this study. The Teacher Description Questionnaire includes 36 items factor analyzed into five scales. No item is scored in more than one scale, and each scale represents a distinct dimension of effectiveness allegedly discriminating between good and poor teaching. The conceptual interpretation of these scales was offered by Hildebrand (1971) and is summarized below.

The first dimension concerns command of the subject. It is scholarship that integrates learning with adventure and is manifest
in the instructor's ability to analyze, demonstrate conceptual understanding, present the origin of ideas and concepts, differentiate between the implications of theories, and participate in the quest for knowledge. This dimension of effective teaching is labeled Analytic/Synthetic.

The second dimension concerns the ability to transmit instructional content. It involves the ability to be clear and as such involves the ability to cut through the elements of entertainment and sheer rhetoric. Clarity of presentation includes appropriately illustrative commentary, a meaningful progression of ideas, significant emphases, and well-timed summary. This dimension is called Organization/Clarity.

The third dimension of effective teaching focuses on rapport with the class as a group. It involves instructor sensitivity to group reactions both verbal and nonverbal and instructor capacity to learn if students are understanding course content and are interested. This dimension is labeled Instructor-Group Interaction.

The fourth dimension is labeled Instructor-Individual Student Interaction. Not restricted to the classroom setting, the component entails the instructor's response to the individual student. The response may be a simple greeting or may involve an extended conversation. Whatever the nature of the exchange, the instructor demonstrates the ability to respond to students as individuals.

The last dimension is labeled Dynamism/Enthusiasm. This dimension concerns those instructor qualities that spark interest, stimulate involvement, and invite student response.
These five dimensions are strongly related to the five dimensions identified by Kulik and McKeachie (1975). As such they are consistent with those characteristics that have been generally associated with effective teaching.

Reliability, Validity, and Usefulness of Student Ratings

Carefully gathered, student opinions rate high as a source of information about teaching effectiveness. Support for their widespread use stems from both theoretical and practical considerations. Theoretical considerations imply the need for firsthand rather than inferential information plus information from sources close to or at the point of learning. Practical considerations involve the need for information that can be systematically, quickly, and economically gathered and that can possess some degree of cross-comparability (Doyle, 1975). Carefully designed and wisely administered, student ratings can meet these specifications (Eble, 1972; McKeachie and Kulik, 1975; Miller, 1974).

Despite the theoretical and practical appeal of student ratings and the accumulated scholarship underlying many current rating forms, objections have been raised concerning the quality of information they can yield. Consequently, the need continually arises to describe the technical characteristics of student rating forms and thereby to indicate the quality of data such forms can provide. For the present, standards of quality will be discussed in terms of reliability, validity, and usefulness.
Reliability

Generally, reliability means consistency. In educational and psychological measurement, perfect reliability—perfect consistency—is a theoretical possibility but rarely if ever achieved in practice. More appropriate considerations, then, are not whether the data are perfectly consistent, but rather whether the data are consistent enough and of the right kind of consistency to serve a particular purpose.

Two kinds of reliability—internal consistency and stability over time—are general descriptive headings under which standard reliability procedures are classified. Internal consistency measures provide an index of the degree to which a set of items measures the same characteristic. When the set of items comprising the rating instrument purportedly measures the same characteristic, internal consistency is generally studied using the split-half method or the Kuder Richardson formula 20. Because many rating instruments are designed to measure multiple characteristics, however, computing a single estimate of internal consistency for the entire instrument is inappropriate. For such instruments, items measuring the same characteristics are generally clustered, and a statistic from a class of internal consistency statistics for non-dichotomous data is applied to each cluster rather than to the instrument as a whole. Examples of such statistics are Cronbach's alpha, the Horst measure, and the Hoyt measure. These procedures reduce the number of items statistically treated and, as a result, may cause the internal consistency of the clusters to drop or to range considerably (Aleamoni and Spencer, 1973; Doyle, 1975). Despite this disadvantage, the
following studies examining the internal consistency of clusters within a single instrument show that, in the sense of internal consistency, student ratings can be very reliable.

Remmers and Weisbrodt (1965) measured the internal consistency of the Purdue Rating Scale for Instructors. Using the Horst measure of internal consistency, they obtained reliability coefficients ranging from .67 to .91 on eleven ten-point rating scales. Hildebrand, Wilson, and Dienst (1971) studied the internal consistency of five scales comprising the Teacher Description Questionnaire. Using Cronbach's alpha as the measure of internal consistency, they obtained reliability coefficients ranging from .80 to .89. Doyle (1971) conducted an extensive study of the Minnesota Student Opinion Survey. The first phase of his study involved extended rating forms which, when analyzed using the Hoyt measure of internal consistency, yielded reliability coefficients ranging from .90 to .96. The second phase of the study involved a short form. Again using the Hoyt statistic, he obtained coefficients ranging from .61 to .92. Aleamoni and Spencer (1973) studied the Illinois Course Evaluation Questionnaire and recorded split-half reliabilities of .92 and .93.

The second operational form of reliability is stability over time. Whereas measures of internal consistency focus on the degree to which a set of items measures the same characteristic, measures of stability focus on the consistency of repeated measures over a period of time. In studies of student ratings, a rating instrument is given usually twice to the same students. The interval between the two administrations may be a matter of minutes or weeks, and the data gathered from the two administrations are correlated.
To the extent that the ratings are similar, the data are said to be stable.

Procedures for determining stability over time may involve the repetition of the same or of equivalent rating forms. Because of the difficulty and expense of constructing equivalent forms, the usual procedure is to administer the same form twice. This procedure is based on the frequently unsupported assumptions that the students have not been changed by the first administration of the rating forms and that the teacher rated has not changed during the interval.

The question of the stability of student ratings has been addressed in numerous studies and has been carefully examined in several comprehensive literature reviews. Costin, Greenough, and Menges (1971), for example, reviewed stability studies from 1954 through 1968. Later McNeil and Popham (1973), Doyle (1975), and Kulik and McKeachie (1975) conducted similar reviews. All concluded that, given proper instrumentation and administration procedures, student rating data can possess reasonably high degrees of stability.

Illustrative studies of the stability of student ratings include several studies conducted by Costin, Greenough, and Menges (1971). In one study student ratings made at mid-semester and again at the end of the semester were examined. For the five dimensions studied, they reported coefficients ranging from .70 to .84 on four dimensions and on the fifth dimension a coefficient of .58. They also reported a second study in which student ratings of four additional dimensions of teaching effectiveness were obtained. Ratings made six weeks before the end of the semester and again at the end of the semester yielded...
coefficients ranging from .67 to .77. Kohlan (1973) studied rating stability of four factors comprising the Nebraska Instructor Evaluation Questionnaire. Ratings made early in the semester and again at the end of the semester were correlated, and the obtained coefficients ranged from .55 to .70.

As mentioned earlier, the basic premise underlying stability measures is that the ratings are the same over a period of time. To the extent that the ratings are similar, they are said to be stable. Not all ratings should be stable, however. Raters can change, and teachers can change. Also, changes in ratings can be a function of random or systematic error attributable to the course, the rating task, or situational factors. Given these considerations, the studies reported above are only marginally adequate as teacher changes are uncontrolled and trait differences are largely unexamined. Although the above studies indicate a reasonable degree of stability, they do so within the confines of their limitations.

Both general literature reviews and specific studies provide considerable information about the reliability of student ratings. Enough evidence exists to support the position that students can rate classroom instruction with a reasonable degree of reliability. In particular, evidence exists to support the assertion that internal consistency and, within limitations, stability are demonstrable qualities of student rating data.

Validation

Generally, validity refers to the process of collecting evidence
which supports the attribution of meaning to data. As a process, validation can be considered incremental in the sense that efforts are continually made to study data and to attribute meaning to them. As an accomplishment, validation is the accumulation of both fact and judgment which, when applied to the data set, attempt to raise interpretive guesses to the level of knowledge.

Validity studies are classified by most major authors as content validity, predictive validity, concurrent validity, and construct validity. Some classifications, however, group both predictive and concurrent validity under the heading of criterion-oriented validity. Studies in this category are similar in that they involve obtaining two criterion measures. If the two criterion measures are obtained simultaneously, then concurrent validity is studied. If there is an interval between the two measures, then predictive validity is studied.

Content validity is established by showing that a statement or set of statements written for a rating form is a representative sample from a universe in which the researcher is interested. The deductive process is commonly used to establish content validity for rating forms. This process involves efforts to define a universe of rating statements and to select statements systematically from within this universe. The procedure generally involves discussion in which individuals and groups attempt to arrive at a judgmental agreement about the meaning of concepts and statements. This procedure is often used at the initial stage of instrument design and development (Doyle, 1975; Wilson et al., 1971) and is, according to Whitely (1975), the kind of validity most easily established for student ratings.
Few quality criterion-oriented studies of validity exist. One reason for this lack is the difficulty of finding ultimate criteria (Cronbach and Meehl, 1967), that is, a standard absolutely adequate for defining the quality measured. A second reason involves the difficulty of developing adequate composite or multiple criterion measures. A third obstacle arises from the growing sophistication of rating research and the recognition that many concepts comprising teaching effectiveness and the rating process itself are extremely complex.

One of the few examples of criterion-oriented validity is a study conducted by Elliott (1950). Elliott measured with four achievement tests the knowledge of recitation instructors and laboratory instructors in undergraduate chemistry classes. He then correlated student ratings of instructor knowledge with instructor scores on the achievement tests. He found a positive but non-significant .30 correlation for laboratory instructors and a positive and statistically significant .40 correlation for recitation instructors. Elliott concluded that students in recitation sections were significantly able to distinguish between instructors who knew more about chemistry and instructors who knew less. He did not, however, allow for the considerations that the recitation instructional mode might provide the instructor with greater and more frequent opportunity to articulate knowledge and that a correlation of .40, although statistically significant, explains little variance and thus may or may not be of educational concern.
A second noteworthy criterion-oriented study was conducted by McKeachie, Lin, and Mann (1971). The researchers studied the relationship between student ratings of teaching effectiveness in changing students' beliefs and actual semester changes in student scores on a test differentiating naive and sophisticated beliefs about economics. They found a positive correlation of .44.

By far the most extensive validation efforts have been made in the area of construct validity. Construct validity is studied when the investigator is interested in some postulated attribute or quality which is not "operationally defined" but which is assumed to be reflected in some performance measure (Cronbach and Meehl, 1967). Here, according to Cronbach and Meehl (1967), the investigator believes no available criterion can fully define the quality measured. Two procedures dominate construct studies involving student ratings of teaching effectiveness. The first procedure is factor analysis and has already been discussed under the heading "Dimensions of Student Ratings."

The second procedure involves an effort to study specific student attributes believed to be reflected in student ratings. The question addressed by these studies is whether students who differ with respect to a given trait render different ratings of teaching effectiveness. Among the student attributes most frequently studied are the so-called demographic variables. Information of this nature is frequently obtained on the face sheet of student rating forms and includes sex, age, major, and college year. Correlations of student ratings with these variables is seldom large and not
always significant (Kulik and McKeachie, 1975). When relationships are found, the influence of the characteristic on the rating is generally regarded as trivial (Doyle, 1975; Doyle and Whitely, 1974; Rayder, 1968).

Students' grades, either received or expected, are frequently the focus of construct validity studies. One set of reviews (Kent, 1967; McKeachie, 1975) examining studies of student grades and student ratings concludes that grades are not related to student ratings of teaching effectiveness. Another set of reviews (Centra, 1972; Doyle, 1975; Menges, 1973) reports inconsistent results with grades being positively related to ratings in some studies but unrelated in others. Feldman (1976), however, conducted a similar but more exhausting review in which he concentrated on research in which the individual student was the unit of analysis rather than the class. Feldman concluded that grade point average was non-significantly related to student ratings, that grades were strongly related to ratings in some classes but unrelated in others, and that a discrepancy between grade-point average and expected class grade showed some relationship to student ratings. Despite his inclination to believe that grades do influence student ratings, Feldman was unable to show convincingly that this was the case.

The relationship between student ability and student ratings has been studied frequently. Whereas in this study the interest is in level of development viewed in terms of cognitive structure, the usual focus is student ability viewed in terms of grade received or performance on an achievement test. Some researchers have found
no relationship between student ability and student ratings (Voeks and French, 1960; Doyle and Whitely, 1974). Other studies have found a relationship, sometimes positive, sometimes negative, and not always significant (Caffrey, 1969; Rayder, 1968; Treffinger and Feldhusen, 1970).

The relationship between student motivation and student ratings has been examined by several researchers (Cohen and Humphrey, 1960; Doyle and Whitely, 1974; Gage, 1960). The usual approach here is to compare ratings by students who were required to take a given course with ratings by students who were not. Results to date generally indicate little relationship between the required versus elective index of motivation and student ratings.

McKeachie and Solomon (1958) studied student ratings in connection with the instructor's ability to stimulate student interest in course content. The criterion, the number of students who elected advanced courses in the same subject area, was found to be positively and significantly related to student ratings. The findings, however, held for only two of five semesters studied.

Frey (1973) studied student ratings and student learning using multiple sections of an introductory and an advanced calculus course. Each course involved a common syllabus and a common examination. Frey found correlations ranging from .14 to .91 between mean ratings and adjusted examination scores for each of the six rating dimensions. For both courses, student ratings of student accomplishment correlated most highly with the learning measure. Ratings of teacher organization and teacher presentation were significantly related to learning in
the introductory course. Workload and grading procedures were more highly related in the advanced courses. The study demonstrated that some and not other ratings tend to reflect student learning.

Doyle and Whitely (1974) also examined the relationship between student ratings and student learning. Using twelve graduate student instructors, each in charge of a section of a course in beginning French, the researchers studied adjusted student scores on a common course examination and student ratings of effectiveness. Mean ratings of general teaching ability and over-all effectiveness were found to be significantly related to the learning measure.

Several studies have examined the relationship between student ratings and student perceptions of instructors. The first study by Day (1969) examined students' perceptions of teacher and self similarity and student ratings of teaching effectiveness. Day found a significantly positive relationship. In a related study, Davison (1973) examined student perceptions of the similarity between themselves and their teachers and the relationship between this perceived similarity and student ratings of teaching effectiveness. He reported that students who perceived teachers as superior to themselves gave higher ratings than did students who perceived the teacher to be similar to themselves. Davison concluded that student ratings are affected by student comparisons of themselves with their teachers.

Recently student rating research has focused on more complex interactions and their impact on student ratings. Although student sex generally has been found to have no consistent effect on ratings, McKeachie, Lin, and Mann (1971) found that male teachers with high
warmth were rated higher by female than by male students and that female teachers with high warmth were more effective with both female and male students than were female teachers with low warmth.

Haslett (1976a) studied the relationship between high school and college student self-evaluations along such dimensions as academic skill, interpersonal effectiveness, quality of scholastic involvement, and their ratings of teaching effectiveness. Among college students, Haslett found student self-assessment of academic ability to be a major variable strongly related to effectiveness ratings. In a second study, Haslett (1976b) investigated the interaction effect between college student knowledgeability, class size, or class level and student ratings of teaching effectiveness. She found significant interactions among the four main effects across the dimensions students used in evaluating instruction. She found that student knowledgeability and class size were strongly related to student ratings. She concluded that the more knowledgeable the student was in an area of study and the larger the class, the higher the ratings of instructors in that area.

Only recently have researchers begun to examine the complex interaction of learner modes of cognition described by dominant cognitive structures and student ratings of instruction. Studies by Hoyt (1969) and Talmadge and Shearer (1969) suggest that this may be a fruitful direction of further study. For the moment, cognitive characteristics theoretically function as internal frames of reference and can influence student ratings (Cohen, Trent, and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Rose, 1971; Feldhusen and Starks, 1970), but the subject has been insufficiently researched.

The problem—what constitutes effective teaching—has been approached through the conduct of validity studies. The above discussion and cited studies suggest extensive and repeated efforts to attribute meaning to student ratings of teaching effectiveness. The most evident meaning derives from judgmental discussion and factor analytic research. To some extent, meaning is derived from learning outcomes, and on a more abstract level, meaning is derived from studied relationships between variables believed to comprise the idea of teaching effectiveness.

Many researchers (Doyle, 1975; Kulik and McKeachie, 1975; McKeachie, Lin, and Mann, 1971) agree that precise definitions of the meaning of student ratings of teaching effectiveness are difficult and elusive. This seemingly tenuous state of affairs overshadows much work that has been accomplished and the continual emergence of new research questions. For the moment, agreement seems apparent that the meaning of student ratings is to some degree contextual and to a very large extent a subject of continued exploration (Doyle, 1975).

**Instructional Usefulness of Student Ratings**

The question of the usefulness of student ratings has been raised repeatedly. The question concerns essentially the benefits that can be derived from the collection of student ratings. Many benefits have been proposed, but the benefit consistently articulated
and widely supported is that of instructional improvement (Dressel, 1976; Eble, 1973; Genova et al., 1975; Kulik and McKeachie, 1975). The belief that student ratings can be used to develop and sustain levels of effective teaching has generated several research questions. For example, can evidence be found that will document the value of reporting student ratings to faculty members? Do faculty members who receive some form of student rating feedback improve their teaching? Do faculty members who are rated at mid-term and who receive feedback shortly thereafter receive significantly higher ratings than faculty who do not receive such feedback?

Most feedback studies employ some variation of a standard procedure involving the collection of rating data at early or mid-semester and again at the end of the semester. Shortly after the initial ratings are gathered, summaries of the data are reported to the instructors involved. Sometimes the rating summaries are provided with interpretive guides or with the option to receive interpretive consultation. Initial ratings and later ratings are statistically analyzed for differences which, if found, are attributed at least in part to the feedback effect.

Considered together, feedback studies show no consistently strong relationship between feedback and improved instruction. Feedback studies by Thomas (1969), Miller (1971) and Pambookian (1974) show no significant improvement in end-of-semester ratings. An often-cited study by Centra (1972) involves experimental and control groups of teachers. Centra found no significant improvement for teachers receiving feedback compared with teachers for whom no
feedback was provided. Centra noted, however, that a small subgroup of teachers did improve. For these teachers a substantial discrepancy existed between self-ratings and student ratings. Using these data, Centra concluded that feedback had value for improving instruction, a conclusion based more on conjecture than on empirical evidence. Without some verification of the self-rating data, the discrepancy could as easily have been attributed to a number of other factors, among them being deflated student ratings.

Some feedback studies reporting positive effects require a cautious view. Findings by Braunstein, Klein, and Pochla (1973) show a positive effect, but the improvement measure was a shift in class median ratings of at least one scale point, a measure which ignored other but less pronounced shifts. Aleamoni (1974) found positive differences between experimental groups of instructors who received feedback and control teachers who did not. Aleamoni did not verify the assumption that comparable students generated rating data under comparable instructional settings. Marsh, Fleiner, and Thomas (1975) found no performance differences on an achievement measure for students of experimental teachers receiving feedback and control teachers given no feedback. Positive differences were found in mean ratings on six of seven dimensions of teaching effectiveness and on three of four summary items. The positive differences obtained for the feedback group, however, were statistically significant for only two dimensions and one summary item.

One study by Overall and Marsh (1976) shows consistently positive feedback effects over a three-semester period. The researchers studied the effects of student rating feedback on
student performance on a final examination and on the end-of-semester student ratings during the fall, winter, and spring semesters of the 1973-74 academic year. The study involved students enrolled in an undergraduate introductory course in computer programming taught by graduate teaching assistants. Pretests at the beginning of each quarter, teaching rating forms at mid-quarter, and posttests and ratings at the end of each quarter were administered to all students. Instructors were randomly assigned to feedback and nonfeedback groups. The feedback group received feedback shortly after the mid-quarter rating and also had access to normative data. The researchers found that students of feedback instructors performed better on the achievement measure than did students of the nonfeedback instructors. Furthermore, the performance difference was statistically significant. The researchers also found that student ratings were more favorable for the feedback instructors and reached statistical significance for four of the seven rating dimensions—Instructor Concern, Learning, Instructor-Student Interaction, and Examinations—and reached statistical significance for the two summary items concerned with overall instructor rating and overall course rating. Overall and Marsh concluded that feedback was an instructionally useful mechanism both for effectiveness development and student learning.

To date, the literature concerned with the usefulness of student ratings has hinted at the instructional usefulness of ratings but has fallen short of identifying consistently strong relationships or of delineating substantive procedures for using rating data to improve teaching. But an appraisal of the feedback studies leads to an
encouraging observation. Although very little research has been reported concerning the developmental effect of student ratings on effectiveness, some evidence does exist to support the assertion that ratings coupled with some form of feedback might make a difference. The possibility that student ratings can improve teaching has encouraged several writers (Astin et al., 1974; Eckert and Steckline, 1961; Karman, 1969; Melnik and Sheehan, 1974) to propose faculty development programs incorporating student ratings, and numerous colleges and universities have initiated developmental projects. Since many of the programs have yet to be implemented or are in initial stages of operation, requests for findings are premature.

**Cognitive Developmental Models**

For a number of decades, researchers have been interested in cognitive differences among college students. The interest has spurred countless studies and studies of studies in search of descriptors for the college students and of the meaning of these descriptors in an educational context. Interest in the personality differences among college students is nothing new, but to conceptualize these differences as manifestations along a single continuum is a contemporary twist and one that has led to the generation of developmental models applicable to the college years.

Current interest in development and developmental models has several sources. Monolithic and reductionist systems proved too simple to account for student behavior and too narrow to encompass the great diversity within the college student population. Moreover, such systems were severely limited in their potential for translation.
into educational ideologies or improved instructional methodologies (Cronbach, 1957; Kerlinger, 1963; Kulik and McKeachie, 1975). A second stimulus probably came from the perceived educational importance of development and its relationship to teaching and learning (Knefelkamp, 1974; Widick, 1975). And a third stimulus no doubt arose from the intellectual challenge of the developmental conception (Loevinger, 1976).

Briefly, the idea of development is constructed on the metaphor of organism and uses structure as the key concept (Loevinger, 1976). Development refers to internal change, to a transformation taking place over time and implying an element of continuity. Structure implies central meaning attributed to elements and parts related to form a well-defined order. When the basic relationships change between elements and parts, the structure changes. Hence, development is conceived as the acquisition of new structures or as the transformation of old structures into new ones (Harvey, Hunt, and Schroder, 1961; Loevinger, 1976).

The view of development as a transformation of structures suggests the importance of Piaget, his study of cognitive development, and his espousal of structuralism as a point of view. Central to Piaget's (1960) developmental framework are such concepts as stage progression, an invariant developmental sequence, an hierarchical integration of stages, and irregular developmental rates. His epistemological framework for developmental psychology and contribution of key concepts are manifest in developmental models applicable to the college years. A review of cognitive developmental
models available for the study of college students, therefore, appropriately begins with Piaget whose study of intellectual development of young children and adolescents is recognized as a major contribution to developmental theory.

The primary focus of Piaget's work has been on cognitive development, especially as manifest in logico-mathematical operations. Piaget identified three major developmental epochs called periods. The periods of sensory-motor intelligence, preparation for and organization of concrete operations, and of formal operations are each elaborated with representative activities and cognitive structures which culminate in thinking described as hypothetical, deductive, and propositional.

A second model concerned with the structure of thought was proposed by Kohlberg (1964). Based on interviews with children who were presented stories posing moral dilemmas, Kohlberg identified six developmental stages and three levels of conceptual structure. The first or preconventional level was conceptualized as a set of rules and judgmental labels external to the child but anchored in a system of rewards and punishments. At midlevel moral value became identified with supporting conventional morality, in helping others, in maintaining law and order. The highest level of moral reasoning identified moral value in terms of contracts and principles of logic and ethical thought.

A third model, proposed by Hunt (1966) and based on the earlier work of Harvey, Hunt, and Schroder (1961), focused on the integrative complexity of conceptual structures. Based on studies involving
primarily young adults and college students, the four-stage model began with a low level of cognitive integration characterized by simple, concrete concepts, a low level of differentiation among stimuli, and absolute rules of concept integration. Theoretically, the more different ways information can be combined and interrelated, the higher the level of cognitive complexity. Thus, the model concluded with the level of high integration where concepts were abstract and where relativism was manifest in thinking.

A fourth model, presented by Perry (1968, 1970), chronicles the succession of forms through which college students construe their experiences, particularly those involving the nature and origin of knowledge, value, and responsibility. Emerging from studies of college students, Perry's nine-stage model concerned a range of cognitive structures beginning with those most dualistic and absolute and concluding with those most relativistic and contingent in nature.

Of the four models discussed above and summarized in Table 1, the model especially appropriate for the study of college students is the Perry model. The model was derived from data collected for a study conducted by Perry through Harvard's Bureau of Study Council. Perry and his colleagues began by studying the thinking process of college students over a four-year period. During the analysis of these thought processes, they began to detect a limited set of cognitive structures and what they believed was a developmental sequence of cognitive structures. To test this observation, Perry replicated the study. He found that the structures and their
sequential appearance confirmed his earlier findings. Thus, the identification of theoretical structures emerged from college student data; for, the ordering of structures was dictated by the logic of student thought, and the developmental sequence of the structure was defined by both theory and empirical study.

Table 1
Summary Characteristics of Four Cognitive Developmental Models

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of Stages</th>
<th>True Linear Development</th>
<th>Typical Age Application</th>
<th>Major Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piaget</td>
<td>3</td>
<td>No</td>
<td>Child</td>
<td>Cognitive Evolution</td>
</tr>
<tr>
<td>Kohlberg</td>
<td>6</td>
<td>Yes</td>
<td>Child-Adult</td>
<td>Moral Reasoning</td>
</tr>
<tr>
<td>Hunt</td>
<td>4</td>
<td>Yes</td>
<td>Adult</td>
<td>Integrative Complexity of Cognitive Structures</td>
</tr>
<tr>
<td>Perry</td>
<td>9</td>
<td>Yes</td>
<td>Adult</td>
<td>Transformation of Cognitive Structures</td>
</tr>
</tbody>
</table>

An Elaboration of the Perry Model

The Perry model presents nine positions or stages along an hypothesized continuum of intellectual development. Each position represents a qualitatively different way of thinking. Table 2 identifies the nine positions and describes the cognitive forms representative of each. As noted by the table, the nine positions may also be grouped into two general categories (Heffernan, 1975), each representing one of the two dominant structures of Perry's development model. Within these two general categories, the stages become variations of the dominant structure as it is transformed.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
from dualism to relativism. The first category called Dualism encompasses positions one, two, and three, positions from which students view learning as the acquisition of a set of truths. In this category, each stage represents an increasing awareness and assimilation of diversity, although diversity is not yet granted the status of legitimacy. The second category called Relativism encompasses positions four through nine, positions from which students view knowledge as being relative and view truth as being contextual. In this category, positions four and five represent the accommodation of diversity as a legitimate phenomenon, and positions six through nine represent the extension of the relativistic view into the personal realms of existence. No structural changes in cognition are evidenced in the Perry model after position five (Perry, 1968, 1970; Widick, 1974; Knefelkamp, 1975); hence, this study whose focus in on structural cognitive change, is restricted to positions one through five.

**Research on the Perry Model**

First reported in 1968 and later published as a book in 1970, the Perry model has generated interest as evidenced in research and developmental literature in higher education. Loevinger's (1976) study of the model led her to conclude that the model was a theoretical structure appropriate for the study of college students. Heffernan (1975) called the model unique in that it described student developmental processes through specific forms of thought and styles of establishing value and personal identity. Knefelkamp (1974) and
Table 2

Perry Model of Intellectual and Ethical Development:
Positions and Representative Thought Structures

<table>
<thead>
<tr>
<th>Intellectual Position</th>
<th>Cognitive Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>The student sees the world in polar terms of we-right-good versus they-wrong-bad. Right answers for everything exist in the absolute and are known to Authority whose role is to mediate or teach them.</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>The student sees diversity of opinion and uncertainty and accounts for them as unwarranted confusion in poorly qualified Authorities or as exercises set by Authority &quot;so we can learn to find the Answer for ourselves.&quot;</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>The student accepts diversity and uncertainty as extensive and legitimate but only temporary in areas where Authority has not found the Answer . . .yet.</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>The student sees legitimate uncertainty and diversity of opinion as extensive, and sees uncertainty in a realm of its own in which &quot;anyone has a right to her/his own opinion,&quot; a realm contrasted with Authority's where right-wrong still prevails.</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>The student perceives all knowledge and values including those of Authority's as contextual and relativistic. The student subordinates dualistic right/wrong functions to a special case in context.</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>The student apprehends the need to orient herself/himself in a relativistic world through some form of personal commitment (this is distinct from an unquestioned or unconsidered commitment to a simple belief in certainty or to a blind submission to Authority).</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>The student makes an initial commitment in some areas.</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td>The student experiences the implications of commitment and explores the subjective and stylistic issues of responsibility.</td>
</tr>
<tr>
<td><strong>9</strong></td>
<td>The student experiences the affirmation of identify among multiple responsibilities and realizes commitment as an ongoing, unfolding activity through which he/she expresses life style.</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
and Widick (1975) among others found the model an especially appropriate conceptualization of individual cognitive differences relevant to the study of teaching and learning in higher education. As an analytical framework, the model integrates knowledge from the areas of cognitive and personality development and is, according to Freedman (1973)

... the most original work on personality development of college students to appear since the field was founded twenty years ago. The book (Perry's 1970 publication) is an intellectual effort of the first caliber and has major implication for the art of teaching. (p. 117)

Structural Properties

Despite its intellectual and theoretical appeal, the Perry model has been criticized primarily for a theoretical shift in focus midway along the developmental scale. The shift occurs in Position 6 with the introduction of the construct of personal commitment. Position 5 in the model reflects relativism as a cognitive structure whose sequel is the emergence of personal commitment occurring in Position 6.

The alleged shift has been noted by Kurfiss (1975), Heffernan (1975), Kitchener (1976), and King (1977) among others. Kurfiss (1975) discussed the model as an integration of Piaget's theory of intellectual development and Erikson's (1963) theory of identity development. Heffernan (1975) reflected these two theoretical perspectives in his categorization of positions 1-5 as "ways of knowing" and positions 6-9 as "ways of being." Kitchener (1976) suggested two parallel rather than sequential constructs—intellectual development and identity formation—and King (1977) argued that the model excluded reflective thinking, a cognitive structure sequential
to relativistic thought. The researchers generally agree, however, that the first five stages of the Perry model concern primarily sequentially-related stages of cognitive development.

Several researchers have studied the structural properties of the model. Loevinger (1976) examined the model with respect to criteria for developmental models. Her criteria were that stages are fixation points defining types of people, that a structural stage conception must exist, that the logic of the stages should arise from the stages and their progression rather than from imposed personal preferences of the researcher, that specific tools and techniques should emerge for advancing knowledge of the model's domain, and that the conception should be applicable to people of all ages. She concluded that the Perry model satisfied the criteria for a developmental model insofar as it was restricted to the college student.

Kurfiss (1975) attempted to validate the structural properties of the Perry model. Specifically, she focused on the sequential, hierarchical, and stage unity properties of the model. Using a structured interview, Kurfiss (1975) studied comprehension scores from fourteen freshmen and fourteen juniors from a public university. The scores were analyzed according to the Guttman scale, and a scalogram analysis yielded a coefficient of reproducibility of .97 and a coefficient of scalability of .71. Kurfiss concluded that a sequence of increasingly complex structures did exist in the Perry model.

The question of hierarchy was approached through an analysis of preference scores derived from position statements ranked by
students from most to least convincing. Kurfiss did not specify which
correlational procedure she used to study the rankings, but she reported
some evidence to support the hierarchical assertion. Methodological and
linguistic problems with this portion of the study, however, greatly
weakened her conclusion.

The question of stage unity was approached by factor analyzing the
data. Using both oblique and orthogonal rotations, a weakness of the
study, she found evidence of stage unity, but the evidence was not strong.

A third study by Meyer (1977) examined Perry's assertion that the
model was content-free; that is, dominant cognitive structures could
be discerned through the articulation of different areas of experience.
Using a structured interview, Meyer studied the responses of twenty
freshmen and twenty seniors to stimuli statements of a religious nature.
Interview data were rated by trained individuals, and interrater reli­
ability estimates derived with the Pearson product-moment formula
yielded estimates ranging from .91 to .93. The reliability of the
sum for the three raters was .97 based on Cronbach's alpha. Meyer
concluded that the model was indeed content-free.

**Instrumentation**

In addition to the structural characteristics of the Perry model,
researchers have investigated several procedures for obtaining estimates
of Perry position scores for students. The first technique, the col­
lection of data through a structured interview, was used by Perry
(1970) to elicit student perceptions of their college experience.
Beginning with the question, "Why don't you start out with whatever
stands out for you about the year," Perry conducted a one-hour
interview with each student during the spring of the academic year. Individuals were then trained to determine Perry position scores based on complete typed transcripts of four-year student protocols, condensed four-year protocols, single interview transcripts, and excerpted statements from the four-year protocols. Interrater reliabilities for each rating procedure were .87 to .96, .74 to .77, .62 to .79 (corrected for the arbitrary rating of one judge), and .36 to .64 respectively. Perry concluded that judges could reliably agree on their application of the Perry model to student interview data regardless of the unit of analysis. He noted, however, that rating reliability tended to decrease as the units of analysis became shorter and less complete. Studies using the interview mode with high interrater reliability include Kurfiss (1975), Meyer (1977), and King (1977).

A second technique for deriving Perry position scores was the Perry Developmental Scheme (PDS), a ten-item self-rating instrument requiring students to rank-order the statements from most to least acceptable. Developed by Hartley (1973), the instrument has been found to possess theoretical, methodological, and linguistic problems which make its use of diminished value for research purposes.

A third technique involved the use of extended written student responses. Developed by Knefelkamp (1974) and Widick (1975) and referred to by others as the "Knewi," this instrument was designed in two parts. The first part was comprised of ten sentence stems, five to be scored using the Hunt model of cognitive integrative complexity and five to be scored according to the Perry model. The second part of the instrument was two one-page essays written
by students in response to topics generated by the researchers. Interrater reliabilities using the Pearson product-moment formula yielded coefficients ranging from .35 to .62 which were statistically significant but lower than the researchers had anticipated in light of extensive rater training. Using only the essay portion of the Knewi, Stephenson and Hunt (1975), however, obtained an interrater reliability of .82.

Research Foci

The Perry model has served as the analytical framework for several studies in higher education. Studies by Knefelkamp (1974) and Widick (1975) focused on the concept of developmental education, that is, instruction designed to promote student movement upward along the Perry scale. Using the Knewi described earlier, the researchers obtained pre-treatment and post-treatment position scores for each student. Position movement was reported for 90 percent of the students, but evidence to support the differential effect of the instructional intervention was not statistically significant.

Stephenson and Hunt (1977) replicated the Knefelkamp-Widick study and used two experimental and two control groups. The essay portion only of the Knewi was used to assess Perry position scores before and after instructional treatment. Using the scoring system reported by Knefelkamp (1974) to reflect within and between position movement, the researchers reported an average stage movement of .85 for the experimental groups receiving development instruction and
an average stage movement of .42 and .12 for the two control groups receiving no developmental instruction. The researchers concluded that developmental instruction produced more upward movement on the Perry scale than did other instructional approaches using similar course content. They did not, however, report whether or not the findings were statistically significant.

Touchton, Wertheimer, and Cornfeld (1977) used the Perry model to assess the impact of an instructional intervention program designed to stimulate cognitive growth in the area of career development. Six sections of a career planning course were used. Instructional content was the same for all sections, but the instructional approach varied from experimental (3 sections) to traditional (2 sections) and mixed (1 section). Based on calculations reported in the studies above, the researchers reported a mean stage movement of .59, .17, and .39 for the experimental, traditional, and mixed treatment groups respectively. The researchers concluded that developmental instruction produced more upward stage movement than did traditional or mixed methodologies. They did not, however, report whether or not their findings were statistically significant.

Unlike this study where only a possible relationship between student position on the Perry scale and student ratings of teaching effectiveness was the research focus, the three studies reported above used student movement along the Perry scale as the measure of teaching effectiveness. Two of the studies (Widick, 1975; Touchton et al., 1977) incorporated student questionnaires in which students rated various aspects of the teaching-learning process.
Student ratings in both studies were generally positive with few significant differences found between ratings given by students at different stages of cognitive development or by students receiving different instructional treatments. The instruments developed and used in these studies, however, were problematic in several respects. The Knefelkamp-Widick questionnaire had format problems which did not allow fine discriminations to be made between groups of students. The Touchton et al. satisfaction surveys were only vaguely described, and in neither study were the psychometric characteristics of the instruments reported.

Cross-Validation and Student Rating Research

Many studies, including this one, are concerned with the relationship between a criterion variable and two or more predictor variables. In such research multivariate analysis is a statistical procedure used which produces the maximum possible correlation and ultimately maximizes the efficiency of prediction for a particular sample.

The problem encountered with correlation coefficients derived through multivariate analysis is that the coefficient is a biased estimate of the population correlation coefficient (Ferguson, 1976). The estimate is biased in that the process of determining regression weights takes advantage of the peculiar characteristics; in essence, by taking advantage of chance. Thus, the multiple correlation coefficient tends to be inflated and biased.

Ferguson (1976) recommended a procedure known as cross-validation to reduce bias or at least to estimate the amount of bias in the
coefficient. With this procedure the multiple regression weights calculated for one sample are applied to a second sample. If the prediction equation derived on the first sample yields the same level of prediction in the second sample, then the investigator can be confident that the level of predictability is not due to chance. In essence, the level of predictability is validated.

Conclusion

This study was designed to explore the possibility of a relationship between student intellectual development and student ratings of teaching effectiveness for which two research questions were posed:

1. Does knowledge of the way a student rates teaching effectiveness enable one to predict the student's stage of intellectual development according to the Perry model, and

2. If predictive power exists, how stable is that power?

Chapter II developed several ideas relevant to these questions. First, over the years researchers have demonstrated considerable agreement over the dimensions of teaching effectiveness considered important by students and by teachers. Second, empirical research has documented that position that, given careful conceptualization and development, student rating instruments can yield data that are both reliable and valid. Third, research to date suggests that the student rating process involves complex interactions of many learner characteristics among which cognitive development is but an emergent subject of study. Fourth, the dominant purpose of student ratings appears to involve instructional development for which feedback studies only recently have begun to show some positive albeit not well understood relationships.
Fifth, a relatively new field, the study of cognitive development has generated several models applicable to the study of students during the college years. Sixth, one of the more useful models available for research involving the college student is the Perry model. And seventh, cross-validation procedures investigate the amount of bias in correlation coefficients derived through the use of multivariate analysis.
Chapter III describes the design and methodological detail of the study. Four major components of the design are discussed: the population and sample characteristics, the measures of intellectual development and teaching effectiveness on which the sample was tested, sample assignment procedures, and data analysis procedures.

Population and Sample Characteristics

The population for this study was comprised of undergraduate men and women attending Western Michigan University during winter semester, 1978. The students, whose major programs varied, were registered for classes offered in the College of General Studies. The college offered courses satisfying university general education requirements for students but offered no major or minor degree of its own. Students were unaware of the inclusion of their classes in the study prior to the fourth week of instruction.

Four criteria directed the selection of classes involved in the study. The first criterion was that the lecture be the dominant instructional mode in each class. This criterion focused attention on the lecture method which is the dominant instructional technique in higher education (Dubin and Taveggia, 1968) and also attempted to control for variance due to instructional technique. The second criterion was that each class be paired with a second class taught by the same teacher using the same instructional content and method. Classes paired accordingly were for the purpose of studying predictive
stability within teachers. The third criterion was that paired classes be of approximately equal size with no class having fewer than fifteen students. The search for classes of equal size was for the purpose of controlling the extent of selection bias in each class due to sample size. A class size of fifteen students would be large enough to fall within the typical range of fifteen to twenty students for winter semester classes in the college. Although classes larger than fifteen students were initially selected, allowance was made for the loss of some students who for various reasons—for example, poor attendance or refusal to participate in the study—would not be included in the sample. The last criterion was that teachers of selected classes be willing to participate in the study. Forced participation was antithetical to the ethics of the investigator, educational research in general, and departmental policy. Given the above criteria, six classes representing three different courses, each with two sections, comprised the sample of classes for this study.

The Intellectual Development Measure

Student intellectual development was measured with the essay portion of the Knefelkamp-Widick (Knewi) instrument. Comprised of two essays (I and II), with Essay I concerning a favorite class and Essay II concerning a decision making experience, the Knewi was field tested during fall semester, 1977, and administered for research and training purposes during winter semester, 1978. Student responses to the Knewi were scored by trained raters, and the derivation of
these scores comprised the intellectual development measures for each student. Procedures for field testing, Knewi administration for research and training purposes, rater training, and scoring are described below.

Field Test

Two Knewi administrative procedures were considered for research purposes, and each was field tested during fall semester, 1977. The first procedure, tried with two undergraduate classes, involved brief introductory remarks, the distribution of the Knewi, and the collection of the completed forms two weeks later. The advantages of this procedure were that students had considerable time during which to think about and then prepare their responses and that minimal instructional time was involved. The major disadvantage of this procedure was a response rate of 19% of the 59 students in attendance.

The second procedure, also tried with two undergraduate classes, involved brief introductory remarks, the distribution of the Knewi for in-class completion, and the collection thirty minutes later of the completed forms. The major advantage of this procedure was a response rate of 100% for the 61 students in attendance. Although the student responses generated with this procedure were on the average three sentences shorter than responses prepared over a two-week period, the quality and length of responses were nevertheless sufficient for scoring purposes. Furthermore, the thirty minutes of class time required for the administration was agreeable to the teachers involved in the study. Consequently, the second procedure
allowing a thirty-minute response period, yielding a response rate of 100%, and providing response quality sufficient for scoring purposes was chosen for the collection of student data during winter semester, 1978. The first procedure, which allowed for a two-week response period, was used for training purposes.

**Knewi Administration**

The administration of the Knewi for research purposes was conducted in early February, 1978, and proceeded in the following fashion. The instruments were distributed among the students, and introductory remarks concerned a general description of the nature of the research, a request for student participation, and a careful delineation of the students' task. All commentary was carefully prepared to avoid cuing student response patterns that might elicit the "guinea pig effect." If necessary, introductory comments were concluded with time allowed for a brief question-and-answer period before students began to write. At the conclusion of the timed writing period, the completed instruments were collected and given a three-digit code identifying the teacher, section, and student. For example, the code 2-2-10 referred to the second teacher, the second of two sections taught by that teacher, and the tenth student in the second section. In addition to identifying the data by teacher, section, and student, the code protected the identity of teachers and students. Administrative remarks and a copy of the instrument appear in Appendix A.

**Knewi Training Materials**

The general unavailability of Knewi materials for rater training
purposes necessitated the collection of an additional set of student responses for training purposes only. Therefore, for training purposes, the Knewi was administered to a second set of eight classes selected according to the same criteria used for the selection of the research sample of classes. The same criteria were used in an effort to obtain training materials comparable to those used for research purposes. The administrative procedure allowing for a two-week response period was chosen for this collection effort because its efficiency allowed entry into classes scheduled simultaneously and because minimal instructional time would be used.

The Knewi was administered to the eight selected classes in late January, 1978, and the completed responses were collected two weeks later. The response rate for this administration ranged from 0% in one class to 21% in another class. The total response yield from these eight classes was 112 completed responses for training purposes. Administrative remarks and the instrument used for this administration were identical to those used for the research sample with the only exception being that students were given two weeks rather than thirty minutes to respond to the instrument.

Rater Training Procedures

Student responses collected for training purposes were rated according to the first five positions along the Perry scale of intellectual development. Only the first five positions were considered for research purposes because no structural changes are evidenced after position five on the Perry scale (Perry, 1968, 1970; Knefelkamp, 1970).
1974; Widick, 1975) and because earlier studies involving the Perry model and undergraduate students in public universities have found few students in position five and no students in positions six or above. The responses were then reproduced and assembled unmarked into twelve training packages. The first two packages contained five student responses exemplifying each of the five intellectual stages considered for research purposes. The next six packages contained between five and seven responses, the majority of which exemplified a single intellectual stage. The remaining four packages contained between twelve and twenty responses representing the full range of intellectual stages. The majority of student responses collected in February, 1978, for training purposes were given a position 2 or 3 rating. In order to build complete sets of training materials representing positions 1-5, most student responses representing upper levels of development were taken from earlier un­published research conducted by Widick (1977) with graduate students at Ohio State University.

Beginning in late February, 1978, and continuing through May, 1978, twelve training sessions were planned for two raters, each of whom had completed a graduate course of study and had experience in the field of educational research. Prior to the first training sessions, the raters studied Perry's (1970) \textit{Forms of Intellectual and Ethical Development}. The first training session consisted of a discussion of the Perry model, an explanation of the research project, an examination of the Knewi, a review of prepared rating guidelines, and a discussion of sample student responses. The
second and each succeeding session involved rating sample student responses, discussing rating problems, clarifying rating cues, and reviewing specific tasks for the next session.

Ultimately eighteen rather than the originally planned twelve training sessions were held. Four of these additional sessions were held to address specific and reoccurring rating discrepancies. The remaining two sessions were for review and discussion purposes. A more detailed description of training materials and procedures appears in Appendix B.

**Scoring Procedures**

Upon completion of the training process, each rater was given a complete set of student responses gathered for research purposes. Each rater was asked to read all Essay I responses and to assign dominant and subdominant ratings to each and then to read all Essay II responses and assign dominant and subdominant position ratings to each.

The position rating was a two-digit code indicating a dominant and a subdominant position. A position was considered dominant if the majority of statements reflected cognitive structures associated with the position. A position was considered subdominant if between one-fourth and one-third of the statements reflected a given position. For example, a 2(3) rating represented a dominant position 2 and a subdominant position 3, and a 3(3) rating represented a response consistently at position 3.

The ratings assigned by each rater were combined to form a single rating for each student. Based on procedures developed by...
Knefelkamp (1974) and Widick (1975), the numerical value of dominant ratings was doubled and then combined with the numerical value of the subdominant rating to reflect the percentage of statement associated with each. For example, a 3(4) rating became 6+4=10, and a rating of 4(4) became 8+4=12. The ratings on the two essays were then combined and averaged. For example, the 3(4) and 4(4) ratings yielded an average position rating of 3.8. In other words, the combined numerical value of the ratings (22) was divided by the weighted number of dominant and subdominant positions (2 dominant + 1 subdominant + 2 dominant + 1 subdominant position scores for a total of 6). Consequently, the intellectual measure for each student was an averaged and weighted position score.

The Teacher Effectiveness Measure

Measures of teaching effectiveness were gathered through use of the Teacher Description Questionnaire. The questionnaire was comprised of five subscales, one each for the traits labeled Analytic/Synthetic (7 items), Organization/Clarity (7 items), Instructor-Group Interaction (8 items), Instructor-Individual Student Interaction (6 items), and Dynamism/Enthusiasm (7 items). Administrative and scoring procedures used with this instrument are described below.

Administrative Procedures

The Teacher Description Questionnaire was administered to students during the week preceding final examinations scheduled for winter, 1978, classes. Administrative procedures developed and standardized
by Hildebrand et al. (1971) and approved by the College of General Studies faculty were followed. The procedures involved the distribution and collection of instruments during a single class period.

Two departmentally-approved modifications were made to the administrative procedures. First, preliminary administrative remarks informed the students that rating data would be used for research purposes only and that the amount of praise or negative ratings was not the thrust of research interest. Rather, the important concern was ratings which were accurate reflections of the way students thought and felt. Second, students were asked to identify themselves by name, social security number, and major on cover sheets accompanying the rating forms. This information was for the purpose of matching student rating data with student questionnaire responses gathered in the semester. The students were told that their individual evaluations would be treated confidentially, that the forms would not be available for individual or collective teacher scrutiny, and that the identification was for research purposes only. Administrative commentary and a copy of the rating form appear in Appendix C.

**Scoring Procedures**

Students rated each of the thirty-five statements comprising the Teacher Description Questionnaire. The ratings were made on a seven-point, one-way scale with a rating of 1 representing unusually ineffective performance and a rating of 7 representing an unusually effective performance. Subscale rating scores for each student were obtained by averaging within subscale scores and were recorded
as mean subscale scores. The reliability of group subscales was calculated using Cronbach's alpha.

Sample Assignment

Six classes, two classes for each of the three teachers, comprised the sample from which student data were collected. For each student in each class, six data units were assembled: rating subscale means ($X_1$, $X_2$, $X_3$, $X_4$, and $X_5$) and the intellectual development measure ($Y$). The five subscale means were considered the predictor variables, and the measure of intellectual development was considered the criterion variable.

Because multiple predictor variables were employed to explain variation in the criterion variable, the data were analyzed by class using the following formula for multiple regression analysis

$$Y = a + b_1X_1 + b_2X_2 \ldots$$

where $Y$ = predicted intellectual development position

$a$ = the regression constant

each $b$ = the regression weight, and

each $X_i$ = standardized subscale mean rating for subscale $i$.

The assumptions underlying the use of multiple regression analysis for this study were that interval scale variables were involved and that each predictor variable, considered separately or in conjunction with other predictor variables in the set, could predict the criterion variable. Standardized mean rating scores were used to control for differences in variation of the predictor variables and also to allow for more interpretable weights.
When applied to a single group of student data, the above formula yielded a multiple correlation coefficient for each class. The coefficient was considered a measure of the power of prediction for a given group. Six multiple correlation coefficients, one for each class, were computed.

The procedures described above were used in an attempt to answer the first research question: does knowledge of the way students rate teaching effectiveness enable one to predict the students' stage of intellectual development?

Multiple correlation coefficients, computed for each of the six classes, provided estimates of the power of prediction. Because the computation took advantage of the idiosyncrasies of a particular class or a particular teacher, the estimates were biased and, as such, did not address the question of predictive stability within or between teachers. The question of stability, the second research question, was approached through the double cross-validation procedures described below.

For the double cross-validation, multiple regression equations were calculated separately for each of two groups in the sample and were then applied back on the other group. The resulting correlation coefficients were examined for the amount of shrinkage, and a determination was made as to whether or not the results indicated stability across the two observed samples.

Two double cross-validation analyses were conducted. The first involved deriving a regression equation for the first class \(A_1\) taught by a given teacher and then applying this equation to the
second class ($B_1$) taught by the same teacher. The reverse, deriving an equation for the second class ($B_1$) and applying it on the first class ($A_1$), was also conducted. The resulting correlation coefficients were used to examine the stability of prediction within teachers.

The second analysis involved calculating a multiple regression equation for the combined classes of a given teacher ($T_1$) and then applying this equation to the combined classes of each of the other two teachers. This procedure comprised the double cross-validation for the stability of prediction across teachers.

In the event that the two double cross-validation analyses provided little evidence of stability, a third analysis was planned. The procedure involved developing a regression equation for the combined first sections of all teachers ($S_A$) and again for the combined second sections of all teachers ($S_B$). The equation derived for each group was applied back on the other group. The procedure was used to examine predictive stability across student groups.

The analyses described above are summarized in Table 3.

Table 3
Summary of Analyses for Stability of Prediction

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Derivation Group</th>
<th>Cross-Validation Group</th>
<th>Estimate of Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$A_1$</td>
<td>$B_1$</td>
<td>Within Teacher</td>
</tr>
<tr>
<td>2</td>
<td>$B_1$</td>
<td>$A_1$</td>
<td>Within Teacher</td>
</tr>
<tr>
<td>3</td>
<td>$T_1$</td>
<td>$T_j, T_k$</td>
<td>Across Teachers</td>
</tr>
<tr>
<td>4</td>
<td>$T_j$</td>
<td>$T_1, T_k$</td>
<td>Across Teachers</td>
</tr>
<tr>
<td>5</td>
<td>$T_k$</td>
<td>$T_1, T_j$</td>
<td>Across Teachers</td>
</tr>
<tr>
<td>6</td>
<td>$S_A$</td>
<td>$S_B$</td>
<td>Across Student Groups</td>
</tr>
<tr>
<td>7</td>
<td>$S_B$</td>
<td>$S_A$</td>
<td>Across Student Groups</td>
</tr>
</tbody>
</table>
Summary of the Research Design

Six classes, two classes for each of three teachers, were involved in the study. Teachers were matched within a single college on the basis of instructional method. Classes were chosen on the basis of course content, college level, and size.

Data were collected on the following measures for each student:

1) Intellectual development, a written response to the essay portion of the Knefelkamp-Widick student questionnaire, and

2) Teaching effectiveness ratings using the Teacher Description Questionnaire.

The intellectual development measure was collected near the beginning of the semester, was rated by trained raters, and weighted mean ratings were derived for each student. Five teaching effectiveness measures were collected near the end of the semester, and mean subscale ratings were derived for each student.

Multiple regression analyses using standardized rating measures were conducted to yield multiple correlation coefficients for each class and for combined class groups. Double cross-validation analyses were designed to estimate the stability of prediction within and between teachers and across student groups.
CHAPTER IV
RESULTS OF THE STUDY

Chapter IV presents the results of the study which investigated the relationship among measures of student intellectual development and teaching effectiveness ratings. The chapter begins with a description of the student sample and presents the results of the analysis of measures of student intellectual development and teaching effectiveness ratings. The chapter concludes with a summary of the findings.

Sample Size and Characteristics

Students participating in the study did so on a voluntary basis. Of the 193 students initially enrolled in the six classes, 161 were contacted in class or by telephone and were invited to participate in the study. The 32 students not contacted in February, 1978, had either dropped the class (16), had not attended more than one of the first twelve consecutive class meetings (15), or were special in that they were enrolled in the class but were doing an independent study and did not attend class (1).

Of the 161 students contacted, 149 or 92% agreed to participate in the study. Those students who declined gave various reasons: not enough time to participate (5), viewed the study as an invasion of privacy (3), gave an unexplained refusal (3), or viewed their personal contribution to the study as insignificant (1).

Each of the 149 volunteer subjects was asked to complete the Knefelkamp-Widick Student Essays in early February, 1978, and the
Teacher Description Questionnaire in mid-April, 1978. A total of 136 or 91% of the volunteer subjects completed both instruments. Table 4 summarizes the number of student responses to each data collection procedure and the number of completed data sets collected for each group.

Table 4

Summary of Student Responses to Data Collection Procedures

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Student Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Potential Subjects</td>
</tr>
<tr>
<td>1</td>
<td>A 36 34 34 31 31</td>
</tr>
<tr>
<td>1</td>
<td>B 35 29 27 26 25</td>
</tr>
<tr>
<td>2</td>
<td>A 30 20 19 19 19</td>
</tr>
<tr>
<td>2</td>
<td>B 31 26 26 24 24</td>
</tr>
<tr>
<td>3</td>
<td>A 32 22 22 22 22</td>
</tr>
<tr>
<td>3</td>
<td>B 29 18 18 15 15</td>
</tr>
<tr>
<td>Total</td>
<td>193 149 146 137 136</td>
</tr>
</tbody>
</table>

Students participating in the study provided descriptive information concerning sex, age, college level, and major field of study. By sex, 60 women (44%) and 76 men (56%) participated in the study. The students ranged in age from 18 to 68 with most students (95% of the women and 72% of the men) being in the age range of 18 to 22. All undergraduate college levels were represented by both men and women. The largest number of women (25 or 42%) were freshmen, and the largest number of
men were either freshmen or seniors (23 and 22 or 29% and 30%, respectively). The students' major fields of study broadly represented business and administration, education, liberal arts, and the natural and social sciences. Table 5 presents by sex the age, college level, and major fields of study of the student sample.

Measure of Intellectual Development

The Knefelkamp-Widick instrument was used to measure intellectual development. Described above in Chapters II and III and presented in Appendix B, the instrument involved two essays, each designed to probe the structural aspect of cognition. This section describes the rating procedure, presents the interrater reliability obtained on the ratings of intellectual development, and discusses the treatment of discrepant ratings.

Rating Procedures

It was mentioned earlier that the Perry model presents no structural changes in intellectual development after position five. For this reason, the two judges independently rated according to the first five Perry positions each of the two essays written by each student. Two judges independently rated the two essays written by each subject. The judges rated all Essay I responses before rating all Essay II responses. The raters were not given any information about the sex, age, college level, or major of the subjects. During the rating of Essay II responses, the judges did not have access to their earlier ratings of Essay I responses.
Table 5
Sample Characteristics: Age, College Level, and Field of Study by Sex

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Characteristic</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>College Level</td>
<td>Freshman</td>
<td>Sophomore</td>
<td>Junior</td>
<td>Senior</td>
</tr>
<tr>
<td>Female</td>
<td>60</td>
<td></td>
<td>25</td>
<td>17</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Male</td>
<td>76</td>
<td></td>
<td>23</td>
<td>16</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age</td>
<td>18-22</td>
<td>23-27</td>
<td>28-32</td>
<td>33-37</td>
</tr>
<tr>
<td>Female</td>
<td>60</td>
<td></td>
<td>57</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>76</td>
<td></td>
<td>55</td>
<td>15</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Field of Study</td>
<td>Business/Administration</td>
<td>Education</td>
<td>Liberal Arts</td>
<td>Natural Science</td>
</tr>
<tr>
<td>Female</td>
<td>60</td>
<td></td>
<td>11</td>
<td>14</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Male</td>
<td>76</td>
<td></td>
<td>38</td>
<td>11</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

a One person age 68 was in this category
Using the rating rules described in Chapter III, the judges rated each student essay response using a three-digit code. The first two digits represented the dominant Perry position reflected in the response and consequently doubled to reflect the position of dominance. The third digit represented the subdominant Perry position reflected in the response. For example, a rating of 2-2-3 represented a dominant position 2 rating and a subdominant position 3 rating for a given student response. The ratings assigned by each judge to the responses for each subject were then combined and averaged according to the scoring guidelines presented in Chapter III. Thus, a weighted and averaged intellectual position score was used for each subject.

**Interrater Reliability**

Comprised of two writing exercises, the Knefelkamp-Widick instrument was designed to elicit an extended written response from students. Each writing exercise presented a subject stimulus allegedly familiar to each student (a favorite class in Essay I and a decision making experience in Essay II). Each writing exercise focused students on specific aspects of the subject. For example, Essay I suggested specific aspects of the classroom learning experience, and Essay II requested commentary on a decision making experience that was both recent and personally significant. Both writing exercises also asked that students present their thoughts and feelings.

Despite the similarities, differences between the two essays existed. The writing exercise for Essay I focused students more quickly on specific content than did the writing exercise for Essay
II. Essay I was expected to elicit more explicit statements concerning the nature of truth and knowledge, the teacher's role and the learner's role and would, therefore, require less interpretive effort. Also, by the fact of its appearance first in the instrument, Essay I was expected to elicit longer written responses. For these reasons, the assumption was made that the two essays would be differentially difficult to rate.

Because the two essays were believed to be different with respect to rating, interrater reliabilities were calculated separately for each essay. The interrater reliability coefficient for Essay I judgments was .91, computed using a Pearson product-moment correlation. The interrater reliability coefficient for Essay II judgments was .67, also computed using a Pearson product-moment correlation. The two coefficients indicated that the two essays were differentially difficult to rate but were nevertheless rated with reasonable consistency.

**Discrepant Ratings**

Interrater reliability coefficients were based on the original and uncorrected ratings by each judge. After the judges had assigned ratings to all essays, and after interrater reliability coefficients had been determined, those essays in which the simple sum of the two judges' ratings were discrepant by one full intellectual stage or more were rerated. The simple sum was used because attention was on the ratings of the essays considered separately. The simple sum was also used because of its ease of computation.
Discrepancy was based on a three-point criterion level. The three-point criterion level was used because a difference of this magnitude reflected a clear-cut discrepancy of a full stage between the raters' scores. Judges' ratings were considered consistent if the difference between their ratings was within a two-point range or not consistent if the ratings differed by three or more points. For example, if judge 1 assigned a score of 2-2-3 (which sums to 7) and judge 2 assigned a score of 3-3-2 (which sums to 8), then the ratings were considered consistent because the difference between their sums was two or less. If the ratings were 2-2-2 and 3-3-3, however, then the ratings differed by 3 points and were considered inconsistent enough to force rescoring. Of the 255 essays (135 for Essay I and 120 for Essay II), only four individual essay ratings or fewer than 2% were not consistent. These rating differences were discussed between the two judges who then independently rerated the essays. In each case, the difference was resolved within a two-point range and was used in subsequent analyses.

**Distribution of Intellectual Development Scores**

The distribution of weighted average scores for intellectual development fell predominantly in the lower and middle range positions on the Perry scale. Most student responses (126 or 93%) were rated in the position 2 range and only 10 responses (7%) were rated in the position 3 range. No student responses were rated in position 1, the lowest intellectual position on the Perry scale; and no student responses were rated in positions 4 or 5, the upper Perry positions.
considered for research purposes. This lack is noted because it suggests homogeneity among students included in the sample despite the fact that scoring procedures allowed for position 1, 4, or 5 ratings. The distribution of weighted average intellectual development scores within and across student groups is presented in Appendix D.

Measures of Teaching Effectiveness

The Teacher Description Questionnaire was used to measure teaching effectiveness. Described above in Chapters II and III and presented in Appendix C, the instrument was comprised of five subscales: Analytic/Synthetic Approach, Organization/Clarity, Instructor-Group Interaction, Instructor-Individual Student Interaction, and Dynamism/Enthusiasm. Comprised of between six and eight items each, the subscales were designed to investigate several dimensions of teaching believed important to effective performance. This section presents the technical characteristics of the Teacher Description Questionnaire.

Instrument Reliability: Internal Consistency

The internal consistency of the teaching effectiveness measures was calculated using Cronbach's alpha. The coefficient alpha was computed for each of the five subscales separately, and the coefficients, listed by number and subscale title, are presented in parentheses in Table 6. The coefficients, ranging from .87 to .94, are high and indicate that the student ratings obtained were reliable.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Subscale Correlations

The relationship between subscales was computed using the Pearson product-moment correlation. Table 6 also presents the correlation coefficients for all pairs of subscales. The obtained correlations were high. The lowest correlation was .59 (scale 2 with 5 and scale 4 with 5), and the highest correlation was .83 (scale 1 with 2). The coefficients indicate that the five dimensions comprising the rating instrument are highly inter-related, that they are not totally interpretable as separate and distinct components of teaching effectiveness. The finding is contrary to the results obtained by Hildebrand et al. (1971) who reported low or negligible intercorrelations between the subscales.

Table 6
Intercorrelations for Teaching Effectiveness Subscales

<table>
<thead>
<tr>
<th>Teacher Effectiveness Subscales</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Analytic/Synthetic Approach</td>
<td>(.87)a</td>
<td>.83</td>
<td>.74</td>
<td>.63</td>
<td>.59</td>
</tr>
<tr>
<td>2. Organization/Clarity</td>
<td>(.94)</td>
<td>.70</td>
<td>.64</td>
<td>.59</td>
<td></td>
</tr>
<tr>
<td>3. Instructor-Group Interaction</td>
<td>(.92)</td>
<td>.69</td>
<td>.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Instructor-Individual Student Interaction</td>
<td>(.91)</td>
<td>.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Dynamism/Enthusiasm</td>
<td>(.91)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a The alpha coefficient for that subscale
Mean Effectiveness Ratings

Mean effectiveness ratings were computed within and across groups for each subscale and are reported in Table 7. Regardless of the unit of analysis—whether by individual group or combined groups—the mean ratings are relatively high and indicate that students tended to use the upper end of the rating scales for teaching effectiveness measures. The mean range for the subscales varied from approximately one scale point (scales 1 and 4) to two scale points (subscale 5) with subscales 2 and 3 showing a one and one-half scale point range. Despite the tendency for mean ratings to cluster near the upper end of the scale, mean ratings tended to vary by subscale both within and across student groups with the widest range reported for subscale 5 measuring Dynamism/Enthusiasm. The data indicate that the students differentially rated their teachers along various effectiveness scales and that Dynamism/Enthusiasm was the subscale along which student ratings most greatly differentiated between teachers. Thus, the ratings provided means necessary for regression analysis.
Table 7

Mean Effectiveness Ratings by Scale and Groups

<table>
<thead>
<tr>
<th>Effectiveness Measure by Rating Subscale</th>
<th>Student Group by Teacher, Section, Total Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1A</td>
</tr>
<tr>
<td></td>
<td>(N=31)</td>
</tr>
<tr>
<td>Analytic/Synthetic Approach</td>
<td>5.31</td>
</tr>
<tr>
<td>Organization/Clarity</td>
<td>5.55</td>
</tr>
<tr>
<td>Instructor-Group Interaction</td>
<td>4.93</td>
</tr>
<tr>
<td>Instructor-Individual Student Interaction</td>
<td>5.55</td>
</tr>
<tr>
<td>Dynamism/Enthusiasm</td>
<td>6.38</td>
</tr>
</tbody>
</table>

Intellectual Development and Ratings of Teaching Effectiveness

The purpose of this study was to investigate two research questions:

1. Does knowledge of the way a student rates teaching effectiveness enable one to predict the student's stage of intellectual development according to the Perry model, and

2. If predictive power exists, how stable is that prediction?

This section presents the findings that were used in an attempt to answer these questions. The findings described and discussed in the following sections were based on analyses of measures of student intellectual development and student ratings of teaching effectiveness. Intellectual development scores were based on weighted average score ratings and were rounded to one decimal place. Effectiveness ratings were reported as average subscale ratings.
The Question of Predictive Power

The question of predictive power was approached through multiple regression analysis. The procedure was used to produce the maximum possible correlation between student development, the criterion variable, and the weighted sum of the five subscale ratings of teaching effectiveness, the five predictor variables. The procedure used beta weights and involved the calculation of multiple correlation coefficients for the criterion and predictor variables. The multiple correlation coefficients are presented for each student group in Table 8.

Table 8
Coefficients Showing the Relationship Between Intellectual Development and Ratings of Teaching Effectiveness for Each Student Group

<table>
<thead>
<tr>
<th>Group</th>
<th>Teacher Section</th>
<th>Multiple Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>.44</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>.53</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>.78</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>.66</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>.36</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>.33</td>
</tr>
</tbody>
</table>

As the table shows, the coefficients are all positive and range from .33 (Group 3B) to .78 (Group 2A). The multiple correlation coefficients obtained indicate that a trend toward a moderate but
positive relationship exists between intellectual development and ratings of teaching effectiveness and can be regarded as evidence of the existence of moderate predictive power.

The Question of Predictive Stability

The question of predictive stability was addressed with three double cross-validation analyses. Figure 1 shows how students were grouped for each analysis. As shown by the figure, the first double cross-validation analysis involved six student groups (A₁ and B₁), two groups for each of the three teachers. The regression equation derived from the first group was applied to the second group (A/B), and the procedure was repeated but with the second group used as the derivation group (B/A). The second double cross-validation involved the combined sections for each teacher (T₁). The regression equation derived from the combined sections for a given teacher was applied to the combined groups for each of the two remaining teachers (Tⱼ/Tₖ). The third double cross-validation involved two groups. The first group was comprised of students from the first section taught by each teacher (Sₐ) and of students from the second section taught by each teacher (Sₖ). The regression equation derived from each group was applied back on the other group, and a simple Pearson product-moment correlation coefficient between the predicted score and the actual score was calculated. These three double cross-validation procedures were used to examine the systematic variance of student response data attributable to individual student groups, individual teachers and combined student groups.
N = Total number of subjects in sample
A_i = Number of students in Section A of teacher i
B_i = Number of students in Section B of teacher i
A/B = Regression equation derived on Section A and applied to Section B
B/A = Regression equation derived on Section B and applied to Section A
T_i = Number of students enrolled in both sections of teacher i
T_j / T_k = Regression equation derived on teacher j's combined section and applied to the combined sections of teacher k
S_i = Combined number of students across teachers in the ith Section
S_i / S_j = Regression equation derived from the ith Sections and applied to the jth Sections
--- = Designated a cross validation analysis
--- = Designates flow of aggregation and disaggregation of sample

Figure 1. Grouping of Students for Cross-Validation Analyses
The first double cross-validation was conducted to investigate the stability of prediction of student response data within teachers and involved the derivation of a single regression equation for each of the two sections taught by the same teacher. The equation derived from one section was applied to the second section. Correlation coefficients were computed using the Pearson product-moment correlation. Each set of coefficients was then studied to determine the amount of shrinkage as the equation from one section was applied to the other section. These analyses were the basis for estimating the predictive stability of student response data within teachers.

The regression coefficients derived for the study of predictive stability of the student response data within teachers are presented in Table 9. The coefficients initially computed for each group were first presented in Table 8. These coefficients are referred to in Table 9 as derivation coefficients. The regression equation derived for one section was applied to the second section for a given teacher. The resulting coefficients are referred to in Table 9 as cross-validation coefficients.

As the data in Table 9 show, the derivation coefficients range from .33 (Section B3) to .78 (Section A2). Cross-validation coefficients range from -.04 (Section B1 on Section A1) to .57 (Section B2 on Section A2). Shrinkage is evident for each set of coefficients with shrinkage being greatest for Teacher 1's sections and least for Teacher 3's sections. Derivation and cross-validation coefficients are strongest for Teacher 2 and can be considered as evidence for the existence of some predictive stability for this set of student responses. The data
suggest minimal stability for Teacher 3 and negligible stability for Teacher 1. Consequently, these analyses present some evidence to support the assertion that within teachers student ratings are stable predictors of student intellectual development. The evidence, however, exists primarily with student data for Teacher 2.

Table 9
Double Cross-Validation Coefficients for Predictive Stability of Student Response Data Within Teachers

<table>
<thead>
<tr>
<th>Group</th>
<th>Derivation Coefficient</th>
<th>Cross-Validation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>.44</td>
<td>-.04</td>
</tr>
<tr>
<td>B1</td>
<td>.53</td>
<td>.03</td>
</tr>
<tr>
<td>A2</td>
<td>.78</td>
<td>.57</td>
</tr>
<tr>
<td>B2</td>
<td>.66</td>
<td>.52</td>
</tr>
<tr>
<td>A3</td>
<td>.36</td>
<td>.25</td>
</tr>
<tr>
<td>B3</td>
<td>.33</td>
<td>.15</td>
</tr>
</tbody>
</table>

\( ^a \) Subscript identifies teacher

The second double cross-validation was conducted to investigate the stability of prediction of student response data across teachers. For these analyses the regression equation derived from the combined sections of one teacher was applied to the combined sections of each of the remaining two teachers. Correlation coefficients were computed using the Pearson product-moment correlation. The coefficients were then studied to determine the amount of shrinkage. These analyses
were the basis for estimating the predictive stability of student response data across teachers.

The data for these analyses are presented in Table 10. The sample source of the equation is identified by teacher across the top of the table. The sample group to which an equation was applied is identified by teacher down the left side of the table. Column-row intersections show the correlation coefficients obtained when the source equation was applied to the different teacher groups. The data show application correlation coefficients that range from .58 (Teacher 1 on Teacher 2) to -.10 (Teacher 2 on Teacher 3). The data provide evidence for the existence of predictive stability of student response data across Teachers 1 and 2 but negligible stability whenever Teacher 3 was involved. Consequently, these analyses provide limited support for the assertion that across teachers student ratings are stable predictors of student intellectual development. Because the assertion is not supported by coefficients involving Teacher 3, these analyses also provide evidence of the existence of teacher differences with a greater similarity found between Teachers 1 and 2 than between Teacher 3 and either Teacher 1 or Teacher 2.

Table 10
Correlation Coefficients for Predictive Stability of Student Response Data Across Teachers

<table>
<thead>
<tr>
<th>Teacher⁠a</th>
<th>T₁ Equation Application</th>
<th>T₂ Equation Application</th>
<th>T₃ Equation Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>(1.00)</td>
<td>.21</td>
<td>.09</td>
</tr>
<tr>
<td>T₂</td>
<td>.58</td>
<td>(1.00)</td>
<td>.08</td>
</tr>
<tr>
<td>T₃</td>
<td>-.08</td>
<td>-.10</td>
<td>(1.00)</td>
</tr>
</tbody>
</table>

a Subscripts identify teacher
b Coefficients in parentheses are derivation sample statistics

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The double cross-validation analyses discussed above provided evidence for predictive stability of student response data within teachers and some evidence for predictive stability of student response data across some teachers. Because evidence of predictive stability was found, the third double cross-validation was conducted to investigate predictive stability across student groups. For these analyses the regression equation was derived from the combined A sections of the teachers and again from the combined B sections of the teachers. The correlation coefficients for predictive stability were computed using the Pearson product-moment correlation.

The data for the analyses of predictive stability across students are presented in Table 11. As shown in the table, derivation coefficients (.44 and .51 for derivation groups $S_A$ and $S_B$ respectively) are moderately high. Furthermore, the coefficients show only moderate shrinkage. The coefficients provide evidence that student ratings have stability across students as predictors of student intellectual development.

<table>
<thead>
<tr>
<th>Derivation Group $^a$</th>
<th>Derivation Coefficient</th>
<th>Cross-Validation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_A$</td>
<td>.44</td>
<td>.33</td>
</tr>
<tr>
<td>$S_B$</td>
<td>.51</td>
<td>.34</td>
</tr>
</tbody>
</table>

$^a$ Subscript identifies combined sections
In addition to the results presented above, another finding merits presentation. Multiple regression procedures were used to derive multiple regression equations to determine the relationship between the criterion variable (intellectual development) and each predictor variable (the five measures of teaching effectiveness) for each section for each teacher. The Pearson product-moment correlation coefficients are presented in Table 12. The data show that after the inclusion of the first predictor variable (Analytic/Synthetic Approach) in the formula, the remaining predictor variables had negligible influence on the predicted score. This finding indicates that the five measures of teaching effectiveness were not totally unique dimensions. Hence, it is doubtful that five unique predictor variables were statistically obtained for research purposes.

Table 12  
Multiple Correlation Coefficients for Predictor Variables

<table>
<thead>
<tr>
<th></th>
<th>Coefficients for Predictor Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1 A</td>
<td>.11</td>
</tr>
<tr>
<td>1 B</td>
<td>.44</td>
</tr>
<tr>
<td>2 A</td>
<td>.25</td>
</tr>
<tr>
<td>2 B</td>
<td>.27</td>
</tr>
<tr>
<td>3 A</td>
<td>-.04</td>
</tr>
<tr>
<td>3 B</td>
<td>-.12</td>
</tr>
</tbody>
</table>
Summary of Research Results

The purpose of this study was to gather evidence concerning two research questions. The first question asked whether or not knowledge of the way students rate teaching effectiveness could enable one to predict intellectual development. The question was investigated through the use of multiple regression analysis and the derivation of multiple correlation coefficients. The coefficients indicated that moderate predictive power existed, but that the predictive power was neither strong within student groups nor consistent for all student groups.

The second question concerned the stability of predictive power of student response data within teachers, across teachers, and across student groups. Double cross-validation procedures involving multiple regression correlation coefficients and Pearson product-moment correlation coefficients were used to study the questions of stability. The coefficients obtained to study the stability of student response data within teachers provided some evidence for stability, but evidence for stability varied from moderate strength for Teacher 2 to negligible strength for the remaining two teachers. The coefficients obtained to study stability across teachers provided evidence of some predictive stability, but for only two of the three teachers. The coefficients obtained to study stability of student response data across student groups also provided evidence of stability.

The investigative process revealed two additional findings. First, coefficients derived to study the intercorrelation between the teaching effectiveness subscales indicated that the scales were
highly related. The finding was contrary to the results of earlier studies which reported low relationships. Second, ratings by trained judges placed student essay responses almost exclusively in the dualistic category of intellectual development. This finding was unanticipated in light of earlier studies conducted at various colleges and universities in the United States. These studies reported a range of position scores that included scores in the relativistic category of intellectual development.
CHAPTER V
DISCUSSION AND CONCLUSION

This study explored the possibility of a relationship between intellectual development and student ratings of teaching effectiveness. Its purpose was to use the intellectual development construct in Perry's (1968, 1970) scheme as the theoretical framework within which to examine student ratings. Chapter V concludes the presentation of this research effort. The chapter begins with a review of the research findings, an interpretive discussion of their meaning, and a review of the research limitations that temper extrapolation. The chapter concludes with an outline of areas for future research efforts.

A Discussion of the Results

The purpose of this study was to explore the possibility of a relationship between intellectual development and student ratings of teaching effectiveness. The investigation of this possibility was designed to provide evidence concerning two research questions:

1. Does knowledge of the way a student rates teaching effectiveness enable one to predict the student's intellectual development according to the Perry model, and

2. Is the power of prediction stable within teachers, across teachers, and across student groups?

The results of this study do not provide conclusive evidence for definitive answers to either research question, but the positive trend of the findings suggests that the simple answer to each question may be "yes", but with qualifications.
The question of predictive power was investigated with multiple regression analysis and multiple correlation coefficients which were presented in Table 8. The correlation coefficients derived for each student group for each teacher provided evidence of the predictive power of the student response data. Predictive power varied, however, with each student group.

The question of the stability of predictive power for student response data was investigated through double cross-validation procedures and Pearson product-moment correlation coefficients. The results of these analyses were presented in Table 9 for stability within teachers, in Table 10 for stability across teachers, and in Table 11 for stability across student groups. Derivation coefficients and cross-validation coefficients derived for the study of predictive stability within teachers provided evidence of stability which varied for each teacher. Coefficients derived for the study of stability across teachers provided evidence of stability, but the evidence existed for only two of the three teachers involved. Finally, coefficients derived for the study of predictive stability across student groups provided evidence of stability. The strength of the correlation coefficients and the amount of shrinkage between derivation and cross-validation coefficients were the bases for identifying predictive stability. Given these indicators, predictive stability was greatest for the analyses within teachers and across student groups.

The investigation of predictive stability also indicated the existence of teacher differences. The student response data for Teacher 2, for example, gave evidence of predictive stability both
for the within and the between teacher analyses. Student response data for the remaining two teachers did not indicate stability consistently evident in the within and the between teacher analyses. Furthermore, equation applications for the across teacher analyses indicated a greater similarity between Teacher 1 and 2 than between Teacher 3 and either Teacher 1 or Teacher 2.

The stability of predictive power for each teacher considered separately or in relationship to other teachers suggested several possible explanations. First, within teacher stability suggested that to the degree that intellectual development and student ratings were related, they were consistently related given stable teacher traits and group characteristics. Given this assumption, the low predictability of Teacher 1 may be attributed to changed teacher behaviors, group differences, or a combination of teacher change and group differences. Because Teacher 1 was fairly predictable using Teacher 2's regression equation, a likely explanation for low within teacher stability in this instance would consider group differences and perhaps some teacher change. Second, stability of predictive power across teachers suggested that to the degree that intellectual development and student ratings were related, they were consistently related. Given this assumption, the predictability existing between Teachers 1 and 2 suggested that, despite the existence of some group differences, these two teachers were similar with respect to one or perhaps a set of teacher traits. Conversely Teacher 3 was different. An explanation of this difference might consider a range of teacher traits including instructional mode.
Teachers involved in this study were matched according to their alleged use of the lecture mode, but the possibility nevertheless exists that significant variations within this mode may account for the teacher differences indicated by the student response data.

Two other findings reported elsewhere in this study merit discussion. First, the premise was incorrectly made that a sample of 136 students would yield a full range of intellectual development position scores. The measure of intellectual development, however, revealed a student sample extremely homogeneous with respect to intellectual development. Second, the premise was also made that the Teacher Description Questionnaire comprised of five subscales would function as the source of five unique predictor variables. This premise was based on research by Hildebrand et. al (1971) who reported for this questionnaire the existence of five subscales, each of which was highly reliable but negligibly correlated with the remaining subscales. Analysis of the effectiveness ratings obtained for this study, however, indicated a high intercorrelation among the five subscales and suggested that the dimensions of effective teaching could not be interpreted as five unique variables. Rather, the dimensions were found to be highly inter-related. These two findings, extreme student homogeneity and a high intercorrelation among rating subscales, may have attenuated the correlations.

The results of this study did provide useful information. Recall that the purpose of the study was to explore the possibility of a relationship between intellectual development and student ratings
of teaching effectiveness. The exploration suggested that a relationship might indeed exist and that such a relationship might possess the qualities of predictive power and predictive stability.

The study also demonstrated the applicability of the Perry model and the Knefelkamp-Widick instrument to research concerned with student ratings of teaching effectiveness. The Knefelkamp-Widick instrument was administered efficiently to a fairly large sample of students. Students did provide responses of reasonable content and length such that ratings could be made. Judges could be trained to rate with respectable interrater reliability the responses generated by students, and evidence was subsequently found to support the possibility of a relationship between intellectual development and student ratings of teaching effectiveness.

Limitations of the Study

Any conclusions drawn from this study must be tempered with a recognition of the limitations that exist. For example, the students in this sample were self-selected in the sense that they registered for classes taught by teachers whose involvement in the study was based on their common instructional approach and department affiliation. Consequently there was no way of determining whether or not the sample was representative of Western Michigan University students or even of students within the College of General Studies since students may select instructors in some systematic way.

The students were also singularly defined. That is, student similarities and differences for research purposes were presented
in light of one descriptive variable--i.e., intellectual development. Research to date, however, indicates that a student's rating of teaching effectiveness involves a complex of student attributes of which intellectual development may be one attribute but is by no means the only one.

Another limitation concerns the possibility of systematic rater bias. In the absence of standardized training materials and procedures for use with the Perry model and the Knefelkamp-Widick instrument, there was no way to check the generalizability of ratings. That is, it could not be said that the judges who rated responses for a different study would have given the student responses the same ratings that the raters for this study gave. Therefore, the possibility exists that ratings were systematically biased.

A fourth limitation involves the usefulness of the Teacher Description Questionnaire for research involving intellectual development. Although the reliability of the instrument was established in the sense that students were consistent in their ratings of teaching effectiveness, the validity of the instrument was questionable in light of the high intercorrelations among the five subscales.

Implications for Future Research

This study provided evidence for the existence of a relationship between intellectual development and student ratings. Although the evidence did not provide conclusive answers to the research questions, the evidence did indicate positive trends and areas for future research efforts.
One implication for future research is the need to study a sample of students more diversified with respect to intellectual development. Because students involved in this study were so developmentally homogeneous, it is not clear whether or to what extent student ratings varied with intellectual development. One way to proceed in future research efforts might be to include college-bound high school students and graduate students in the sample. Assuming that intellectual development, age, and education level are related, the inclusion of such students might eliminate the possibility of attenuated correlations due to student homogeneity.

A second implication is the need to select a more representative group of students and more heterogeneous teachers. Because students in this study were not representative of students in the department or in the university, the generalizability of results is severely restricted. A random sample of students enrolled in lecture classes from throughout the university would certainly increase the generalizability of results. A selection of more heterogeneous teachers with respect to one or possibly several characteristics would make teacher differences explicit and would most likely strengthen predictions based on student response data.

A third implication is for continued research with instruments for rating teaching effectiveness. Because the instrument used in this study failed to yield five distinct subscales, the question of interest is whether or not the failure was attributable to this specific rating instrument, to rating instruments in general, or to intellectually dualistic students who for various reasons might
tend to render high ratings or who might be unable to differentiate between components of teaching effectiveness. An investigation of this question might involve a student sample for whom intellectual development scores were known. The students could be asked to make teaching effectiveness ratings of the same teacher(s) but using a variety of instruments, each purported to have subscales measuring various and unique effectiveness dimensions. The rating data might enable a researcher to examine the extent to which the subscales in each instrument were or were not distinct and thus to compare the instruments in terms of their potential to yield multiple and distinct predictor variables. A second approach might involve in-depth interviews with students at different stages of intellectual development. The interviews could present students with teaching effectiveness items, and student responses to these items could be probed in an effort to search for differentiated cognitive processes and rating judgments of students at different stages of intellectual development.

Still another implication is for continued development of training programs for use with the Perry model and with the student essay response mode. The development of rating guides is particularly important. Currently rating guides are sketchily described or comprised of rules only generally outlined. Although useful, existing guidelines need explication, refinement, elaboration, and whenever possible illustrative materials so that rating criteria are as precise as possible and so that the comparability of rating results across studies can some day be demonstrated rather than merely assumed.
A final implication concerns longitudinal rather than cross-sectional research involving the interaction effect of multiple variables. For example, if students can be described according to the Perry model—and this and other studies suggest that they can be—and if students' cognitive structures change during their college experiences—and earlier studies by Knefelkamp (1974) and Widick (1975) show that they do—then a longitudinal investigation might illuminate our understanding of the interaction between student intellectual development, student ratings of teaching effectiveness, and other educationally relevant variables such as instructional content and method.

Summary

This study investigated student ratings of teaching effectiveness, their predictive power and the ability of that power relative to student intellectual development. Evidence was found to support the existence of predictive power and stability within teachers, across teachers, and across student groups. Although the evidence was for some teachers sometimes negligible, at other times low, but often moderate, a positive trend was identified. This trend and the limitations of this study were considered in the section suggesting areas wherein continued research is not only possible but needed.
BIBLIOGRAPHY


Craeger, J.A. A multiple-factor analysis of the Purdue rating scale for instructors. LaFayette, Ind.: Purdue University, 1950 70, 75-96.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Melnik, M.A. & Sheehan, D.S. A diagnostic model of teaching improvement. Amherst, Mass.: University of Massachusetts, Clinic to Improve University Teaching, 1974.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Rees, R.D. Dimensions of students' points of view in rating college teachers. *Journal of Educational Psychology*, 1969, 60(6), 476-482.


Widick, C.C. Personal communication, December 21, 1977.


APPENDIX A

THE MEASURE OF INTELLECTUAL DEVELOPMENT

The Knefelkamp-Widick Student Questionnaire
and
Administrative Procedures
To the Student:

On the following pages is an instrument that has two parts. The first part concerns how you as an individual think about your classroom experiences. The second part concerns an important and personal decision-making experience.

We are asking you to respond in writing to the questions asked on the next two pages. Please write your response to each question in the space provided. You may use the back side of each page and include additional pages if you wish. Please try to write at least three paragraphs for each question.

We would like to thank you for your help with this research project. In particular, we appreciate your willingness to think about and to discuss these subjects.

Thank you.
Dr. ________ will be here in a few minutes. In the interim, I am here in his place to ask for your assistance in an educational research project at Western Michigan University.

My name is ________, and I am affiliated with the Evaluation Center here at Western. We are currently engaged in a project that attempts to study the way students view teaching in higher education. Identifying student preferences in the classroom and the way students resolve certain problems is the first step in our research.

During the next thirty minutes we are asking that you respond to this instrument which deals with how you as an individual think about several subjects. The first subject concerns your favorite class and what made it so. The second subject concerns a recent decision you have made. We will also need to obtain some basic information: your name, age, sex, and college major. This information is for research purposes only. The basic information and your response to the instrument will be held in the strictest confidence.

Please think about the questions, and then write your responses in the space provided. There are no right or wrong responses. We are not concerned with such things as spelling or grammar or writing style. What is important is the way you think and feel about the subjects presented.

At the end of thirty minutes I will collect the instruments and answer any questions you might have regarding the project.

Are there any questions?

You may begin.
We would like you to write an essay in response to the questions posed on the following two pages. It will help us greatly if you can be as specific and complete in your answer as possible.

A. Topic Question: Please describe the best class you have taken since you have been in college; or if you are a first-semester freshman, you may describe your best class in high school or another learning institution. What made it positive for you? Please be as specific as possible. Feel free to go into as much detail as you think will give us a clear idea of the class; for example, you might want to discuss areas such as what the teacher was like, the subject, the particular content (readings, films), the atmosphere of the class, grading, procedures, etc. We want a description of your experience and how you thought and felt about it.
B. Topic Question: Think of the last time when you had to make a decision about something that had major importance to you or the last time you had to choose between some significant alternatives. 1) How did you feel about having the alternatives? 2) How did you go about making the decision? 3) How did you feel about it afterwards? Be as detailed as possible in your description.
APPENDIX B

OUTLINE OF TRAINING CONTENT AND STRUCTURE

I. General Guidelines
II. Position Cues
III. Anchor Essays
IV. Schedule of Sessions

105
I. GENERAL GUIDELINES

I. Remember always that regardless of content, the primary question is: how does the complexity of structure place the response on the Perry scale?

II. Assess structural cues.

A. Examine the writing sample for basic assumptions about the world.

1. Is it dualistic? Two categories exist: good v. bad.

2. Is it multiplistic? There are many categories, but they are essentially equivalent and unordered—e.g., a good teacher is an expert, is warm, tells good jokes, is friendly, talks clearly, lectures well, and is interesting.

3. Is it relativistic? Multiple categories exist, but they are ordered in content—e.g., a good teacher is warm, expert, and clear in terms of presentation. His warmth is conveyed in lots of ways, for example, he may tell good jokes or simply seem friendly.

B. Examine writing sample for qualitative differences in use of concepts. Multiple dimensions exist and may include the student's:

1. use of absolute v. qualified statements.

2. awareness of alternative perspectives.

3. complex analytic thought applied to various topics including self.

4. sense of dependence v. autonomy; internal v. external locus of control.

5. ability to empathize with others, especially those quite diverse.

6. ability to synthesize, integrate ideas, especially those ideas that are similar in a concrete sense.

7. ability to assume responsibility and to take on new roles.
III. Assess the writing sample for the pattern of attitudinal and/or behavioral correlates which seem related to different Perry stages.

A. Look especially for attitudes about and preferences expressed in regard to:

1. the proper role of a learner.
2. the characteristics of a good teacher.
3. the purpose and appropriate method of evaluation.
4. the type of atmosphere that ought to exist in a classroom.
5. when and where frustration occurs.

B. Look at writing sample for cues in the form and style of language use. Examine differences in:

1. words used: concrete v. abstract words.
2. length of response.
3. complexity of sentence structure.

IV. Rating Strategies and Reminders

A. Reminders

1. Keep your focus on structure rather than on content.
2. Consider position as a point of outlook from which the student views the world.
3. Look at coherence of forms rather than the endurance of form.
4. Look for forms of assumptions about knowledge and value.
5. Don't overinterpret—i.e., give undue importance to multiple ideas regardless of whether or not those ideas are expressions of some truth.
6. Don't ignore the Authority implicit in a statement.
7. Remember time cues.
8. Don't overemphasize parenthetical expressions—e.g., maybe, I suppose.
9. Look for cues of uncertainty, especially uncertainty that is believed to be pervasive or temporary or illegitimate.

B. Strategies

1. Use diagrams for identifying broad categories, dualistic v. relativistic thought.

2. If a range of forms is found within a single response, look for the position of central tendency.

3. Build a defense for position ratings—e.g., identify specific cues, defend the position rating against ratings above and below the one you have assigned.

4. Review each student response and your rating at a second sitting and without access to your earlier rating.
Position 1

1. Cognitive simplicity is dominant. The student uses the simplest set of assumptions.

2. In-group vs. out-group dichotomies may exist. The student uses unconsidered differentiations.

3. The student is authority oriented: Authority = right = we = I.

4. Authority is equated with absolute Right. People in authority are right by definition.

5. Morality and personal responsibility consist of simple obedience.

6. What authorities want is, of course, truth and right.

7. Depending upon the topic, teachers, instructors, even peers can function as Authority.

8. Detachment is impossible. At this stage a student can not stand back and observe himself, cannot detach himself from categories.

9. The world is free of conflict. One's differences from Authority are not given legitimacy. There are dumb people, wrong people, and solutions, but no real conflict. The individual experiences no conflict. Things are clearly right or wrong.

10. An act is either right or wrong. It cannot be better or worse except in a quantitative sense. There are no shades of gray.

11. Judgments refer to quantity not quality.

12. No opinions or acts are neutral; all can be put into one of two categories.

13. Teaching methods are criticized as too diffuse, bad, failed to give the (right) answers.

14. Learning to be independent is learning self-controlled obedience.

15. No distinctions are made between the teacher as Authority and absolute truth.

16. In education, morality is hard work, memorizing, correct answers, doing what Authority asks.

17. Teacher Authority differentiated between those who mediate well and those who do not, those who know and those who do not know, real Authority from fraudulent types.
Position 2

1. The student can begin to see differences between absolutistic opinions.

2. Rebellion can occur in this position. It is often in opposition to the diffuseness of teachers: for example, theories are bad.

3. Adhering students (non-rebelling students) guess that teachers present things diffusely so students will have to find the right answer.

4. Poor Authorities introduce diversity and complexity. Thus, authority gets separated into good Authority who is truthful and good or into bad Authority who is a false god.

5. Complaints against diffuseness are often unfocused.

6. A liberal education is resented: it doesn't do any good.

7. If activism occurs, it is dogmatic in nature and often involves indiscriminate protest.

8. Dogmatic inactivism can also occur and often in the form of indiscriminate and global love, happiness, humanism, or psychadelic experience.

9. The student may feel out-of-it, immature, having missed the boat. The student may observe that if other students can and do enjoy something like theorizing or interpreting and she/he does not, where does that leave her/him?

10. Whether perceived in the teacher, course, peer culture, multiplicity is a mere appearance and is not real. Teachers convey multiplicity to make students search for the right answer.

11. Teachers use multiplicity through instructional content and/or method to elicit more work.

12. Unable to tolerate uncertainty, the student searches for certainty and often expresses approval of or preference for the sciences as opposed to the humanities.

13. Interpretive exercises have no meaning because uncertainty is not a legitimate reality.

14. For some students, the discovery of the legitimacy of uncertainty as a temporary phenomena is experienced as liberating, as a feeling of greater freedom for self-regulation.
Position 3

1. Room is made for legitimate human certainty and uncertainty. Uncertainty is, however, only temporary.

2. The tie is loosened between Authority and the Absolute—even more so than in Position 2.

3. Uncertainty is unavoidable, even in the sciences.

4. The salient questions are: how are answers judged? Is not one answer as good as another?

5. Evaluation becomes the prime issue. Authorities pass judgment even when they themselves are ignorant of the right answer.

6. Reward for hard work does not necessarily exist anymore. Old rules do not work. This can cause some confusion and resentment.

7. Quantity becomes important. The initial impact of complexity and diversity is often experienced in terms of sheer quantity. Students respond to the amount of reading, the length of papers, the number of details, all seen as indissoluble discretes.

8. Uncertainty implies the legitimacy of the multiplicity of answers.
Position 4

1. Position 4 can be experienced through opposition or through adherence.

2. If the oppositional mode is chosen, then the student may manifest the following:
   a. No one has a right to call anyone's opinion wrong.
   b. The domain of freedom may be extended at the expense of Authority's claims.
   c. The game is: give Authority what you figure he wants.
   d. Authority may be seen as bigoted and dogmatic.
   e. There may be some perception of the difference between an opinion and a supported opinion.

3. If the adherence mode is chosen, then the student may manifest the following:
   a. Multiplicity is something Authorities want students to work on.
   b. There is movement from the position of what they want to the way they want us to think. Students begin to have a structural element in their thinking.
   c. The way they want us to think forces a comparison of patterns of thought. The student begins to think about thinking.
   d. A distinction is made between an unconsidered belief and a considered judgment. The distinction is explicit.
   e. The student believes independent-like thought should get good grades.
   f. Responsibility becomes explicit and is experienced primarily in matters of petty conduct and always with reference to Authority.
   g. Responsibility for conduct emerges with expanding responsibility for studies.
   h. The student begins to relate courage (putting oneself on the line) with choice.
   i. The student can also demonstrate a quality of detachment which allows avoidance of responsibility.
Position 5

1. Before relativism was a special case, a way of thinking about a certain class of problems, a way of making sense in an otherwise chaotic multiplicity, a special procedure, something teachers wanted.

2. Now, relativism is perceived as the common characteristic of all thought, all knowing, all of one's relation to one's world.

3. Now, dualistic right/wrong thinking and even ideas of absolutes become special cases in the new relativistically structured context.

4. The student has not yet faced or come to grips with the personal and social implications of his discovery: responsibility and commitment.

5. There are four salient qualities of this position:
   a. breakdown of the old structure and identity, balanced by a realization of growth and competence in a relativistic world.
   b. changed relation to authorities.
   c. new capacity for detachment.
   d. unawareness of a path toward a new identity through personal commitment.

6. Students experience new breadth; they contrast this with a sensed narrowness of the past. They also fear any narrowness in the future. They fear loosing the breadth and becoming too specialized or set in one's ways.

7. They experience a sense of expansion which has a lasting quality.

8. Commonly, the first compensation for insecurities of relativism is that of immediate competence in management of studies.

9. A sense of lostness and loneliness threatens. It is commonly believed that the community can provide the required strength.

10. Authority now rests on experience and expertise in groping. Authority (large caps) becomes authority (small caps) and serves a social function.

11. The student finds it easier to ask for help. The sense of community and a different view of learning makes asking easier.

12. There emerges a new meaning and respect for individuality.
III. ANCHOR ESSAYS

Position 1

No examples of position 1 essays were found among the student responses collected for research purposes.

Position 2

The best class I have is the class that makes me work the hardest. Through working hard you never stop reaping the rewards nor do you feel you could do less; just more.

Position 3

One of my best classes I have taken is Marketing 370. I'm a business major so naturally the class would be of some interest to me. What I really liked about the class was my instructor. He knew the topic well and could explain things to the students' level.

Besides having a good instructor, it also helped that marketing tends to have a large common sense factor. Also a lot of the topics we talked about were observable in the everyday world.

Summing it all up, I think a good instructor who can relate to the students is important, but also the class material must be interesting to the student.

Position 4

The best college course I took was entitled "The Religious Implications of Recent American Fiction." It was an upper level religion course taught by the most sensitive man I have ever met. He was intimately committed to his subject and to transmitting his commitment to his students. We explored the subject of setting really and how characters were either attached or detached from their setting and drew several conclusions from it. We also explored maternal imagery and mind/body dualism. We dealt with personal theories of the professor which had a great deal to do with the exciting atmosphere of the class. Some of our readings were Vonnegut's Cat's Cradle, Kesey's Cuckoo Nest, Nabokov's Transparent Things, Sontag's Death Kit, Mailer's American Dream, Keroac's Drarma Bums, Pynchon's The Crying Lot 49, and one book each by Bellow and Barth. So the readings were contemporary and fast paced, as well as controversial which contributed to much of the charged atmosphere of the class. We each had to lead a discussion with 2 others on a book, but it wasn't a report because everyone always had something to say. The professor would give his personal theory on another day, and on the next we read an essay on Transcendence which was the real focus of the course—how could man transcend his environment? The essays tied in with the particular book we'd read that week and were written by noted theologians and philosophers and were quite stimulating.
The experience for me was one of complete joy. I am a natural student, and may be most happy when challenged and stimulated in a course. I am quite sure that it was the feeling. In fact, the whole class sensed a different sort of atmosphere. I liked it better than some because I agreed with the professor's schemata and since it had such a personal investment in it, he did not take all to kindly to criticism. But it was still the kind of class where we got together at various people's houses for cook outs, wine parties, we gathered for coffee before class—it was beautiful.

My strong commitment to the subject helped me. I was an English and Religion major and loved to see the 2 interrelated. But, as you know from teaching, some classes just seem to have that magic. It's almost inexplicable and I guess most students and profs wish it could be bottled.

Position 5

The best class I have taken since I have been in college was Comparative Literature 101. We read such books as Candide by Voltaire, The Stranger by Albert Camus, the Aeneid by Virgil, the Invisible Man by Ellison, and Soul on Ice by Eldridge Cleaver.

The interesting part of the course was the class discussion which followed the reading of each book. The teacher gave us limited guidance, but we had to interpret the material for ourselves. We took part in debates, in which no right answer was ever found. This class was one of personal growth for me. I found a thread which wound throughout each of the books as well as throughout my own life. I found that ultimately we are all very much the same, regardless of the race, sex, or culture. We, as human beings have the same joys and pain. After the class, I became very compassionate about my fellow man. For after all, we as human beings even share the same fears and weaknesses.

Because of that class I decided to major in Comp. Lit as an undergraduate. Through the subsequent literature I learned much about my fellow man, as well as about myself.
IV. SCHEDULE OF SESSIONS

Session 1

B. Hand out copy of Knewi instrument.
C. Hand out general guidelines and cue sheets.

Session: The Perry model was presented and illustrated with specific examples. One set of student responses was examined, position cues were applied, and structures were analyzed and discussed.

Session 2

Preparation: A. Analyze five student responses.
B. Rate the student responses.

Session: The raters analyzed and discussed their ratings. Differences and similarities between the ratings were examined. Position cues and guidelines were reviewed.

Comments: Initial ratings are overly responsive to the length of the student response and an over-all impression by the rater. The need exists to read the responses sentence by sentence to help focus on structure and to help see multiple structures within a single response.

Session 3

Preparation: A. Analyze five student essay responses.
B. Rate each response and note rating cues used.

Session: Individual ratings were discussed in terms of position cues and general guidelines. The discussion focused on the role of authority explicit and/or implied in the essays.

Comments: Raters identified correctly and agreed on dominant position cues for the high and low positions. Differences between positions 2 and 3 and positions 3 and 4 were problematic. The tendency persists to rate on an over-all general impression.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Session 4

Preparation: No preparation.

Session: Raters were given a series of completed sentence stems from an earlier and related study. Each sentence was rated individually and discussed. Sentences were then ordered back into the original response from which they were taken.

Comments: This was a useful strategy for helping raters slow their reading paces, examine individual statements, and focus on structure. Sentences formerly from single student responses and recombined illustrated well multiple position statements found in some responses.

Session 5

Preparation: Examine and rate 8 student responses in the 2-3 position range.

Session: Sentence-by-sentence analysis and discussion helped reveal multiple positions within a single response. Raters agreed on the stages found within a single response but sometimes reversed the assignment of dominant and subdominant numbers. Guidelines for determining subdominant emphasis were discussed.

Comments: One rater is consistently high and the other low when responses involve position 2-3 combinations.

Session 6

Preparation: Examine and rate 8 student responses in the 2-3 position range.

Session: Individual ratings were discussed and analyzed. Position 3 cues were carefully reviewed.

Comments: Increased sophistication of position 3 language tended to influence raters who focused on content again rather than structure. Raters agreed on 75% of the responses.
Session 7

Preparation: Rate two sets of responses, 29 in all.

Session: Responses were in the 2-4 range. Position 2 and 2-3 transitions were discussed and analyzed. Positions 3 and 4 were reviewed in detail.

Comments: Positions 2 and 3 are no longer problematic. Raters tend to over-emphasize multiple categories regardless of the meaning conveyed, to focus on authority assumptions only if one or two categories are identified, and to over-emphasize the significance of adverbs such as very, always, maybe, and perhaps. Position 3-4 transitions need work. Rater agreement is .62.

Session 8

Preparation: Rate one set of responses in 2-4 position range.

Session: Raters discussed cues regarding uncertainty and the meaning of specific words such as risk, factors, evidence, and belief. The differences between ordered and unordered categories was discussed.

Comments: Raters still have problems distinguishing between multiple categories, especially identifying the balanced perspective of relativistic as opposed to multiplictic perspectives. Transitions between positions 3 and 4 need work.

Session 9

Preparation: Rate one set of responses in 2-5 position range.

Session: Position 3-4 transitions discussed in detail. Raters have a 90% agreement on positions 2 and 3. Agreement drops on responses with a mix of position 3 and 4 statements.

Comments: Raters need to work on cues implying temporary or pervasive uncertainty. Perhaps yearning to find position 5 responses, raters tend "to reward" any evidence of complexity with a 4 or 5 rating regardless of the existence of cues for the legitimacy of uncertainty or the balanced perspective of the relativistic thinker.
Session 10

Preparation: Raters are given diagrammatic illustrations of position 2, 3, and 4 responses.

Session: Raters were given 10 responses and were asked to diagram categories present in each before giving ratings.

Comments: Rating accuracy was high (90%) for positions 2 and 3. The use of diagrams seemed to help clarify most categorical differences. Work is still needed on transitions from position 3 to 4.

Session 11

Preparation: Rate one set of responses in the 2-4 position range.

Session: Individual ratings were discussed and analyzed.

Comments: Accuracy was high (85%) for position 2 and 3 responses. Positions 4 and 5 need work, especially with the issue of the legitimacy of uncertainty.

Session 12

Preparation: Rate one set of responses in the 2-3 position range.

Session: This was a short review session for the purpose of selecting essays illustrating positions 2 and 3. These would be referred to as anchor essays and would represent typical responses for these positions.

Comments: The next session will concentrate exclusively on positions 4 and 5.

Session 13

Preparation: Rate one set of responses in positions 4 and 5.

Session: A discussion of the differences between positions 3 and 4 and positions 4 and 5 concentrated on belief and considered judgment, balanced perspectives, and commitment within a relativistic framework.

Comments: Position cues for 4 and 5 are inadequate. Perry needs to be reread and more functional cues need to be developed.
Session 14
Preparation: Review literature for positions 4 and 5.
Session: Raters delineated more functional position 4 and 5 differentiations.
Comments: Primarily a working session to clarify 4 and 5 cues. Raters agreed that position 4 students will acknowledge the legitimacy of complexity but will not necessarily move beyond acknowledgement to exploration (position 5), will see discrete and differing points of view but will not necessarily see variation with a given view (position 5), might see the need for commitment within a relativistic framework but will not necessarily acknowledge a personal need to do so (position 5), and might recognize the complexity of issues but would not reduce the issues to their complex components or discuss the components complexly.

Session 15
Preparation: Review position 4 and 5 cues and continue rating student responses.
Session: Student responses ranged from 3-5. Refined position cues were discussed and applied. Position 3-4 transitions were reviewed at length.
Comments: Raters need to be reminded that diagrammatic structures can help reveal distinctions between positions 4 and 5. Rater agreement is 70%, the drop attributable to the problem of 3-4 transitions.

Session 16
Preparation: Rate responses in 2-5 position range.
Session: Raters reviewed position cues, rated responses, and discussed differing ratings.
Comments: Accuracy was high (95%). Two responses in the 2-3 transitional stages were problems.

Session 17
Preparation: No preparation
Session: Ratings were assigned to a sample of position 2-5 responses.
Comments: Interrater agreement was high (90%). Appropriate.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Session 18

Preparation: Raters reviewed training materials and prepared questions.

Session: The rating task was described and rating materials distributed. The discussion concerned rating cues and general differences between the positions with respect to specific characteristics of intellectual development. Discussion focused on 5 responses that had been particularly ambiguous and difficult to rate.
The Teacher Description Questionnaire has two parts. The first part is the cover sheet and asks for your name, social security number, major, and whether or not you completed the student questionnaire distributed in class in February. This information is for research purposes only; it will not be reported to the teacher.

The second part is the actual rating form. Beginning on the front and continuing on the back, the rating form is comprised of a number of statements concerning the instruction you have received. Please respond to every statement by assigning a number rating. A rating of 1 is the lowest rating you can give. A rating of 7 is the highest rating you can give.

Please remember to fill out the front and the back side of the form. Please keep in mind that the amount of negative or positive praise is not the thrust of the research. What is important is that your rating accurately reflect how you think and feel about the quality of instruction you received in this course.
EVALUATION COVER SHEET

Name ________________________________________________________________

Social Security Number ____________________________

Major ________________________________________________

Did you complete the student questionnaire?  ____ Yes  ____ No
### Teacher Description Questionnaire

1. **My age:** 1. under 18;/2. 18-19;/3. 20-21;/  
   4. 22-23;/5. 24-25;/6. 26-30;/  
   7. 31-35;/8. 36-40;/9. 41+  
   
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

2. **My year in school:** Freshman; 2. Sophmore; 3. Junior; 4. Senior; 5. Other

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3. **I am:** 1. male; 2. female

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

4. **My cumulative CPA is:** 1. 0-.9;/2. 1.0-1.4;/  
   3. 1.5-1.9;/4. 2.0-2.4;/5. 2.5-2.9;/  
   6. 3.0-3.4;/7. 3.5-4.0

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

6. **My reason for taking this course:** 1. fulfills  
   General Education requirement; 2. fulfills Major/Minor requirement; 3. recommended to me; 4. out of personal interest.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**DIRECTIONS:** Each of these statements describes a basic component of teaching. Give the instructor an overall rating for each component, reserving the highest scores for unusually effective performance. If items do not apply, leave blank.

7. **Discusses points of view other than her/his own.**

<table>
<thead>
<tr>
<th>Low Score</th>
<th>High Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234567</td>
<td>000000000</td>
</tr>
</tbody>
</table>

8. **Contrasts implications of various theories.**

<table>
<thead>
<tr>
<th>Low Score</th>
<th>High Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234567</td>
<td>000000000</td>
</tr>
</tbody>
</table>

9. **Discusses recent developments in the field.**

<table>
<thead>
<tr>
<th>Low Score</th>
<th>High Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234567</td>
<td>000000000</td>
</tr>
</tbody>
</table>

10. **Presents origins of ideas and concepts.**

     | Low Score | High Score |
     |-----------|------------|
     | 1234567   | 000000000 |

11. **Gives references for more interesting and involved points.**

<pre><code> | Low Score | High Score |
 |-----------|------------|
 | 1234567   | 000000000 |
</code></pre>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Low Score</th>
<th>High Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.</td>
<td>Presents facts and concepts from related fields.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>Emphasizes conceptual understanding. (Stresses major ideas rather than isolated facts.)</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>Explains clearly.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>Is well prepared.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>Gives lectures that are easy to outline.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>Is careful and precise in answering questions.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>Summarizes major points.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>States objectives for each class session.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>Identifies what he/she considers important.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>21.</td>
<td>Encourages class discussion.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td>Invites students to share their knowledge and experience.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td>Clarifies thinking by identifying reasons for questions.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>24.</td>
<td>Invites criticism of own ideas.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td>Knows if the class is understanding him/her.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>26.</td>
<td>Knows when students are bored or confused.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td>Has interest and concern in the quality of his/her teaching.</td>
<td>1234567</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Has students apply concepts to demonstrate understanding.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------</td>
<td>-----------</td>
<td>------------</td>
</tr>
<tr>
<td>28</td>
<td>Has genuine interest in students.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>29</td>
<td>Is friendly toward students.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>30</td>
<td>Relates to students as individuals.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>31</td>
<td>Is accessible to students out of class.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>32</td>
<td>Is valued for advice not directly related to the course.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>33</td>
<td>Respects students as persons.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>34</td>
<td>Is a dynamic and energetic person.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>35</td>
<td>Has an interesting style of presentation.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>36</td>
<td>Seems to enjoy teaching.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>37</td>
<td>Is enthusiastic about his/her subject.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>38</td>
<td>Seems to have self-confidence.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>39</td>
<td>Varies the speed and tone of his/her voice.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
<tr>
<td>40</td>
<td>Has a sense of humor.</td>
<td>Low Score</td>
<td>High Score</td>
</tr>
</tbody>
</table>
| 41| If you wish to make any further comments on the course and/or effectiveness of the instructor, please use the sheet of paper that is provided. | 1234567 | 0 0 0 0 0 0 0 0 0

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
APPENDIX D

DISTRIBUTION OF WEIGHTED AVERAGE INTELLECTUAL DEVELOPMENT SCORES OVER BOTH ESSAYS WITHIN AND ACROSS STUDENT GROUPS
Distribution of Averaged Intellectual Development Scores
Over Both Essays Within and Across Student Groups

<table>
<thead>
<tr>
<th>Averaged Position Score</th>
<th>1A (N=31)</th>
<th>1B (N=25)</th>
<th>2A (N=19)</th>
<th>2B (N=24)</th>
<th>3A (N=22)</th>
<th>3B (N=15)</th>
<th>Total (N=136)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>D 2.2</td>
<td></td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>U 2.3</td>
<td></td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>A 2.4</td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>L 2.5</td>
<td></td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>I 2.6</td>
<td></td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>S 2.7</td>
<td></td>
<td>10</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>M 2.8</td>
<td></td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>2.9</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.0</td>
<td></td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>3.3</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R 3.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E L 3.7</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>A T 3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I V 3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I S 4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.