McKethan, Robert N.

THE DEVELOPMENT AND EVALUATION OF A BEHAVIORALLY ANCHORED RATING SCALE FOR SECONDARY PHYSICAL EDUCATION TEACHERS

The University of North Carolina at Greensboro ED.D. 1985

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark __√__.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages __√__
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Dissertation contains pages with print at a slant, filmed as received _____
16. Other _____

_____

_____

University
Microfilms
International

THE DEVELOPMENT AND EVALUATION OF A BEHAVIORALLY ANCHORED RATING SCALE FOR SECONDARY PHYSICAL EDUCATION TEACHERS

by

Robert N. McKethan

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

Greensboro
1985

Approved by

Sarah M. Robinson
Dissertation Adviser

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Dissertation Adviser Sarah M. Robinson

Committee Members Gail M. Hennessy
Marian K. Bolleter
William H. Turkey
Marie Riley

October 14, 1985
Date of Acceptance by Committee

October 14, 1985
Date of Final Oral Examination

MCKETHAN, ROBERT N. The Development and Evaluation of a Behaviorally Anchored Rating Scale for Secondary Physical Education Teachers (BARSSPET). (1985)

Directed by Dr. Sarah M. Robinson

The purpose of this study was to develop a behaviorally anchored rating scale for obtaining student feedback about teaching behaviors of physical education teachers. A second purpose of this study was to use additional rating data to ascertain the psychometric properties of the BARSSPET.

Four hundred and eighty-eight students and 14 teachers in six independent subject groups participated in the development of the BARSSPET. In part one of the study, the steps included (a) generation of behavioral statements, (b) editing of the behavioral statements, (c) rating of the behavioral statements, (d) dimension identification, (e) allocation and value assignment, and (f) final selection of anchors and dimensions.

In the first step, 735 behavioral statements, depicting three performance levels, were generated. After the editing step, 383 behavioral statements remained for further analysis. In the rating step, Cronbach's alpha analysis was used to retain 135 behavioral statements which showed high internal consistency. Eleven dimensions of teaching behavior were identified from the literature and nine were presented to student raters for verification through a rating process. In the initial attempt to allocate statements to dimensions

and assign values to statements, students were unable to satisfactorily allocate statements to dimensions. An alternative procedure was undertaken in which dimensions were identified using the pool of 135 statements. The new dimension identification process yielded seven dimensions for use in the alternative procedures developed for allocation and value assignment. After the use of new allocation and value assignment procedures, 52 behavioral statements and five dimensions were retained. Twenty-eight anchors and five dimensions were selected for use in the BARSSPET.

In the second part of this investigation, 176 secondary students rated seven physical education teachers using the BARSSPET. The data were collected from classes at two schools. These rating data were used to evaluate the psychometric qualities of the BARSSPET.

When compared to School One data, the ratings from School Two exhibited greater leniency and smaller levels of halo and central tendency bias. Test-retest procedures indicated the scale reliability was low.

A type of discriminant validity was ascertained by Pearson Product Moment correlations of student satisfaction scores and BARSSPET ratings. Self-satisfaction appeared to be independent of the students' ratings on the BARSSPET with correlations ranging from .058 to .314. The correlations of the other dimensions of satisfaction to the BARSSPET subscales ranged from .160 to .450. These results indicated

some degree of independence between the BARSSPET and student satisfaction. Because of the iterative process of scale development, the investigator claimed the presence of content and construct validity for the BARSSPET.

ACKNOWLEDGEMENTS

Many persons have contributed to the completion of this study. The valuable assistance and cooperation of these persons are acknowledged at this time.

The investigator wishes to express sincere appreciation to his committee members, Dr. Gail Hennis, Dr. William Purkey, Dr. Marie Riley, and Dr. Marian Solleder. The committee provided much guidance, support, and needed criticism during the development of the BARSSPET.

The support of the administrators, principals, and physical education teachers in the Cumberland County and Robeson County School Systems was invaluable. Special appreciation is extended to my principal, Mr. Ed Baldwin, for his patience in granting the investigator frequent early dismissal to travel to Greensboro and to the schools where data were collected.

The investigator also wishes to acknowledge the help and support received from the panel of judges. The panel of judges willingly accepted a time consuming task without hesitation. These individuals included Dr. James McKethan, Dr. Thomas Martinek, Dr. William Purkey, Ms. Arcelia Jeffreys, and Ms. Jane Brumbaugh.

The assistance provided by the statistical consulting center was essential to the completion of this study. The investigator is grateful for the consultation provided by Dr. David Ludwig and Ms. Robin Panneton.

It is difficult to express the investigator's appreciation for the work of his advisor, Dr. Sarah Robinson. In addition to her able and creative criticism of the BARSSPET development, Dr. Robinson's compassion and moral support were uplifting.

Finally, the investigator wishes to recognize and express his love and appreciation for his faithful and patient wife, Cindy.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

INTRODUCTION

A number of factors influenced the decision to develop a Behaviorally Anchored Rating Scale for Secondary Physical Education Teachers (BARSSPET). Support for development of the BARSSPET was found through identification of (a) the investigator's belief in the need for evaluative processes, (b) concern for the present sources of data for teacher evaluations, and (c) an understanding of the nature of behaviorally anchored rating scales (BARS).

Needs promoting evaluative processes. The logic supporting the use of evaluation is quite prevalent today, appearing in both lay and professional literature. The literature suggests that performance evaluation seek to achieve two purposes. The first purpose of the evaluation rests with supervisory judgments which are used for making administrative decisions about employees. The second purpose of the evaluative process serves as a developmental function for the individual (Cummings and Schwab, 1973).

Typically, the needs which promote evaluation include economic considerations, accountability (Popham, 1975), and the provision of information for making educational decisions (Bolton, 1973). Beggs and Lewis (1975) provided a concise illustration of the accountability rationale in the following statement.

> ...the tax payers and others responsible for the funding of an educational institution require that educators be able to account for expenditures. That is to provide an illustration that students are deriving the desired benefits from the money that is being spent (p. 6).

Bolton (1973) identified additional purposes served by the judgmental appraisal. The first of these purposes is concerned with administrative functions. Bolton suggested that the evaluation serves as a source of information for use in the modification of assignments which may include placement in another position, reduction of load, promotion to a leadership position, or termination of employment. An additional function of evaluative appraisals is explained by the desire to protect the system and its employees from incompetence, whether from capricious administrators or inept teachers. Finally, Bolton suggested that the evaluation, functioning as a judgemental device, serves to recognize superior performances and to validate the selection process.

The second general category of the performance appraisal identified by Cummings and Schwab (1973) focuses upon employee strengths and weaknesses. Bolton (1975) identified three specific areas to which the developmental evaluation may be addressed. These include improvements in (a) the teaching system, (b) the teaching environment, and (c) teaching behavior. Additionally, Bolton has stated that

evaluations provide a basis for the employee's career planning and growth and development.

There exists one additional factor which not only promotes the evaluative process, but also promotes research activity in teaching behavior. Doyle (1981) stated that the need for evaluation may provide a justification for research on teaching. Also, Hall (1979) recommended in Research and Development Agenda in Teacher Education that research in teaching should be extended and that the existing knowledge base about teaching should be considered in terms of its implications for teacher education practice.

Student evaluations. Since secondary students have been excluded from participation in evaluative processes (Nation's Schools, 1970; Sullivan-Kowaski, 1978), the involvement of secondary students in scale construction and administration tasks seemed to be a fruitful topic for research. The use of data from secondary students to construct the BARSSPET was an additional supporting factor for this investigation.

Students in physical education, as in other subjects, interact with their teachers on a daily basis. Hence, a desirable source of data might be found within the student population. Popham (1975) has suggested that the desirability of including secondary students in the evaluative process stems from the fact that the students have access to a greater data base than most administrators.

Data, describing the accomplishment of educational goals may be obtained from students. These goals include (a) the development of motivation for continued learning, (b) areas of rapport, (c) degrees of communication, and (d) the existence of problems between students and instructors. Previous research focusing upon performance appraisals have indicated that secondary students might provide similar information (Denton, et al., 1976; McKethan, 1978; and Wilkinson, 1979).

The goals and objectives that are typically associated with physical education programs may promote the necessity of teacher performance appraisals by students. Marks (1976) indicated that evaluation of teacher behavior is appropriate when judging the quality of teaching in cases where pupil achievement cannot readily be measured. Marks cited for his examples the problems of accurately and immediately assessing the objectives of "Citizenship" and "Health". In physical education, it is conceivable that the representative objectives of "Discovery and development of psychological potential" and "Clarifying values of gaining and maintaining physical health" promoted by the National Assocation for Sport and Physical Education (1979) may present problems in assessment similar to the examples cited by Marks (1976).

Concerns may be expressed when an investigator chooses to utilize secondary school students in the development and administration of a rating scale such as the BARSSPET. One

measurement concern is with questions of reliability and validity of student information. Grasha (1977) suggested that student ratings of faculty are both reliable and valid. However, Masters and Weaver (1977) cited conflicting research concerning the validity and reliability of student ratings. In their own study, using the Student Observation of Teachers and Teaching Techniques Instrument, Masters and Weaver found that students of the two genders rated teachers differently. The results of McKethan's study (1979) supported Masters and Weaver's (1977) conclusions regarding the inconsistent nature of student data. In physical education, McKethan (1979) found that attitudes toward instructional processes were not significantly different according to first semester letter grades. However, the ratings from the two gender groups were significantly different. Thus, there would appear to be some reason to study further rating difference patterns in the high school age group.

Although research findings indicate that students and administrators may have similar perceptions of performance effectiveness (Denton, et al. 1976), a second theoretical concern arises from the idea that students' perceptions will differ from those of teachers and administrators. Students, because of their collective experience (Purkey, 1978) may own a different perspective of behaviors occurring in the classroom, just as the teachers' perspective may be different from that of the administrator. Wilkinson (1979) suggested

that student perceptions of teaching behaviors are limited to the more visible aspects of teaching skill. The concept that one is "really" measuring only differences in perceptions is further supported by Whitney and Doyle (1976) in which they stated:

> ...many studies may have identified dimensions which were implicit in student raters rather than in the instructors (p. 241).

Even though differences may exist in the perception of teaching performance, the utilization of student input would provide greater scope in the evaluative process. Bolton (1973) stated that the standards for teaching success should be acceptable to those people who are affected by teacher behaviors, including students. Similarly, Blood (1974) suggested that utilization of different points of view regarding standards of teaching behavior creates a more comprehensive rating instrument.

Rationale for the BARS. Other considerations which supported the development of the BARSSPET are the purported advantages of the BARS. These advantages are found in the following list:

1. Behaviorally anchored rating scales are based upon observable behaviors.
2. Behaviorally anchored rating scales are constructed by populations similar to those who will be using the rating scale.
3. Anchors to the scales consist of non ambiguous qualities and incidents (Bernardin, et al., 1976; Schwab, et al., 1974; and Smith and Kendall, 1963).

4. Teachers who are to receive ratings from the BARSSPET can have a clear and concise understanding of the expectations that are associated with the instrument.

5. As a result of its multi-dimensional categories, the BARSSPET could be a comprehensive guide to ratings of teacher behavior.

6. Instrument construction, behavioral anchors, and dimensions that are found in a behaviorally anchored rating scale promote its applicability as a guidance and remediation instrument (Blood, 1974).

Finally, support for the development and use of the BARSSPET is evidenced by criticism of the widespread use of evaluation instruments designed for classroom teachers which are utilized across all teaching areas, including physical education (Oliver, 1980). Support for changing this practice, expressed by Oliver, is found in research on teaching behaviors which suggests that effective teaching behaviors appear to be situation specific (Graham and Heimerer, 1981).

The findings from recent research investigating whether or not there are both generic and specific teaching variables are inconclusive. MacDonald and Elias (1970) examined teaching behaviors within differing contexts. The contexts included math and reading in the second and fifth grades. Their results indicated that there were no performance variables which were significant indicators in both grade levels and subjects. However, Berliner (1975) investigated generic and specific variables in contexts identical to the

MacDonald and Elias (1970) study. Berliner's findings indicated that there were 21 variables which discriminated between more effective and less effective teachers in second grade reading, second grade math, fifth grade reading, and fifth grade math. Rosenshine (1976) reviewed two studies in which teacher behaviors were examined in similar contexts. Each of these studies produced conflicting results about the existence of generic teaching skills. The conclusions of Rosenshine (1976) seem to indicate support for the continued development and administration of subject specific instrumentation. His conclusions were as follows:

> The differing results are puzzling. One can argue that differences in coding procedures, selection of sample, and length of instruction are so large that comparing these (sic) studies are meaningless. At any rate, additional studies are essential before we are clear about which skills are generic, grade level specific, or subject area specific (p. 64).

Thus, there is no compelling professional agreement to suggest that "generic" evaluations are preferable to subject specific scales such as the BARSSPET will provide.

Statement of the Problem

The purpose of this investigation was to use student and teacher input to develop a rating scale to obtain student feedback on teaching behaviors of secondary physical education teachers. The investigation included the following phases:

1. To identify and select behavioral incidents that are associated with the teaching of secondary school physical education.

2. To administer the BARSSPET in order to evaluate the psychometric qualities of validity and reliability of the tool.

3. To assess the independence of ratings of teacher behaviors (using the BARSSPET) by comparison with a measure of student satisfaction (Kneer, 1972).

Definition of Terms

The following terms are defined from the context of their use in this investigation.

BARSSPET. The behaviorally anchored rating scale for secondary physical education teachers constructed using retranslation procedures (Smith and Kendall, 1963).

Behaviorally anchored rating scale - BARS. A performance evaluation instrument which incorporates, in each dimension, a range of observed teaching actions to ascertain performance effectiveness as perceived by the raters.

Behavioral incidents. Teacher actions in the context of instruction, illustrating a continuum ranging from effective to ineffective, that are identified and selected by the student population.

Dimension. A category descriptive of a homogenous group of teaching behaviors.

Effective behaviors. Teacher actions which, according to the raters, promote the goals and objectives of the secondary school physical education program.

Ineffective behaviors. Teacher actions which, according to the raters, do not promote the goals and objectives of the secondary school physical education program.

Rater population. Secondary physical education students and teachers in the Cumberland County and Robeson County schools in North Carolina.

Retranslation. Procedures in which a second independent group of students is provided with dimensions (and definitions) and critical incidents and asked to assign each incident to a dimension (Schwab, et al., 1975).

Secondary physical education student. High school students participating in physical education classes as a requirement for graduation.

Assumptions

The assumptions upon which this research was based are listed as follows:

1. Teaching behaviors can be identified, described, and placed into dimensions by secondary physical education students.
2. The dimensions of the BARSSPET represent the major categories of teaching behaviors found in current literature and in the perceptions of secondary physical education students.
3. Scales used to evaluate each dimension adequately sample the full range of behaviors related to that dimension.
4. The students and teachers who will be identifying behavioral incidents are able to discriminate between observable incidents and general impressions of behavior.

Scope of the Study

Due to the nature of this investigation, the scope will be explained in two parts. First, the scope of the instrument will be considered in terms of establishing the criteria for distinguishing between effective and ineffective teaching behaviors. The second portion of this section will be a listing of the parameters for procedures used to develop the BARSSPET.

Behavioral criteria. Colvin and Roundy (1977) and McKenna (1981) suggested that evaluations of the level of performance of any endeavor should be based upon how nearly the expressed goals and objectives of that endeavor are achieved. Since the goal of this investigation was the construction of a rating scale, it is appropriate that the criteria for ascertaining effective and ineffective teaching behaviors be identified.

A sampling of the curriculum materials in physical education literature usually reveals five to eleven objectives outlined for physical education programs. Although sets of objectives are stated in differing terminology, the underlying meanings appear to be essentially the same. In a position paper published in 1979 the National Association for Sport and Physical Education has identified desirable objectives for secondary physical education programs. These objectives focus upon (a) development of personal skills, (b) development of physiological and

psychological potentials, (c) positive social behavior, (d) values clarification promoting a healthy lifestyle, and (e) an understanding of the mechanics of movement and the effects of exercise.

The Curriculum Guide published by the North Carolina Department of Public Instruction (N.C.D.P.I.) (1979) listed similar objectives for physical education programs. These objectives include (a) to develop efficient and effective motor skills, (b) to develop and maintain the best possible level of physical fitness, (c) to develop an interest and skills for recreational programs, (d) to develop desirable social behaviors, and (e) to develop interest and proficiency in using skills. Finally, the Self Study Reports from three Cumberland County (N.C.) high schools contained program objectives that were congruent with those listed in the preceeding paragraphs (Cape Fear Senior High School, 1982; Douglas Byrd Senior High School, 1974; and Seventy-First Senior High School, 1972).

The program objectives found in the Self Study Reports for the three high schools, where the data were gathered, were used to promote a uniformity in asking participants to ascertain the effectiveness of teaching behaviors. When judging the effectiveness of behavioral statements, students and teachers were asked to consider whether the observed behavior promoted or inhibited a set of representative program objective.

Procedural boundaries. The boundaries for this investigation are found in the following paragraphs.

The students contributing data for the development of the BARSSPET were limited to the population of secondary physical education students in the Cumberland County and Robeson County (N.C.) Schools. Teachers contributing data for construction of the BARSSPET were limited to the Cumberland County Schools. The raters who responded to the BARSSPET were limited to the Cumberland County Schools.

The data used for development of the BARSSPET were limited to secondary physical education teaching behaviors. The edited data deliberately precluded administrative or coaching behaviors.

Data were collected for the development of the BARSSPET in the 1983-1984 and 1984-1985 academic years. Administration of the BARSSPET (to ascertain its psychometric properties) took place in the 1984-1985 academic year. Limits to generalization of specific results beyond the participant population are recognized.

Significance of the Study

The importance of the development of the BARSSPET originates from its potential application in two areas of performance appraisal. First, the efforts represented by BARSSPET construction and evaluation contributed to the body of knowledge in the evaluation of secondary physical

education teaching performance. Secondly, the issue of secondary student involvement in performance appraisals was addressed.

The BARS technique has been developed for a wide range of applications, including the evaluation of grocery clerks (Fogli, et al., 1971), college teaching (Harari and Zedeck, 1973), dormitory resident assistants (Knouse and Rodgers, 1983), police officers (Landy, et al., 1973), and nurses (Smith and Kendall, 1963). The only BARS application for use in the public schools was developed by Price (1978) for the evaluation of special education teachers. Through 1984, there were no BARS tools developed for the evaluation of physical education teachers in the public schools. Hence, an important aspect of the development of the BARSSPET lay in its potential to provide an instrument that is focused directly to the public school physical education teacher. The development of the BARSSPET holds the potential to reduce the practice of subjecting (Oliver, 1980) special area teachers to evaluation exclusively by instruments which are designed for traditionally bounded classroom use.

Although the merits of student input for evaluations are identified in the literature (Norris, 1980; and Popham, 1975), students in public school physical education typically are not involved in the process. Grasha (1977) indicated that many teachers distrust students when it comes to providing input concerning their teaching. The inclusion of

student samples in the development and administration of this rating scale is expected to uphold opinions and research claims regarding the advantages of securing valid and reliable student evaluations while at the same time (because of the method of tool construction) reduce expressions of teacher distrust. Teacher evaluation should be a comprehensive process (Norris, 1980); hence, the development and administration of the BARSSPET introduces an important data source.

REVIEW OF THE RELATED LITERATURE

The review of literature presents topics associated with the development of the BARSSPET. These topics include research addressing (a) the quality of student evaluations of educational processes, (b) teaching behaviors in varying contexts, (c) applications of behaviorally anchored rating scales (BARS), and (d) comparisons of the BARS methodology to other rating formats.

Student Evaluations

From a survey conducted in 1970, Shaw indicated that apparently the use of student data historically has not been valued by school officials as a source of information. The survey, distributed to 14,000 administrators, indicated that only 4.5 percent of the respondents reported an annual evaluation program incorporating student data. The limited utilization of student data may stem from questions about reliability and validity of student information.

Many educators question the reliability of student data when evaluating teacher performance (Eastridge, 1976). In rating teacher effectiveness, objections have focused upon students' age, maturity, and lack of experience as impediments to reliability. Presumably, research has shown

that these factors are related to a low level of stability in student's ratings of adult action. Also, there are concerns that the variables of grade expectations, subject matter, class level, and grade point average may negate the discriminant validity found in student assessment (Zakrajsek, 1978).

However, support for use of student evaluations can also be found in the literature. Thompson (1975) found that sex, year in school, grade point average, and expected course grade were not related to student ratings of teacher performance. The following sections present research findings on sources of measurement bias most commonly associated with student ratings.

Attitudinal set. Closely related to the evaluative process is the issue of student attitudes. This section is concerned with student attitudes toward the rating process, attitudes toward school, and any influences of attitudes upon the process of teacher rating.

According to Traugh and Duell (1980) students feel that their evaluations can have an impact on the ways in which teachers teach. Four hundred and eighty-one junior and senior high school students responded to an eight statement scale regarding their roles in the evaluative process. A Chi-square test of independence indicated that (a) that there were no differences between grade levels and responses, (b) that students do feel that their opinions influence the ways

in which teachers teach, and (c) that evaluations are not a waste of time. Attitudes toward instructional processes in physical education were the focus of McKethan's (1979) dissertation. Among 278 tenth grade students in southeastern North Carolina, males and females demonstrated significantly different attitudes. Female students did not like lectures to the class nor did they like lectures on game strategy. The attitude of male students were related to the content of the teacher's verbal behavior and its target. Male attitudes also varied, depending upon the sex of the instructor and the activity. Also, McKethan found that there were no significant differences in attitudes when compared to first semester grades.

Masters and Weaver (1977) found among 925 tenth, eleventh, and twelfth grade students significant relationships between attitudes toward learning, school, and teachers and the kinds of ratings given. Smith and Brown (1976) obtained similar results with 436 students in grades seven through twelve. Here, students responded to two instruments, Attitude Toward Teaching (ATT) and the Course and Instructor Rating Scale (CIRS). Students displayed a strong relationship between ATT II (a general attitude about teachers as presenters and facilitators) and all but one of the CIRS factors (overall challenge). Just as attitudes may vary according to gender (McKethan, 1979) the gender factor may be involved in the assignment of ratings to teacher actions.

Thus, some research findings indicate that students have favorable attitudes toward the rating process and its impact upon teaching practices. The results found in student ratings appear to be linked to attitudes toward learning, school, and teachers.

Gender and Ratings. The gender variable is one of the characteristics most often studied. Landy and Farr (1980), in a review of performance appraisals, suggested that there is no consistent effect of rater sex on ratings obtained in an educational context. However, there were indications that student gender effects the response to different aspects of teacher behavior.

McKeachie, et al. (1971) examined the relationship between teacher ratings and test results, gender, and attitudes. The authors found that women rated effective teachers higher on "skill" and "structure" dimensions. Five hundred and seventeen business law students participated in a study to identify effective and ineffective behaviors of secondary business law teachers. After the behaviors were identified and placed into dimensions, Wilkinson (1979) administered a Chi-square test of independence to ascertain the relationship between perceived teacher behaviors and student characteristics, including gender. Wilkinson found that female students recognize teacher behaviors dealing with student-teacher interactions and the effective use of supplimentary materials and activities. Male students more

readily perceived behaviors related to student discipline, control, and student evaluations.

In another study, the responses of secondary students to the Pupil Evaluation of Teachers scales were different with respect to gender. Biggs and Chopra (1979) used ANOVA techniques to identify a significant main effect between student sex and the "teaching clarity" dimension. Girls rated teachers, regardless of teacher sex, as clearer than did boys ($p<.02$). Male and female students in tenth grade English and mathematics classes responded similarly in ratings of their teachers.

Some investigators have suggested that there is no consistent effect of gender upon ratings. However the literature illustrated that some dimensions of teacher behavior may be perceived differentially by the two gender groups.

Race and Ratings. Another potential source of bias in the student evaluation of teachers may be found in social and cultural differences that are associated with racial groups. Yet, the race variable does not appear to be investigated to the same extent as other variables. Patrick (1978) attempted to identify personality variables that may be attributed to effective teachers. Using a working definition, a multi-racial sample (N=308) of secondary students in three Utah school districts identified a most effective teacher. The selected teachers responded to the Edwards Personal

Preference Schedule (EPPS). The data were analyzed using means and standard deviations. The data indicated that different racial groups perceived the personality variables attributed to effective teachers to be of almost equal importance. A similar study conducted by Sizemore (1981) yielded results of a differing nature. In Sizemore's study, ninth and twelfth grade Black and White students (N=480) from eastern Virginia attempted to identify the most important differences between effective and ineffective teachers. Analysis of variance procedures indicated racial differences in the perception of warm teacher behaviors ($p < .001$), well organized teacher behaviors ($p < .016$) and stimulating behaviors ($p < 001$). Specifically, behaviors which are well organized, sympathetic, and stimulative were perceived as most important by White students in the ninth and twenth grades. Black students more frequently perceived warm behaviors. As with gender and attitudinal set, the role of race in the evaluation process has not been clearly established.

Grades and Ratings. In the evaluation of secondary and college teaching performance, some researchers have focused upon the relationship of assigned grades to performance rating. Thompson (1974), in support of the stability of student ratings suggested, without documentation, that secondary student evaluations are not related to assigned grades.

McKethan (1979), in his investigation of student attitudes toward the instructional processes in physical education, found that there were no significant differences in attitudes when compared according to first semester letter grades. Wilkinson's (1979) results paralleled McKethan's findings where anticipated grades by secondary students were not related to the perception of effective and ineffective teaching behaviors.

Among college students, Frey et al. (1975) found that grade point average did not covary with respect to ratings assigned to instructors. However, with 436 secondary students, Smith and Brown (1976) found that student anticipated grades were significantly related to the ratings that instructors receive.

Temporal stability. Temporal stability represents another criterion which must be achieved for ratings to be considered for evaluative purposes. Grasha (1977) defined temporal stability as "...ratings given on one occasion tend to correlate well with those given on a second occasion (p.24)."

Data indicative of the stability of ratings produced by secondary students were provided by McKethan (1979) in his procedures for developing an attitudinal assessment instrument. In his research, 278 students responded to a 75 item pool of statements regarding instructional processes in physical education. After a five week interval, students'

second responses showed a .72 correlation with the first rating. Similarly (Masters and Weaver, 1977), secondary students responding to an instrument designed to provide teacher feedback produced a high agreement with a second rating occuring after a one month interval. A .70 correlation ($p.<.01$) was obtained on the test-retest check for reliability. Steele, et al. (1971), in a pilot study to develop an instrument to assess meaningful dimensions of educational climate yielded high test-retest correlations. After a two week interval, the authors readministered a 25 item class activities questionaire to six junior and senior high school classes. On the four subsets of the questionaire, the test-retest correlations ranged from .59 to .91. Additional evidence of the temporal stability of student ratings was provided by Noble and Cox (1983) in the development of an instrument to assess instructional effectiveness of lifetime sports classes. The rating scale, consisting of 18 items was administered to a class (N=32) on two occasions with a one week interval. With the exception of two of the items, the reliabilities of all individual responses were above .70.

The literature surveyed provided indications that not only do student ratings of their own attitudes remain stable over time but also their ratings of instructional climate and teacher behavior can be shown to remain stable. The last

factor to be considered in the quality of student ratings is concerned with the question of validity.

Validity of student ratings. Generally, validity refers to the ability of an instrument to measure what it is purported to measure. Grasha (1977) provided a concise definition and description of obtaining indications of validity.

> In its most general usage, validity is supposed to measure whether an instrument measures what it is supposed to measure. That is, do the student rating questionaires really measure instructor characteristics and ability adequately? One way to test this is to see what student ratings relate to. If they really measure instructor characteristics and ability, then one should be able to demonstrate relationships between student ratings and external criteria for these characteristics and abilities (p.24).

In research using secondary students as raters, validity is usually ascertained by comparing student responses to responses of teachers. In some instances, the validity has been "alluded to" while in other cases a specific criterion was established in order to clearly ascertain the level of validity.

In Caruso's (1982) efforts to identify enthusiastic teaching behaviors, secondary physical education teachers reported their own enthusiastic teaching behaviors while students attempted to report the same behaviors. From the incidents provided by students and teachers, twenty categories of enthusiasm were identified. Of the seven highest ranked categories, students and teachers agreed upon

five (Participation, humor, encouragement, praise, and momentum). In a study of instructional climate, Steele, et al. (1971) provided a similar assessment of validity. Student observations were compared to teacher data and to observations using the Flanders Interaction Analysis System. The authors recorded responses to the question, "on the average, the teacher talks how much of the time: 90%, 75%, 60%, 40%, 25%, 10%?" The median student estimate was within five percent of the actual talk in 30 percent of the cases and within 10 percent of the actual talk in 58 percent of the cases. In contrast, no teacher estimates were within five percent of the actual talk. Only 16 percent of teacher estimates fell within 10 percent of the actual talk.

Tuckman (1970) examined the validity of student ratings by comparing their scores on the Student Perception of Teacher Style (SPOTS) instrument to teacher judgments on the Observer Rating Scale (ORS). A correlation of .53 was obtained (significant, $p < .01$) between the teacher and student ratings. Denton, et al. (1976) obtained evidence of validity in student estimates of student teacher effectiveness. Student scores were compared to the ratings of supervising teacher and college supervisors. The authors offered no statistical values to support their claim, that "...correlation coefficients determined between the eight variables of the instrument and the rating scales completed by university and classroom supervisors signify a significant

relationship between the scales on the instrument and the supervisor rating scale (p.185)."

Finally, Noble and Cox (1983) attempted to establish the validity of college student ratings of instructional effectiveness in lifetime sport classes. They compared mean scores on instructional dimensions with perceived course outcomes. From the first part of their validity study, results indicated that student ratings of instructional methods were predictive of student ratings on course outcomes only in relation to satisfaction with the course and the instructor.

Summary. The literature illustrated contradictory findings regarding the reliability and validity of student ratings. Some researchers indicated that secondary students are both reliable and valid in their ratings of classroom climate, teacher behavior, and course effectiveness. However, McKeachie (1971) suggested that when secondary students are given the task to ascertain effective teaching behaviors, considerations should be extended to the varying agendas that students possess when coming to class. Finally, Smith and Brown (1976) stated that student attitudes toward school, their opinions regarding course difficulty, and the grade which they expect to receive should be considered or accounted for in the interpretation of teacher rating data.

Teacher Behaviors

This portion of the review of literature examines teacher behaviors and scale dimensions similar to those identified in the development of the BARSSPET. The literature exhibited a wide range of topics which are related to the BARSSPET development. This section of the literature review also illustrates the multifaceted nature of behavioral dimensions.

Data gathering instruments which are both categorical and descriptive were found in the plethora of teacher behavior literature (Cheffers, et al., 1980; Lombardo and Cheffers, 1983; Pieron and Hacourt, 1979; and Siedentop and Hughley, 1975). Instruments of this type are utilized by independent observers and are designed to promote organization and objectivity in the recording of data (Amidon and Flanders, 1971). Data from these instruments often must be compared to some criterion variable so that inferences can be made (Phillips and Carlisle, 1983; Bookhout, 1967).

Also, there are instruments which provide rating data generated by students (Biggs and Chopra, 1979; Colvin and Roundy, 1977; and Zakrajesk, 1978). These instruments are designed to provide feedback for teachers. Also, the literature details studies, using these instruments, that compare perceived teacher behaviors to a criterion such as student attitudes, teacher and student characteristics, and class climate (Denton, et al., 1967; and Noble and Cox, 1983).

In the literature there are examples of different systems of classifications of teacher behavior. In this section of the review, three different classifications of teacher behavior are examined which illustrate different methods for deriving a system of categorizing teacher actions.

Observer instruments. Amidon and Flanders (1971) stated that the Flanders Interaction Analysis System (FIAS) provides the teacher with insights about how children perceive their teacher. Similarly, the teacher gains insights into his/her behavior, as well as student behaviors. According to Amidon and Flanders, the FIAS classification provides attention to the amount of freedom that a teacher grants to students. Verbal statements occurring in the classroom are classified into one of three sections: (a) teacher talk, (b) student talk, and (c) silence and confusion. Also, the two subdivisions of teacher talk (direct and indirect) are subdivided into additional categories: (a) accepting feelings, (b) praising and encouraging, (c) accepting ideas, (d) asking questions, (e) lecturing, (f) giving directions, and (g) criticizing or justifying authority. The CAFIAS represents efforts to overcome the verbally based limitations of the FIAS. Nonverbal categories were added which increased the scope of data included in interaction analysis (Cheffers, et al., 1980).

The use of observational systems to describe the effects of teacher behavior upon selected variables was accomplished with the use of instruments which measure dyadic interactions. According to Martinek, et al. (1982) dyadic interaction represents the interaction of teacher behaviors with a single student. Using a modification of the CAFIAS to measure dyadic interaction, Martinek and Mancini (1979) examined the relationship between teacher behaviors and expected student outcomes. Elementary physical education teachers rated their students according to levels of expected performance. The high expectancy group received more encouragement, acceptance of ideas, and analytic type questions than did students in the low expectancy group. In another study with high school physical activity classes, Crowe (1979) examined selected teacher behaviors toward high achieving students and low achieving students. The results showed that the high achieving students received 62 percent more praise than did low achieving students. Contrary to the use of praise, teachers did not differentiate in the use of touch with the high and low achieving groups.

Other instruments have been designed to provide indications of teaching effectiveness. According to Phillips and Carisle (1983) some instruments have failed to provide data about behaviors which can be altered. Bloom (1980) indicated that a significant change in research methodology has been in the focus upon variables which can be changed.

According to Bloom, behaviors are considered to be alterable if changes in practices occur as a consequence of a training program.

Siedentop and Hughley (1975) described the O.S.U. Teacher Behavior Rating Scale as an instrument developed for use in gathering descriptive data on teaching behaviors and for research in the modification of student teacher behaviors. The O.S.U. Teacher Behavior Rating Scale included eight categories: (a) input teaching acts, (b) managerial, (c) monitoring, (d) no activity, (e) positive response to a skill attempt, (f) negative response to a skill attempt, (g) positive reaction to on task behavior, (h) and negative reaction to off-task behavior. The authors did offer evaluative comments regarding some of the categories. Concerning the negative skill attempt, Siedentop and Hughley stated "...attempts should be made to have teachers focus on positive aspects of performance and to deliver more positive feedback (p.45)." Similar comments were given with respect to the eighth category, "negative reaction to off task behavior".

The Physical Education Teaching Assessment Instrument (PETAI) was developed by Phillips and Carlisle (1983). The PETAI was developed as a direct response to the "alterable variable" (Berliner, 1975) concept. The instrument, like the other instruments discussed in preceedinng paragraphs, depended upon an independent observer for its administration.

Unlike these instruments, the PETAI observer responds to a rating scale rather than tallying actions. The instrument contained three teacher behavior categories, including: (a) Analyzing student needs, (b) teacher instructional time, and (c) teacher management time. These categories represented a wide range of behaviors that are found as categories in other rating or data gathering instruments. Such behaviors included (a) flexibility behaviors, (b) presentation skills, (c) monitoring time, (d) feedback behaviors, (e) record keeping, (f) equipment management, (g) organization, and (h) other task categories (time used in tasks other than class organization and instruction).

Observation instruments and rating scales have been used to provide descriptions of what physical education teachers do and to relate these behaviors to critical aspects of the teaching process. Pieron and Hacourt (1979) recorded verbal and nonverbal behaviors to describe the teaching behaviors of physical educators at different grade levels. Data for the descriptions were obtained from 24 male and female physical education teachers from elementary, junior, and senior high levels of instruction in Belgium. The authors found that both male and female teachers talked more frequently as the grade level increased. With females, there was a decrease in evaluative functions as the students' grade level increased. The reverse was true of the male teachers. Organizational behaviors comprised the highest percentages of all teacher

interventions. Behaviors in this category ranged from 25 to 35 percent of all observed behaviors.

Lombardo and Cheffers (1983) used multiple observations with the CAFIAS recording instrument to obtain behavioral data on four elementary physical education teachers. Teachers were observed two times each day over a 20 day period, producing 112,054 tallies. ANOVA techniques were used to assess the variability of teaching behaviors. Only two of the 51 CAFIAS parameters and categories varied significantly on a day to day basis. The greatest frequency of behaviors were represented by information giving, lecture, and teacher direction categories. Also, the teachers in this study rarely required students to utilize higher levels of cognitive functioning.

The studies cited here were similar in that the data were gathered using observational techniques and that the results are purely descriptive. The two studies which follow were also based upon observation by independent observers. However, the data were related to some criterion.

Bookhout (1967) completed one of the first research efforts with teacher behavior in physical education using an observational technique. The author attempted to identify the teaching patterns of 36 ninth grade physical education teachers in North Carolina. To assess class climate, students responded to the twelve item Reed's Pupil Inventory. Teacher behavior data were collected using Medley and

Mitzel's OScAR, a schedule for recording clearly defined teaching behaviors. Two important findings were identified as a result of her analysis. First, the author noted the amount of teacher movement in the classes. Bookhout offered two explanations for teacher movement. Since gymnasiums and fields are larger than classrooms, a physical education teacher must, in order to interact with pupils and to observe, move more. Also, since movement is both the subject matter and the means of teaching physical education, demonstrations require that teachers move. The second result which Bookhout reported was that teacher behaviors which support students and foster their initiative ("inviting" and "accepting") behaviors appear in relationship to supportive class climates. Also behaviors which direct and restrain appeared in relation to defensive climates.

Phillips and Carlisle (1983) used the Physical Education Teaching Assessment Instrument (PETAI) to collect data on the alterable variables of 18 physical education teachers. All teachers taught a 10 lesson volleyball unit consisting of the same skills. A cluster analysis of student volleyball skills achievement was used to categorize the teachers as effective (N=5) or ineffective (N=13). Significant differences between the two groups ($p < .10$) were found for the overall "analyzing student needs" scores as well as four of the six subparts. With the exception of "awareness of skill levels", most effective teachers surpassed the less effective teachers in

"knowledge of content", "use of objectives and testing", "flexibility", and "appropriateness of instruction". The authors also found that effective teachers were significantly different in providing "positive performance feedback", and in "management time" (beginning class, equipment management, organization, and ending class).

The use of observational instruments in examination of physical education teaching behaviors has provided the researcher with the potential to examine behavior in relation to varied contexts and criteria. Also, the use of behavioral instruments has provided the researcher and practioner the means to describe the events in the gym. Observation instruments have allowed the researcher to examine the interaction of teacher behaviors with entire classes or individuals within a class. Various researchers using these instruments have compared their data with teacher characteristics, student characteristics, class variables, and achievement to better understand the effects of certain teaching behaviors. The use of student appraisal of teaching behaviors would provide an additional avenue of feedback for the teacher.

Student ratings of teacher behavior. College and secondary students have participated in the development of rating scales to evaluate teachers and to rate the performance of their teachers. Although the literature reviewed in this section pertains only to college physical

activity classes, there are similarities noted between dimensions and behavioral items found in scales developed and used by secondary students in other subjects.

More than twenty years ago, Isaacson, et al. (1964) identified dimensions of teacher behaviors in an introductory psychology class. The ratings of 1260 students were factor analyzed by sex and semester. The findings showed that there were no factors which were unique to sex or semester. The factors included (a) general teaching skills, (b) overload (the amount and difficulty of course work), (c) structure, (d) feedback, (e) group interaction, and (f) student-teacher rapport. With the exception of the "group interaction" factor, all of the factors were associated with teacher behaviors. Items within the "group interaction" factor were related to the freedom of expression within the class. Wilkinson (1979) identified five categories of critical requirements for successful teaching in secondary business law classes. These included (a) use of supplementary materials and activities, (b) organization and presentation of materials, (c) student participation, (d) student discipline and control, (e) student-teacher interaction and (f) student evaluations.

Biggs and Chopra (1979) used secondary students to develop a seven dimension rating scale for teachers. Dimensions included in their work, but not represented in the previous literature were (a) classroom management, (b)

negative focusing, and (c) interest in teaching. The authors also included two categories identified as "faulty verbalization" and "classroom preparation".

Denton, et al. (1967) focused upon teaching style in the development of an instrument to include student assessment in rating student teaching competence. The categories of the instrument, designed to assess the competencies of the final field experience, included (a) inquiry style of teaching, (b) tolerance of divergent behavior, (c) use of technology, (d) nature of class questions, (e) encouragment of independent thinking, (f) expository teaching, (g) teacher led discussions, and (h) teacher openness.

Noble and Cox (1983) developed an instrument designed for student evaluation of physical education activity courses. Six major categories of instructional effectiveness were included: Stimulation of interest, enthusiasm for subject matter, knowledge of subject matter, preparation and organization, clarity and understandability, and sensitivity to class process.

Systems of classification. Church (1974) developed a catalogue of competencies of physical education teachers based upon a theoretical model of outcomes of physical education instruction. The model of outcomes contained three components: A motor skills component, a physical fitness component, and a social behavior component. For each component, Church identified desirable outcomes for the

student. For example, an outcome for the physical fitness component read as follows, "knows the basic effects of exercise on the organic systems of the body and the exercise benefits of selected recreational sports (p.49)." For each of the components and its associated desired outcomes, Church identified nine behavior catgories and 52 specific behaviors. Each of the behavioral categories is listed below.

1. The physical education teacher is responsible for instructional planning.
2. The physical education teacher evaluates students, the program and self.
3. The physical education teacher participates in the development of the physical education curriculum.
4. The physical education teacher communicates with students, parents, and peers.
5. The physical education teacher performs administrative duties.
6. The physical education teacher directs the learning experiences in physical education.
7. The physical education teacher designs learning experiences that contribute to socialization.
8. The physical education teacher develops programs which enable students to maintain a predetermined level of physical fitness.
9. The physical education teacher designs and conducts intramural programs.

Examination of the preceeding categories illustrated that the scope of Church's Catalogue of Core Competencies extends beyond the instructional arena. Not only did the catalogue include instructional behaviors, but it also

included behaviors concerned with public relations, curriculum development, and program evaluation. Conversly, Oliver (1980) provided an informal listing of behavioral categories which are limited exclusively to instructional processes. Other systems of teacher behavior categories have limited categories only to verbal behaviors (Amidon and Flanders, 1971) while still others may have limited their categories to both verbal and nonverbal behaviors (Cheffers, et al., 1980). Categories for the classification of teaching behaviors were found to be varied. Just as one system exhibits similarities and differences when compared to another system, the categories in two different systems of classification may show vast differences in scope.

The definition of a construct within categories also have been widely variable. Price (1979) categorized "Instruction" as "The ability to implement the educational program through the use of appropriate materials and techniques (p. 43)." Church's (1974) sixth category defined Instruction as directing learning experiences in physical education. To define Instructional activity, Oliver (1980) utilized three separate categories. These included "Explanations" (directions and examples), "Demonstrations" (skill portrayal), and "Feedback" (provision of evaluative information to the student).

For the dimension of "Communication", both Church (1974) and Price (1979) provided essentially identical definitions

within similar contexts. Price defined the Communication dimension as "The ability to exchange information with professional persons, parents, and other adults (p. 43)." Price extended Church's definition by including students in the information exchange process. Price utilized a "Rapport" category to deal with student communication. Oliver's consideration for Communication was apparently in the Explanation, Demonstration, and Feedback categories.

"Management" was another teacher behavior category frequently reported. However, only Oliver (1980) directly addressed a category identified as Management. This category was defined as "The amount of time spent in getting the class organized, taking roll, and passing out equipment (p. 83)." Church (1974) assimulated management behaviors into an "Administrative" category. Administrative behaviors were defined as selecting and maintaining facilities, equipment and supplies and record keeping. Price (1979) identified two categories that encompassed management behaviors. The first category, "Record keeping and recording" referred to the ability to keep accurate and up-to-date records which provide the basis for comprehensive reports. The other category, "Behavior management", represented the ability to establish and maintain an atmosphere conducive to learning.

The intent of the preceeding discussion has been to illustrate the diversity that exists in the classification and definition of teacher behaviors. Some authors have

omitted a category which was retained by another author. In other instances a category was explicitly defined in one system while it was represented as a subcategory in another system. Only rarely can one judge whether categories or dimensions have been established on the basis of a priori judgment or independent statistical factoring.

Summary. Teacher behavior assessment instruments clearly have been the subject of much research activity in education. The scope of the instruments is as varied as the behaviors which are categorized. The scope of the instruments described in this section focuses on the behaviors which are associated with instruction. These instruments provide data for descriptions and feedback, for evaluative purposes, and for analysis of relationships with other class variables.

BARS Applications

The BARS instrumentation, at its inception, was conceived as an alternative to the then current rating scales (Smith and Kendall, 1963). Uses of the BARS have ranged from performance evaluations of laborers to evaluations for professional employees. Also, the literature illustrates suggestions for the use of the BARS (and its developmental procedures) for tasks other than performance evaluation. These suggestions focus upon training programs and the assessment of organizational functions.

The literature provides indications of the appropriateness of the use of the BARS in secondary schools. Alternative uses of the BARS may provide information regarding student expectations of teacher behaviors, student perceptions of rules, or other school functions. Also, the behavioral anchors and psychometric qualities reported in the literature may make the BARS suitable for use in student evaluations of teaching. The first section of this portion of the review considers the literature discussing alternative uses of the BARS.

Alternative uses. Since the BARS contains observations of effective and ineffective job behaviors, recommendations for alternative functions have centered upon job training. Kearney (1979) advocated the use of the BARS as a measure of remediation for employees who are unable to perform satisfactorily due to a lack of role perception. Similarily, Knouse and Rodgers (1981) suggested that BARS anchors would be helpful in eliminating unrealistic expectations for applicants aspiring to become assistants in college resident life programs. Also, Harari and Zedeck (1973) proposed that agreed upon incidents from BARS developmental procedures be used as the basis of training for college instructors.

Behaviors to be used in job training programs should include all behaviors in which there is agreement. The behaviors, according to Blood (1974) should not be limited to those found in the final rating scale. Harari and Zedeck

(1973) offered more specificity in suggesting that behaviors to be included in training programs should possess low standard deviations. Both Kearney (1979) and Blood (1974) discussed the rationale of using ineffective as well as effective behaviors. Kearney offered the following assumption for the inclusion of both effective and ineffective behaviors.

> The underlying assumption is that if more effective ways of doing a job can be distinguished from the less effective, adoption of the former will raise the probability of getting results (p. 227).

Blood (1974) suggested that ineffective behaviors be specified so that trainees may become aware of which behaviors to avoid.

Another use of behavioral items generated in BARS developmental procedures rests with promoting familiarity with the evaluation system (Blood, 1974). Kearney stated that employees, through the use of the BARS, have at their disposal proven behavioral prescriptions for better job performances (1979).

Other nonevaluative functions proposed for the BARS were concerned with two intraorganizational functions. The first function was concerned with the assessment of the level of agreement on specified organizational policies. The second focused upon the accuracy of communication between levels of an organization. Blood (1974) and Fogli, et al., (1971) stated that both tasks are accomplished by ascertaining the amount of variance in the ratings of

behavioral items. When scaled items are generated by managers, item variances on the assigned values might indicate the level of agreement on the appropriateness of certain items. Items with large variances expose areas where organizational policy is unclear (Blood, 1974; Fogli, et al., 1971). Similarity of item ratings given by the members of two different levels of an organization may indicate the effectiveness of policy communication. Discrepant ratings suggest the need for increased communication, regarding the behaviors found in policies being questioned (Blood, 1974). Discrepant ratings may also indicate justifiable differences in perception of policies.

Noneducational applications. Smith and Kendall (1963) originated the Behaviorally Anchored Rating Scale (BARS). The authors' purpose in developing the BARS was to construct a rating instrument that might be used in a variety of situations without losing specificity. The original use of the Smith and Kendall scales was in the evaluation of staff nurses by their superiors. Similar scales for evaluating nursing performance were developed by Goodale and Burke (1975) and Zedeck, et al. (1974).

The Smith and Kendall scales were developed from data provided by nurses (N=623) representing different geographical locations. Groups of nurses submitted observed behaviors, edited the behaviors for expectancy and specificity, categorized the behaviors to dimensions and

assigned values to the behaviors to indicate levels of effectiveness. Of the 141 items generated, a total of 89 were retained in six dimensions. Discrimination of the scales was checked by comparing average ratings assigned to outstanding and satisfactory nurses. Scale reliabilities ranged from .972 to .999. Zedeck and Baker (1972) utilized the Smith and Kendall (1963) scales to evaluate nursing performance. Head nurses (N=9) and supervisory nurses (N=5) rated 98 subordinate nurses. The correlation coefficients between ratings (values ranged from .57 to .47 for the scales) indicated that head nurses and supervisory nurses were in agreement in their ratings of the subordinate nurses on all five dimensions. Unlike the results of Smith and Kendall, the discussion revealed no basis for claiming discriminant validity in the scales.

Zedeck, et al. (1974) used a modification of the Smith and Kendall procedures to compare the development of a BARS using data obtained from supervisory nurses and subordinate Registered Nurses. Both groups generated 420 behavioral examples. Supervisory nurses and RNs each identified seventeen dimensions of performance. Between the two groups, twenty-two dimensions were agreed upon. Of the 240 items remaining in the final pool, 177 items were assigned higher mean values by the subordinate nurses. There were very few examples depicting satfisactory or average performance. Finally, Zedeck, et al. (1974) concluded that head nurses

overwhelmingly evaluated the expected behaviors of RNs to be less effective than the RNs themselves. In another scale development procedure to construct a supervisory ratings scale for subordinates, a BARS, applicable to multiple job descriptions within a hospital setting, was constructed (Goodale and Burke 1975). Twenty supervisors representing nine support services generated 360 behavioral incidents and identified 10 dimensions of work performance. Forty percent of the original pool of incidents survived the retranslation procedure. In view of the diverse range of job behaviors represented in the incidents, an ample number of incidents met both the percentage of agreement and standard deviation criteria. The median standard deviation of the scale values assigned to the 149 items was .75 with a range of .30 to 1.50.

Fogli, et al. (1971) used an interview process to generate behavioral statements from 43 grocery store personnel in order to establish job criteria for grocery store clerks. One hundred and sixty-two behavioral statements were retained (from 251) following category and value assignments. Value assignments of items were completed by 97 personnel from three separate districts. Correlations of the mean item scores by personnel from each district revealed high agreement with the lowest correlation being .97.

Other studies produced BARS applications intended for the measurement of morale and work motivation. Motowildo and

Borman (1977) generated 1,163 examples of morale from 190 U.S. military personnel. Seventy-two incidents were retained in nine dimensions. Forty-seven platoons were rated using the BARS. The interrater reliability ranged from .34 to .70 (median = .55). There was evidence of moderate leniency with the mean ratings on the eight scales ranging from 5.47 to 6.70 on a nine-point scale.

In a similar study, the BARS was used to measure the work motivation of engineers (Landy and Guion, 1970). The instrument contained 45 behavioral anchors in seven dimensions. Two peer ratings on each of the seven scales were obtained for 19 chemical engineers. Also ratings on each of the seven scales were obtained from 141 professional engineers. Interrater reliabilitiies were obtained on each of the scales for the 19 chemical engineers ranging from .51 to .76. For the 141 professional engineers, the range was from .55 to .69. Dimension intercorrelations indicated that halo bias was in effect.

Peer and supervisory scales were developed by Landy, et al., (1976) for the evaluation of police officers. One hundred and eight supervisors contributed 147 behavioral items, of which 80 were successfully allocated to eight dimensions. Two hundred and forty peer officers contributed 164 items of which 110 were successfully allocated to nine dimensions. The resulting scales were field tested on 4,575 supervisors and police officers. Indications of positive

leniency were shown by displacement of means to the positive end of the scales. The median intercorrelations between agencies were indicative of a high degree of halo error on both the peer and supervisory scales.

Educational applications. There have been a number of BARS applications within educational settings. Three scales were developed to evaluate teaching performance while one scale was developed to evaluate college coaching performance. Also, one scale was developed to evaluate resident assistants in a college residence life program.

Harari and Zedeck (1973) utilized 38 students to generate 310 behavioral incidents depicting observed behaviors of psychology instructors. One hundred and ninety-three students placed the behaviors into dimensions and assigned values to each statement. The final scale included 78 behaviors as anchors in nine dimensions of teacher performance. The authors noted that students had difficulty in generating mid-range behaviors. Also, the results of the assignment of values showed dimensions containing few items with mean values between 3.5 and 4.5 (on a seven-point scale).

Zedeck, et al. (1976) developed a parallel form of the BARS constructed by Zedeck and Harari (1973). The parallel form utilized the same dimensions as in the preceeding BARS. Two samples of students rated their introductory statistics instructors using one of the two forms of the BARS. Three of

the four criteria necessary for establishing equivalent forms were established: Equal means, equal variances, equivalent correlation with other variables, and correlation between forms. The fourth criterion was not possible to assess since each subject responded to only one form of the BARS. In a similar effort, (Kaufman and Madden, 1980), 53 psychology students generated 243 behavioral incidents of teaching behaviors. Three hundred and eighty-four engineering and humanities students assigned behavioral incidents to dimensions while 71 students assigned values to each statement. Of the 243 incidents generated, 165 behaviors did not meet the retranslation criteria. Retranslation showed that the raters were unable to distinguish between those behaviors intended to illustrate the dimensions of "knowledge of subject matter" and those of "preparation" and "accessibility" and "responsiveness".

A behaviorally anchored rating scale was developed to evaluate women's coaching performance in a multisport setting. Out of 299 editorially refined behaviors, 123 were retained in ten dimensions. Principal component factor analysis revealed that the 10 dimensions were highly related as dimension intercorrelations ranged from .31 to .72. Psychometric analysis revealed a high degree of leniency and moderate halo effect (Knoppers, 1979). In another BARS developed in the higher education context, Knouse and Rodgers (1981) followed the Smith and Kendall procedure in using 16

resident assistants to develop a scale for resident assistants. Their purpose in developing this BARS was not for rating job performance, but to develop a job description for use in place of the job description developed by the administrator.

Summary. The review of literature indicated that the BARS seemed applicable for a wide range of performance evaluations. The literature also provided indications that the BARS may provide evidence of constructs such as morale and motivation through ratings of behavioral performance. Finally, the literature suggested that the BARS and/or its procedures used for construction may be useful in the assessment of certain organizational functions.

Comparison of Rating Formats

The value of a rating instrument rests with its psychometric properties ascertained through statistical analysis of raters' responses. This section of the literature review examines selected psychometric properties of the BARS in relation to those of other rating formats. One should heed the caution extended by Bernardin, Alvares, and Cranny (1976) regarding the comparison of psychometric properties of rating scales.

> ...In comparing rating formats, the method of scale development should first be examined. If comparable effort has not been exercised to insure equally rigorous scales, the implications of the results are hopelessly confounded (p. 569)."

Comparison of anchors. Nunally (1967) suggested that before rating scales can be utilized, the steps on the scales must be defined. The definitions of scale steps are referred to as "anchors". Rating scales typically have one of three types: (a) numerical, (b) adjectival, or (c) behavioral (Landy and Farr, 1980). Some studies have compared the kinds of anchors that may be used with behaviorally anchored scales while other studies compare BARS anchors to anchors found in other rating formats.

Bernardin, et al. (1976) compared rating scales that were anchored continuously and noncontinuously. Two scales were developed using the critical incident/retranslation technique with the scales differing in the scale continuum. Eighty-one instructors were rated using the two versions of the scales. The authors found no significant differences between the formats with ratings in comparing interrater agreement, discriminability, and leniency error.

In another study comparing scales with behavioral anchors, scales with no anchors, and trait anchored scales, differences were found. Borman and Dunnette (1975) developed three scale types for the evaluation of naval officers. The BARS was developed using the critical incident/retranslation technique. The nonanchored scale used the same fourteen dimensions and definitions found in the BARS. The trait anchored scales used dimensions which were not found in the preceeding two scales. Leniency bias was found to be lower

in the behaviorally anchored format while standard deviation scores indicated that the nonanchored format showed less discriminant validity. Finally, the authors concluded that the behaviorally anchored format was less susceptible to halo bias than a trait anchored format.

Peters and McCormick (1966) compared the reliabilities obtained from job task anchored scales and numerically anchored scales. In their study, judges rated items for a particular dimension using either a job task anchored rating scale or a numerically anchored rating scale. Analysis showed that four of the five dimensions of the job task anchored scale had greater reliability coefficients. Peters and McCormick subjected the coefficients to further analysis to determine the overall significance of the obtained reliabilities. It was found that scales constructed of job-task anchors could be used with greater reliability than scales anchored simply by numbers.

Comparisons to summated scales. Summated rating scales contain a set of items, all of which are considered approximately equal in attitude or value loading. Responses are indicated in varying degrees of intensity on a scale ranging between extremes (Isaac and Michael, 1980). Three studies compared the BARS to summated scales. Two comparisons were obtained from the ratings of college instructors by their students while the other study obtained comparisons from ratings on store department managers.

Bernardin et al. (1976) obtained ratings from 154 college students. The investigators compared the BARS to two kinds of summated rating scales. One summated scale was derived from performance dimensions generated in the dimension development phase of BARS construction. The second summated scale was comprised of behavioral items surviving the retranslation process. The BARS format illustrated greater interrater reliability coefficients. However, there was no significant difference between the three scale formats with respect to leniency bias, halo error and discrimination. In the second study analyzing the ratings of college students of their instructors, the BARS format was compared to an item analyzed summated rating scale and a dimension developed summated rating scale. Data obtained from the ratings of 27 college instructors showed that item analyzed summated rating scales showed less leniency bias than the BARS. In terms of discrimination, no differences were found between the BARS format and the summated rating scales. Also, the smaller standard deviations of mean ratings obtained from the summated rating scales indicated a greater interrater reliability for the summated rating scales.

Somewhat conflicting results were obtained by Campbell, et al. (1973) in their comparsions of the BARS format to the summated rating scales. On the summated scales, the maximum possible rating score was 4.0. Six of the nine summated dimensions yielded means between 3.0 and 4.0. In contrast,

the mean ratings for the BARS format clustered around 6.0 on a nine-point scale which was closer to the midpoint of 5.0. The results also indicated more discriminant validity in the BARS than in the summated scales. Finally, the summated rating scales showed greater halo error in the resulting scores.

Comparisons to mixed standard scales. Blanz, and Ghiselli (1972) proposed the Mixed Standard Scale (MSS) format to control for leniency and halo bias in rating responses. In the MSS format, three items are constructed for each dimension to reflect low, medium, and high amounts of the dimension. All items were randomized in their order of presentation, and the raters responded to the items without knowledge of the items' dimensionality.

Finley, et al. (1977) developed three scales for the evaluation of store managers. A BARS was developed using the critical incident/retranslation technique; behaviorally general scales were derived from the BARS; and the mixed standard scales employed statements used to anchor the behaviorally general scales. Store managers were rated on two separate occasions by first and second line supervisors. The results indicated that there were no differences between rating methods with respect to leniency. In terms of the standard deviations of the mean ratings by first line supervisors, the mixed standard method produced larger scores. Second line supervisors produced no significant

differences in standard deviation scores with respect to rating formats. The authors concluded that the differences between the two supervisor groups were due to learning which occured after the first ratings. Interrater reliabilities were significantly greater for the scales incorporating obvious continua when compared to the mixed standard format. Convergent validity was equal for the behaviorally general and BARS formats. The behaviorally general scales exhibited greater discriminant validity. The BARS exhibited less halo bias than the behaviorally general scales or the mixed standard scales.

Dickinson and Zellinger (1980) examined scale formats in a questionnaire containing BARS, mixed standard scales, and Likert-type scales. The questionnaire was administered to 86 students in a veterinary medicine program. Format comparisons showed that the mixed standard format produced as much discriminant validity as the BARS format. Also, when both formats employed specific behavioral anchors, the mixed standard format was equally desirable with respect to its psychometric properties. Finally, the BARS format was preferred in terms of meeting assessment goals and providing the best feedback to students and faculty members.

Comparisons to graphic rating scales. According to Nunnally (1967) rating scales are usually thought to be presented graphically. The pictorial representation of a

scale allows the respondent to make a mark instead of writing in a number.

The BARS has been compared to graphic rating scales. Keaveny and McGann (1975) compared the ratings obtained from BARS and graphic rating scales. The BARS was developed using the critical incident/retranslation technique. The graphic rating scale was based upon the dimensions and definitions obtained in the BARS procedures. Students of four different professors in seven classes responded to both scales. Support for the contention that the BARS control for leniency was not found in the results. Standard deviation scores for nine of the 13 BARS scales were significantly smaller than the corresponding scales in the graphic rating scales. The authors used three checks in determining that the BARS illustrated superior discriminant validity. Zedeck, et al. (1976) compared the BARS, checklist, and graphic rating scales from ratings of college instructors. As in other studies cited in this review, all three formats were based upon procedures used to develop the BARS. The authors stated that conclusions did not support the purported superiority of the BARS over the checklist and graphic rating formats.

Summary. Comparisons of the BARS with different scale formats has been reported. The data showed mixed results with some studies illustrating psychometric superiority for the BARS while other studies showed no differences or

psychometric superiority favoring other rating formats. However, the studies reviewed in this section did not compare the psychometric qualities of rating formats of instruments designed for use by secondary students.

METHODS

The behaviorally anchored rating scale (BARS) has been applied to many different areas of performance evaluation (Landy, et al., 1976; Latham, 1977; and Motowildo and Borman, 1977). The BARS has been used in colleges and universities for the evaluation of teaching and coaching (Bernardin, 1979; Harari and Zedeck, 1973; and Knoppers, 1978). Development and utilization of the BARS has been promoted by factors which are not found with more conventional rating techniques. Two factors, descriptive of the methodology to be employed, are listed below.

1. The raters' own terminology is used to describe and define specific behavioral examples attributed to performance (Kaufman and Madden, 1980).

2. The scales are more meaningful for the raters because of their unique understanding of the jargon and context found in the behavioral anchors (Borman and Vallon, 1974).

While research that has investigated the nature of student evaluation examines evaluations in colleges and universities, little research has centered upon the qualities that secondary students possess as potential raters. In spite of the lack of research directed toward secondary

student evaluations, secondary students are considered to be a substantial source of data by some educational assessment authors (Popham, 1975).

The procedures used for this research reflect the concerns and issues identified in the preceeding sections. Both purported advantages of the BARS and needs regarding instrumentation for the evaluation of secondary physical education teachers were addressed as research decisions were made about the steps used to construct the BARSSPET. Also, the issues of student development and utilization of evaluation instruments were addressed through the psychometric evaluation of the BARSSPET.

Some procedures used in developing the BARSSPET differed from procedures reported in the literature. In order to clarify the procedural decisions, a description of typical BARS development procedures is followed in this chapter by a description of procedures specifically unique to the development of the BARSSPET. Finally, this section concludes with a description of the BARSSPET and steps used in its psychometric evaluation.

Procedures for BARS Development

Smith and Kendall (1963) first developed the BARS to evaluate nursing performance. Subsequent BARS have incorporated either identical procedures or some derivation of the Smith and Kendall technique. Consequently, the

procedures described in the following sections are the procedures used by Smith and Kendall (1963) to construct the first BARS. The procedures used to develop a BARS are essentially iterative. Work which is performed by one group is checked and revised by another group.

Dimension identification. Dimensions which are to be evaluated are listed by each group. Dimensions which are listed with the most frequency are selected for further analysis. Critical incidents are generated and classified to provide additional dimensions. The terminology of the rater is retained in the dimension labels and critical incidents (Smith and Kendall, 1963).

Defining performance levels. A group, independent of the first group, identifies general statements representing definitions of high, low, and acceptable performance for each dimension (Smith and Kendall, 1963).

Generation of behavioral statements. The scale developers edit the critical incidents into behavioral statements prefaced with expectations of specific behavior (Smith and Kendall, 1963). That is, each behavioral statement is prefaced with "...can be expected to...."

Assignment of statements to dimensions. In this step judges are required to assign the statements to dimensions. Statements are eliminated if modal agreement is not clear in the assignment of statements to dimensions. Also, dimensions are eliminated if the statements are not consistently

reassigned to the same dimension (Smith and Kendall, 1963). In the literature, the modal agreement criterion ranged from 50 percent to 80 percent (Bernardin and Walter, 1977; Fogli, et al., 1971; and Harari and Zedeck, 1973). Bernardin, et al. (1976) investigated the results of scales constructed using varying levels of agreement for item retention. The investigators found that variations of 50 percent, 60 percent, and 80 percent in percentages of agreement did not affect the degree of rating variability across dimensions with ratees. When behavioral statements are consistently assigned to a dimension, the dimension is retained for further analysis. In BARS literature, assignment of four behavior statements constitutes the criterion of consistency (Bernardin, et al., 1976; Fogli, et al., 1971; Knoppers, 1978; and Zedeck and Blood, 1974).

Determination of discrimination value. Another group of judges rates an example of outstanding performance and unsatisfactory performance using the statements. The difference between the outstanding and unsatisfactory peformance is computed for each pair of ratings to determine the discrimination value for each example.

Assignment of values. Each scale, with the general definition, is presented with a list of examples previously judged by other raters as belonging to that dimension. Judges rate numerically each statement according to the desirability of the behavior illustrated. Statements are

eliminated if (a) the standard deviations of the ratings are large or (b) if the distribution is multimodal. Next, statements meeting these criteria are assembled on the scale at points indicated by mean ratings (Smith and Kendall, 1963).

The standard deviation criterion has varied from one BARS study to another. Values for the standard deviation ranged from 1.0 (Bernardin, 1977) to 2.0 (Dickinson and Zellinger, 1980).

In investigations after the Smith and Kendall study (1963) a criterion other than the modal distribution criterion was used. Sometimes between group agreement was established for each behavioral statement by means of a Chi-square test of independence. Behavioral statements which show agreement between racial groups were retained in certain studies (Jordan, 1976; Knoppers, 1978).

Procedures for Development of the BARSSPET

The steps used in construction of the BARSSPET were based upon the procedures developed by Smith and Kendall (1963) and other BARS developers. Part one of this investigation is described by steps used to develop the BARSSPET. These steps include (a) generation of behavioral statements, (b) editing of the behavioral statements, (c) student ratings of the behavioral statements, (d) dimension identification, (e) allocation of statements to dimensions

and assignment of values to statements, and (f) final selection of anchors and dimensions. Part two of this investigation included the collection of rating data from secondary students using the BARSSPET. The rating data were used to ascertain the pyschometric properties of the BARSSPET.

Subject selection. Subjects selected for development of the BARSSPET were students and teachers in the Cumberland and Robeson County (N.C.) School systems. Cluster sampling was used in student selection for generation of behavioral statements, the initial attempt in dimension identification, rating of behavioral statements, and the initial attempt in allocation and value assignment. Teacher samples were randomly selected. A panel of judges was used for the second attempt in identification of dimensions. Whole classes were selected, on a nonrandom basis, for the second attempt in the allocation of behavioral statements to dimensions and assignment of values to statements.

As is usually the rule in research conducted in nonlaboratory schools, random selection of individuals to be included in student samples was deemed not practical because principals rarely will allow students to be called out of classes on a random basis. This assertion was supported by Bicknell (1974) who stated:

> ...populations in general and school populations specifically, are not at the unqualified disposal of the educational researcher (p.34).

The advice of Kerlinger (1973) was heeded that although there are identified limitations preventing the use of random sampling, the cluster sampling procedures retain some of the virtues of randomness (Kerlinger, 1973).

The three high schools from which the cluster samples were drawn were selected on the basis of three factors: (a) replication of the composition of the student population, (b) replication of the gender and military dependent distributions of the student population, and (c) replication of the rural and urban composition of the total student population. In addition, written approval to collect data in schools was secured from the appropriate school agency.

Orientation. Orientation of the student samples was provided by the investigator prior to data collection at each stage. The contents of the orientation detailed the nature of BARSSPET development, the nature of rater responsibilities, and a guarantee of anonymity. Students participating in the data collection returned consent forms, indicating their approval. The text of the orientation and consent form, as approved by the Human Subjects Review, are found in Appendix A.

Comparison of procedures. The following section illustrates the sequence of typical BARS procedures and the proposed BARSSPET procedures.

BARS Procedures	BARSSPET Procudures
A. Dimension identification	A. Generation of statements
B. Generation and editing of of statements	B. Editing of statements
C. Assignment of statements to dimensions	C. Rating of statements
D. Determination of discrimination value	D. Dimension identification
E. Assignment of values	E. Allocation of statements to dimensions
	F. Assignment of values to statements

Procedural alternatives. The initial procedures, steps a) to d), seemed to proceed according to design. However, the results of the first allocation of behavioral statements to dimensions were unsatisfactory. The investigator concluded that the unsatisfactory results were related to earlier procedures used in d) to identify dimensions and allocate behavior statements to dimensions. Consequently, alternative procedures were developed for identifying dimensions and the allocation of behavioral statements and assignment of values to behavioral statements. The course of these research decisions is detailed here and explained further in the findings chapter.

Generation of behavioral statements. In the first step of BARSSPET development, students generated statements which contained observed incidents of physical education teaching behavior. In the generation of behavioral statements students were required to ascertain the effectiveness of each

statement as compared to school physical education goals. To insure an adequate representation of teaching behaviors, a sample of teachers also generated behavioral statements.

Editing of Statements. Similar to the Smith and Kendall (1963) editing procedures, the investigator and a group of students made editoral changes while retaining student terminology in the statements. Unlike typical BARS editing procedures, the behavioral statements were not prefaced with "...can be expected to..."

Rating of behavioral statements. Numerous BARS developmental efforts report problems associated with a lack of midrange anchors (Knoppers, 1978; Landy and Guion, 1970; and Zedeck, et al., 1974). DeCotiis (1978) suggested that this problem is a developmental problem contributing to the BARS failing to realize its purported potential. For this reason, in the present study, ratings of the pool of behavioral statements were used as an initial criterion for item retention (DeCotiis, 1978).

The students rated the original pool of behavioral statements for accuracy in describing effective teaching behaviors. An example of the rating instrument is found in Appendix D. Cronbach's Alpha (Carmines and Zeller, 1979) was used to examine the internal consistency of the ratings.

Dimension identification - initial procedures. This portion of BARSSPET development deviated from typical BARS developmental procedures. There were three reasons for

deviation from the typical protocol. Since time requirements for the development of a BARS have been considered a detriment to its widespread use (Green, et al., 1981), the investigator's approach to dimension development was designed to reduce the hours that are required for development of a BARS. The second reason supporting a different protocol was the abundant descriptions and definitions of teacher behavior in the literature (Beebee, 1980; Church, 1974; Oliver, 1980; Pieron, 1981; Pieron and Hacourt, 1979; and Siedentop and Hughley, 1975). The third reason for utilizing a different method of dimension development was in minimizing excessive student absences from class.

The chosen protocol required (a) a survey of current literature to identify dimensions of teacher behavior and (b) compilation of dimensions with definitions, and (c) the rating of dimensions by students. Ratings of dimensions were based upon the students perception of the dimension's contributions to effective teaching. Average ratings were utilized to determine the retention of dimensions for the retranslation process. An example of the rating instrument is found in Appendix E.

Dimension identification - alternative procedures. In the initial procedure, students were unsuccessful in allocating statements to dimensions. The unsuccessful allocation of statements was attributed, in part, to unfamiliar dimension labels, overlapping dimensions, and

definitions which did not reflect the context of the behavioral statements earlier generated by the student subjects. Consequently, new procedures were developed to identify and define dimensions. These new procedures were used to identify dimensions for the alternative procedures in allocation and value assignment.

Similar to the procedures used by Motowildo and Borman (1977), dimensions were identified and defined from the context of the behavioral statements themselves. The investigator sorted the behavioral statements into clusters and developed dimension labels and definitions from the context of the clustered statements. A panel of judges replicated the sorting process by allocating the statements to the new dimensions. Also, the judges evaluated the dimension labels and definitions. The criterion for retention of dimensions was identical to that used in initial and alternative procedures for allocation of behavioral statements to dimensions which was a minimum of four behavioral statements being assigned with at least a 60 percent agreement among the judges.

Allocation and value assignment-initial procedures. The pool of behavioral statements retained from the student ratings was allocated to dimensions identified by a panel of judges. Statements were retained for further analysis if there were a 60 percent agreement on assignment to a dimension.

Dimensions into which a minimum of four statements were allocated were retained for further analysis.

In the assignment of values, numerical values were assigned to the behvioral statements to indicate the students' perception of the behavior's contribution to the fulfillment of physical education program objectives. The standard deviation and Chi-square test of independence were used as criteria for retention of statements for additional analysis. Mean scores of the value assignments were used to determine scale values for statements used in the BARSSPET.

Allocation and value assignment-alternative procedures. In the initial procedure, students were unsuccessful in allocating statements to dimensions. The unsatisfactory allocation of statements was attributed, in part, to inadequate time during the class period to allocate statements to dimensions and assign values to statements. In the alternative procedures for allocation of statements to dimensions and assignment of values to statements, steps were taken to reduce the number of tasks required of each student during the class period.

Construction of the BARSSPET

Behavioral statements and dimensions which met the criteria in allocation and value assignment were used for construction of the BARSSPET. Within a particular dimension, statements with the same mean value or similar behavioral

content were eliminated. In such cases, the statements with the lower standard deviation value were retained.

BARSSPET scales. BARS literature is not consistent in providing guidelines for establishing an optimal number of scale points. Schwab, et al. (1975) stated that typical scale values ranged from seven to nine points. Bernardin, et al. (1975) found no significant differences in results from BARS applications with different numbers of scale points.

Guilford (1954) stated that scale reliability tends to increase as the number of scale points is increased from two. Nunally (1967) suggested that gains in scale reliability become minimal beyond seven points.

Selection of the seven point scale was based, in part, upon the use of the same by Knoppers (1978) and Zedeck and Harari (1974) in BARS developed for use in an educational context. Placement of the anchors on the scales was based upon the mean value -- to the nearest .25 -- assigned by the sample.

Each dimension was placed on a separate page following the form utilized by Price (1979). A vertical, continuous format was incorporated for each scale. Examination of the literature illustrated differences in the order that anchors appeared on the scales. Knoppers (1978) placed ineffective anchors at the top of the scale in an attempt to suppress leniency errors. In her results, she reported low leniency errors for her instrument. Conversely, Campbell, et al.

(1973) obtained similar results with most effective behaviors being placed at the top of the scale. Most scales illustrated in the literature placed effective anchors at the top of the scale (Campbell, et al., 1973; Fogli, et al., 1971; Motowildo and Borman, 1977; Peters and McCormick, 1976; and Price, 1979). In construction of the BARSSPET, the placement of effective and ineffective anchors at the top of the scale was alternated between scales in order to reduce response sets.

Raters' guide. A Raters' Guide was developed which consisted of three parts. Part one of the Raters' Guide consisted of an orientation which detailed rationale and purposes for the BARSSPET. The second part of the Raters' Guide incorporated a set of Likert type scales for rating BARSSPET dimensions. The BARSSPET was found in the third part of the Raters' Guide.

Part one of the Raters' Guide outlined the merits of the BARSSPET in obtaining ratings of physical education teachers. A brief discourse about the necessity of students providing reliable and valid feedback about aspects of their educational experience was also included.

Part two of the Raters' Guide was a set of seven point Likert scales corresponding to the dimensions in the BARSSPET. Following the argument found in measurement literature, the assumption was that ratings using these global scales would allow the rater to eliminate overall

positive feelings toward the teacher. Consequently, results from subsequent administration of the BARSSPET should result in a reduced halo effect (Lawler, 1967 and Zedeck, et al. 1976). Likert scales of the dimensions contained in part two of the Raters' guide were listed on a single page.

BARSSPET. Part three of the Raters' Guide consisted of the BARSSPET with its dimensions of physical education teaching behavior. Each dimension was placed on a single page. Appearing below the caption identifying the dimension was a statement or short paragraph defining the dimension (Price, 1979). According to Bernardin, et al. (1976) the provision of definitions for scale dimensions promotes decreased leniency and greater discrimination. The seven numerical points found on each scale were anchored by behavioral statements retained from the iterative processes. Behavioral anchors function to aid the rater in assessing the effectiveness of teaching behaviors. For each dimension, the rater judges which scale value best describes the teacher's past behavior.

Administration of the BARSSPET

Part two of this study began with collection of data using the BARSSPET. These additional data were used for the psychometric analysis of the BARSSPET.

The completed BARSSPET was administered to tenth grade physical education students in two Cumberland County (N.C.)

high schools. The BARSSPET was administered in the cafeteria at each of the two schools.

The investigator discussed the concept of rating errors and illustrated precautions for avoiding rating errors prior to the administration of the BARSSPET. The purpose for a discussion on rating errors was to assist in the reduction of errors of leniency, halo effect, and central tendency (Bernardin, 1978; Bernardin and Walter, 1977; Borman, 1975; and Zedeck and Baker, 1972). Bernardin (1978) suggested that raters who can properly identify types of rating errors will be more careful to avoid such errors in the rating process.

Following the discussion of rating errors, the investigator read aloud the written instructions while the raters read concurrently. Upon completion of instructions, the raters responded to the BARSSPET.

Except for one instance, all procedures used in administering the BARSSPET were identical at each of the two schools where data were collected. The procedures differed with the presence of teaching personnel at the rating site. No teaching personnel were present during the administration at School One. However, at School Two, one faculty member was present at each of the four administrations of the BARSSPET, this fact being a condition of obtaining data from the physical education faculty. This fact changed the protocol of describing the data.

Statistical Analysis

Data generated from student responses to the BARSSPET were analyzed using the Digital VAXX 11/780 computer system at the University of North Carolina at Greensboro. Mean scores, standard deviation scores and scoring ranges were generated using the SPSSX (1983) statistical package. Cronbach's alpha analysis (Carmines and Zeller, 1979) was used to ascertain the internal consistency of student ratings of the behavioral statements. The extent of student agreement in the assignment of values to behavioral statements was assessed by the Chi-square test of independence and Fisher's Exact Probability Test (Siegel, 1956). Relationships between BARSSPET dimensions, dimensions of student satisfaction, and scale reliability were ascertained by the Pearson Product Moment correlation technique (Carmines and Zeller, 1979). The statistical analysis generated data for use in the psychometric analysis of the BARSSPET.

Psychometric Analysis

Kerlinger (1973) stated that the rating scale, because of its ease in construction and ease in administration, is widespread in use. Ease of construction and administration, however, exacts a heavy price on the rating scale. The heavy price exacted amounts to a lack of validity due to rating errors which enter into rating measures. Consequently, the

final portion of this chapter considers leniency bias, halo bias, central tendency bias, scale reliability, content validity, discriminant validity, and construct validity.

Leniency bias. According to DeCotiis (1977), leniency error is the response set generally attributed to easy raters (positive) and exacting raters (negative). Statistically, leniency bias is characterized by extreme high or low ratings with little variability. For the BARSSPET, leniency error was ascertained by examination of mean ratings and standard deviation values from each scale.

Halo bias. DeCotiis (1977) suggested that halo error reflects a rater's unwillingness or lack of ability to discriminate among dimensions of job behavior. Accordingly, ratings exhibiting halo error generalize to all aspects of a ratee's behavior based upon impressions of one aspect of the ratee's behavior. In this study, the magnitude of intercorrelations among dimensions of performance ratings provided information for the assessment of the amount of halo error (Motowildo and Borman, 1977).

Central tendency. Central tendency effect is the rating of subjects toward the mid point of the scale (Isaac and Michael, 1980). Central tendency bias in rating scores is reflected by small amounts of variability (DeCotiis, 1977). Examination of standard deviation (dispersal of scores) provided indications of the magnitude of central tendency bias for the BARSSPET.

Scale reliability. The procedures described by Carmines and Zeller (1979) were used to ascertain the stability of BARSSPET scales. This procedure utilized test-retest to obtain indications of scale stability over a time interval.

Discriminant validity. A type of discriminant validity was ascertained by comparisons of the BARSSPET to a measure of student satisfaction (Kneer, 1976). One would expect student satisfaction to vary in ways normally associated with the patterns of young adults. Factors influencing student satisfaction may include interest in physical activity, peers, class environment, and kinds of experiences. One would not expect the BARSSPET scores to be influenced in the same way the student satisfaction is influenced. For the BARSSPET to be independent of student satisfaction, BARSSPET scores and student satisfaction scores should not exhibit parallel trends with respect to leniency, halo bias, and central tendency.

Content validity. The question of content validity for the BARSSPET was considered by examination of the methodology of scale construction (Brown, 1976). First, incidents used in scale construction were solicited from a sample that had observed and interacted with physical education teachers on a near daily basis for a greater portion of a year and perhaps over a number of years. Secondly, the iterative process by which anchors and anchor values were derived contributed to a high degree of content validity. Essentially, the

development and format of the BARSSPET provide logical assurances of content validity.

Construct validity. Construct validity in the BARSSPET was a judgmental function based upon iterative processes in which behavioral statements with homogeneous responses are retained (Brown 1976). These procedures included the analysis for internal consistency of student ratings using Cronbach's alpha and the process of assigning behavioral statements to dimensions by both the judge pool of professionals and the student sample.

Summary

This section represented a summary and brief explanation of the procedural steps used to develop and evaluate the BARSSPET. In a number of instances, the procedures differed from usual BARS development procedures. Findings from each of the developmental steps are explained in the chapter to follow.

RESULTS

This investigation was designed as having two major parts. The data from the first part were collected from secondary students and teachers for the purpose of developing the BARSSPET. Four hundred and eighty-eight students and 14 teachers contributed data for the development phase of the study.

The second part of this investigation consisted of an evaluation of the BARSSPET. One hundred and seventy-six secondary students rated their physical education teachers. Statistical analysis of the rating data was used for a psychometric analysis of the BARSSPET.

Generation of Behavioral Statements

The first step in the development of the BARSSPET was the generation of behavioral statements. The findings from this step are detailed in the following sections.

Subjects. A cluster sampling technique was used to identify four physical education classes in three schools. A total of 114 secondary students volunteered to participate in the generation of behavioral statements. The racial

breakdown of the subject population appeared to be similar to the total population statistics for Cumberland County (N.C.). The percentages for the Black and White segments of the sample were 36.8 and 57.0, respectively. The percentages for Blacks and Whites in the population were 30.7 and 64.1, respectively (Census of Population and Housing, 1980). A breakdown of the grade level and gender percentages appears in Table 1. These latter figures were not compared with population statistics.

Behavioral statements. Six hundred and twenty-nine responses were generated by the students. The instructions required that three neutral behaviors be generated for each two positive and each two negative behaviors. To ascertain the effectiveness of each behavioral statement, students determined whether or not the behavior promoted or interfered with program objectives listed in the instructions. These objectives were gleaned from school reports and documents.

Neutral behaviors represented 35 percent of the total number of behavioral statements. Effective and ineffective behaviors represented 32 percent and 33 percent, respectively, of the total number of behaviors. The greater percentage of ineffective behaviors generated by the students was in contrast to the results obtained in other studies in which the tendency of the raters was to submit slightly more effective behaviors (Knoppers, 1979 and Harrari and Zedeck, 1974).

TABLE 1
Student Sample
Generation of Behavioral Statements

Race	N	%	Sex	N	%	Class	N	%
Black	42	36.8	Female	69	60.5	Soph	105	92.1
Others	7	6.2	Male	45	39.5	Jun	8	7.0
White	65	57.0	----	--	--	Sen	1	.9

Teacher generated statements. Letters were mailed to a sample of 20 randomly selected secondary physical education teachers in the Cumberland County Schools (See Appendix B). Fourteen teachers responded to this request. Table Two provides a description of the teacher sample.

TABLE 2
Teacher Sample
Generation of Behavioral Statements

Race	N	%	Sex	N	%
Black	2	14.3	Male	8	57.1
Others	--	--	Female	6	42.9
White	12	85.7	------	-	--

A total of 106 behavioral statements were generated. Forty-eight percent of the teacher behavior statements represented effective teaching behaviors. Neutral teaching

behaviors accounted for 24 percent of the total while ineffective behaviors accounted for 28 percent.

Editing of Statements

After all of the behavioral statements were generated by the first sample of classes and teachers, an independent group of students, working with the investigator, edited the statements. The editing group consisted of two female students and three male students. The racial composition consisted of one black and four white students. Grade level composition of the editing group consisted of one senior, three juniors, and one sophomore.

The editing process took place in two phases. In the first phase, the investigator eliminated statements which were expressions of opinions or attitudes. Also, repetitious statements were eliminated. This phase of the editing took approximately three hours. The second phase of the editing process utilized the input of the student editing group. The time requirement for the second phase of editing was five hours. The duration of each session was one hour and fifteen minutes.

In the second phase, there were a number of objectives to be met in editing the behavioral statements. First, the group identified and eliminated syntax which they believed would be unfamiliar to sophomore students. Secondly, statements with meanings specific to racial groups or gender

were clarified. Finally, specificity was added to generalized, global statements. Three hundred and eighty-three statements were retained after the editing process.

Dimension Identification - Initial Procedures

Identification. In the original design, dimension identification called for a survey of the literature to identify potential dimensions for the BARSSPET. Dimensions were selected from contexts similar to those found in the BARSSPET. These contexts included dimensions identified by secondary students, dimensions found in secondary teacher behaviors in physical education and other subjects, and dimensions found in ratings of college teacher behaviors in activity courses (Masters and Weaver, 1977; Miller, 1978; Oliver, 1980; Patrick, 1978; Pieron and Hacourt, 1979; Price, 1979; Wilkerson, 1979; and Zakrajsek, 1978). Dimensions thus identified for further analysis included "Grading," "Empathy," "Rapport," "Communication," "Feedback," "Management," "Organization," "Discipline," "Enthusiasm," "Instruction," and "Aggression."

Subjects. The second group of subjects was obtained from a cluster sample of four different classes in three schools. A total of 146 secondary students volunteered to rate each dimension. This group also rated behavioral

statements. The Black and White racial groups constituted the major portion of the second group of subjects. These groups represented 32.2 and 62.3 percent, respectively, of the subject group. In comparison, the local area population statistics for the same racial groups are 30.7 and 64.1 percent, respectively (Census of Population and Housing, 1980). A complete breakdown of the sample appears in Table 3.

TABLE 3
Subjects Rating Dimensions
and Behavioral Statements

Grade	N	%	Sex	N	%	Race	N	%
Ten	130	89.0	Male	55	37.9	White	91	62.3
Eleven	13	8.9	Female	88	60.7	Black	47	32.2
Twelve	2	1.4	M.V.	2	1.4	Other	6	4.2
M.V.	1	.7	----	--	----	M.V.	2	1.4

Note M.V. = missing values

Rating of dimensions. Students rated each of the 11 dimensions gleaned from the literature on its importance to the teaching of physical education. Below the dimension label and definition was a Likert-type, seven point scale. On the scale, the number "one" represented a most important dimension while the number "seven" represented a least important dimension. Students checked the space above the number of their choice. The instrument appears in Appendix E.

Three criteria were used to select dimensions to be used in allocation of behavioral statements to dimensions. Mean and standard deviation values were used as the first criteria. Analysis of variance was used as the third criterion. The mean score for the "Aggression" dimension was 4.39 while the mean scores ranged from 2.44 to 3.12 for the remaining 10 dimensions. The differences in values assigned to Aggression when compared to the other dimensions indicated that the sample did not feel that it was important to teaching. Also the standard deviation scores were examined. The standard deviation score for the Aggression dimension was 2.09. The standard deviation scores for the other 10 dimensions ranged from 1.59 to 1.86. The Aggression dimension was removed from further analysis because of its perceived unimportance and lack of consensus regarding its value. Using an Analysis of Variance procedure, if the observed value (2.19) of the F-ratio (9df, p .05) was less than the critical value of 2.44, the null hypothesis was accepted. Since there were no differences between the measures, the remaining 10 dimensions were used for allocation of behavioral statements.

Rating of Behavioral Statements

Subjects. The sample selected for the rating of behavioral statements was the same sample that participated in initial procedures of dimension identification. This

sample is described in the previous section. Table 3 illustrates a complete breakdown of the sample.

Rating of statements. Following the editing process 383 statements had been retained. Since students would not be able to rate 383 behavioral statements in a fifty-five minute period, the statements were assigned to one of five scales. The scales were identified as "Scale A," "Scale B," "Scale C," "Scale D," and "Scale E".

Participating students rated the behavioral statements contained in one of the five scales. The number of students responding to each scale ranged from 28 for Scale D to 30 for Scale E.

After receiving instructions, students rated the behavioral statements for accuracy in describing effective behaviors of physical education teachers. The behavioral statements were rated on a seven point scale with one being most accurate, four being neutral, and seven being most inaccurate. The instructions and an example of the scales are found Appendix D. During the 55 minute period each student rated approximately 80 items. All students were able to complete the task before the end of their class period.

Statistical analysis. The covariance method of Cronbach's Alpha (Carmines and Zeller, 1979) was used to eliminate behavioral items which did not contribute to the internal consistency of the scales. Table 4 provides a more

complete description of the analysis. Behavioral statements which depressed the Alpha coefficient were eliminated.

TABLE 4
Analysis
Rating of Statements

Scale	Begin items	End items	Begin Alpha	End Alpha	No. of Analysis
Scale A	78	26	.6560	.9249	7
Scale B	61	18	.6900	.8448	12
Scale C	84	39	.8353	.9034	7
Scale D	81	21	.6607	.8534	12
Scale E	79	31	.7943	.8992	12

In Scale A and Scale C, seven applications of the analysis were required to eliminate items which depressed the Alpha. The remaining scales required twelve applications of the analysis. After the first analysis, the Alpha values ranged from .6607 to .7943. The Alpha for each scale following the last analysis ranged from .8448 to .9249. As items were eliminated, the reduction in the scale variance and the increase in the Alpha coefficients indicate increases in internal consistency of the scales.

Allocation and Value Assignment - Initial Procedures

Analysis using Cronbach's alpha to ascertain the internal consistency of student ratings of behavior statements yielded 135 items for allocation to dimensions and value assignments. The ten dimensions previously evaluated by the students were the dimensions to which behavioral statements were to be allocated. Behavioral statements retained as anchors in the BARSSPET must possess statistical specifications indicated by three criteria which are described in the discussion of the results. Also, a dimension must have received at least four statements to be included in the BARSSPET.

Subjects. The subjects for allocation and value assignments of behavioral statements were volunteers from four classes selected in the cluster sampling process. Asians, Hispanics, and American Indians accounted for eight students. The small numbers represented by the racial groups described in the preceeding statement necessitated a label of 'Others' for Chi-square analysis.

Similar to the first two subject groups, the racial composition of the third group approximated the population statistics for Cumberland County (N.C.) (Census of Population and Housing, 1980). A more comprehensive description of the subject group appears in Table 5.

Table 5
Subjects for Statement
Allocation and Value Assignment

Race	N	%	Sex	N	%	Class	N	%
Black	24	30.0	Female	50	62.5	Soph	70	87.5
Others	8	10.0	Male	30	37.5	Jun	6	7.5
White	48	60.0	------	--	----	Sen	4	5.0

Allocation. After statistical analysis using the percent agreement criterion, the data showed that students were not able to allocate behavioral statements satisfactorily to nine of the dimensions identified in the initial procedures. Only three behavioral statements were allocated to dimensions using the 60 percent agreement criterion. All three behaviors were allocated to the "Discipline" dimension. In order to investigate the value assignments alone and agreement between racial groups, behavioral items were considered for retention using a 40 percent criterion.

By this procedure, thirty-five behavioral statements were allocated to six dimensions. Three dimensions did not receive statements. Dimensions receiving no statements included "Rapport," "Feedback," and "Managerial". The "Communication" dimension received only three behavior statements, eliminating it as a category in the BARSSPET. Five categories received sufficient statements to be retained

after allocation. Table 6 illustrates the item allocation and value assignment.

Value assignments. Analysis of the value assigments of the behavioral statements revealed only 14 behaviors which met the 2.00 standard deviation criterion. Six behavior statements were allocated to the "Enthusiasm" dimension. Standard deviation values for the Enthusiasm dimension ranged from 1.72 to 1.99. The "Organizational" and "Empathy" dimensions each retained two behavioral statements following application of the standard deviation criterion. The standard deviation values ranged from 1.83 to 2.07 and 1.76 to 2.09, respectively. Of the eight behavioral statements retained in the "Discipline" dimension, only three met the standard deviation criterion. The standard deviation values for the Discipline dimension ranged from 1.70 to 2.55. One behavioral statement met the 2.00 standard deviation criterion in the "Grading" dimension. The standard deviation values for this dimension ranged from 1.88 to 2.37.

DeCotiis (1975) indicated that a lack of behavioral statements representing the midrange of effectiveness was a major problem with BARS. Prior to subjecting the behavioral statmehts to the standard deviation criterion, there was no lack of behavioral statements representing the midrange. Seven behavioral statements were retained which had mean ratings between 3.20 and 3.80. However, only two of these behavioral statements possessed standard deviation values

TABLE 6
Analysis of Initial Allocation
and Value Assignment

Statement	Dim	%	$\bar{X}$	SD	X^2	Df
15.	Grad	51	3.42	2.08	12.142	4
36.	Grad	40	2.76	1.88	5.995	4
68.	Grad	40	4.48	2.21	1.414	4
69.	Grad	48	4.54	2.15	10.808	4
75.	Grad	53	3.78	2.28	7.333	4
100.	Grad	52	3.28	2.05	14.003	4
108.	Grad	43	4.87	2.37	3.511	4
109.	Grad	47	2.92	2.17	6.677	4
14.	Disc	61	3.61	1.70	.926	4
26.	Disc	40	2.98	1.82	9.097	4
30.	Disc	45	3.43	2.15	14.366	4
43.	Disc	40	3.53	2.37	6.185	6
50.	Disc	51	5.01	2.18	5.229	4
54.	Disc	64	2.80	2.13	17.723	4
82.	Disc	57	5.44	2.55	2.310	4
102.	Disc	61	2.76	1.85	7.794	4
3.	Enth	53	2.75	1.85	16.239	4
25.	Enth	57	2.72	1.72	6.078	4
64.	Enth	45	2.73	1.86	4.207	4

Underlined values are significant p .05

TABLE 6 CONTINUED

Statement	Dim	%	$\bar{X}$	SD	X^2	Df
85.	Enth	43	2.71	1.99	12.073	4
87.	Enth	42	2.78	1.79	14.003	4
105.	Enth	42	2.80	1.89	6.779	4
32.	Emp	44	3.11	2.09	10.331	4
65.	Emp	45	2.45	2.08	8.015	4
70.	Emp	45	3.32	1.76	7.589	4
101.	Emp	44	2.45	1.83	8.351	6
5.	Comm	52	2.95	2.10	17.388	4
13.	Comm	51	3.72	2.16	4.323	4
49.	Comm	40	5.11	2.19	1.772	4
7.	Org	40	3.06	1.83	1.623	4
16.	Org	49	3.09	2.0	12.142	4
19.	Org	56	2.70	2.07	19.908	4
55.	Org	44	2.75	1.94	17.652	4

Note Grad = Grading : Enth = Enthusiasm : Emp = Empathy : Comm = Communication : Org = Organization

equal to or less than 2.00. Generally, among the 35 behavioral statements retained and the remaining 101 behavioral statements, as the mean values approached and exceeded the midrange value, there was less agreement which was illustrated by greater standard deviation values.

The Chi-square analysis from the SPSSX Crosstab procedure was used to determine significant differences between Black and White racial groups. Significant differences were ascertained at the .05 level of significance. Of the 35 behavioral statements meeting the 40 percent criterion, 22 behavioral statements exhibited no differences. Of the 14 behaviors meeting the standard deviation criterion, the racial groups agreed upon 11 with respect to value assignment.

Discussion. There were several factors which could be attributed to the students' inability to allocate statements satisfactorily to dimensions. Factors attributed to the problem were associated with time constraints, student characteristics, and category labels.

The first factor, time constraint, was associated with allowance of inadequate time to complete the task. Students were required to place the 135 behavioral statements into a dimension and assign each statement a numerical value. This task required students to make decisions after studying dimension labels, definitions, and program objectives. Students were required to make 270 decisions in 110 minutes (two 55 minute class periods). Consequently students would

have to make a decision every 24.6 seconds. Most likely, the time factor was a signficant contributor to the inability of students to allocate behavioral statements to dimensions.

The factor of student characteristics centered upon varying perceptions, level of committment, and reading ability. Differences in perception may have contributed to students placing the same behavior statement into different dimensions. The lack of committment by students was illustrated in two ways. Although students had received an orientation and signed a consent form, some students at one school chose not to participate in the task after receiving directions. At another school a large number of students chose not to participate for the second day of the task by being absent from class. Finally, even though the statements had been edited with the help of a student panel, the reading levels of students participating in the task could have contributed to the lack of consensus found in the results of the allocation process.

The factor, dimension labels, along with definitions were identified from the literature describing teaching behaviors. An independent group of students rated the importance of each dimension to teaching. It is possible that the dimensions identified by the second group of subjects were not congruent to the statements generated by the first group of subjects. These results suggest that the dimensions, as defined, were not congruent with the context

of the behavioral statements and/or student perceptions of teaching behaviors in physical education.

Dimension Identification - Alternative Procedures

Initial efforts to identify dimensions of physical education teaching behaviors centered on a review of the existing teacher behavior literature as is explained in Chapter Three. Through the use of those procedures, categories of teaching behaviors were identified. However, students were unable to agree on the allocation of behavioral statements into dimensions.

In an effort to achieve greater consensus in the allocation of statements to dimensions, the investigator utilized a new set of procedures. The development and use of these procedures was based upon the assumption that categories of teaching behaviors might be identified from the pool of behavioral statements which had been already generated and edited and subjected to statistical evaluation. The steps followed are listed below.

1. The investigator presorted into clusters the 135 behavioral statements into clusters which had been retained following Alpha and Chi-square analysis.

2. Labels and preliminary definitions were assigned to each cluster based upon the context appearing in each cluster of statements.

3. The statements and a set of preliminary definitions were sent to a panel of judges who placed the 135 statements into the newly defined dimensions.

4. The investigator examined the feedback provided by the panel of judges and made final decisions on the dimension labels and defintions.

Presorting of statements. In the first step of this dimension identification, the investigator hand sorted the student generated and edited behavioral statements into clusters. Placement of statements into clusters was based upon two considerations. These considerations included (a) the context of the behavior found in the statement and (b) the function(s) of the behavior as perceived by the investigator. The dimension labels and definitions were based upon the investigator's perception of the functions of the behaviors in each cluster.

Through the presorting process, seven dimensions were identified. The labels for these dimensions included "Discipline," "Grading," "Managerial," "Communication," "Enthusiasm," "Teacher sensitivity," and "Instruction." The dimension labels and preliminary defintions are found as a part of the instrument sent to the panel of judges (see Appendix G). The judges were asked to evaluate the labels and preliminary definitions by allocating the 135 statements to the newly identified dimensions, to provide a critique of the preliminary definitions, and to suggest additional dimensions.

The judging panel. The panel of judges consisted of five professional educators, all familiar with the area of teacher behavior. The panel included two University of North

Carolina at Greensboro professors from the School of Health, Physical Education, Recreation and Dance (School of HPERD) and from the School of Education. The third member of the panel is the Director of Exceptional Children's programs in the Cumberland County Schools. The remaining two judges were experienced secondary physical education teachers who were advanced UNC-G graduate students in the School of HPERD.

Each member of the panel was first contacted by telephone. After securing verbal agreement to participate, each judge received a packet by mail. The packet contained a cover letter, 135 behavioral statements, instructions, and worksheets illustrating the dimension labels, preliminary definitions, spaces for listing statement numbers, and spaces for criticizing category labels and definitions. The instrument for sorting and critique is found in Appendix G.

Seventy-nine percent (107) of the statements were placed by the judges into categories with a minimum of 60 percent agreement. All of the dimensions received statements as a result of the sorting process. The Grading dimension received the least number of statements (8) while the Teacher sensitivity dimension received the greatest number (31). Table Seven illustrates the dimensions with the number and percentage of statements assigned.

Only 3.77 percent of the behavioral statements were not placed into any of the seven dimensions. These behavioral statements, identified as counterproductive, were all from

the same judge. Another member of the panel identified a new dimension called, "Respect". Twenty-two behavioral statements were assigned to this category (all by the same judge). Since only one member of the panel identified and/or suggested an additional dimension, no further iterations were undertaken to add to the dimensions.

TABLE 7

Sorting of Behavioral Statements into Dimensions by the Panel of Judges

Dimension Label	Statements Assigned	Percent of items Assigned to dimensions	Percent of total number
Instruction	14	13.0	10.4
T. Sensitivity	31	29.0	23.0
Enthusiasm	16	15.0	11.9
Communication	9	8.4	6.7
Managerial	17	15.9	12.6
Grading	8	7.5	5.9
Discipline	12	11.2	8.9

Criticism of preliminary definitions focused on overlapping concepts between the dimensions. One panel member suggested that an overlap existed between Instruction and Communication. Another panel member suggested that there was overlap existing between Instruction and Managerial. In

an attempt to minimize the perception of overlap and aid in the allocation of behavioral statements in the next step of scale construction using students, generic definitions were added to the preliminary definitions.

Summary. Dimension labels and preliminary definitions were identified through a presorting process conducted by the investigator. The panel of judges confirmed the dimension labels and definitions via their sorting process. As a result of criticisms offered by the panel, generic definitions were added to the preliminary definitions. Seven dimensions were identified, labeled, and defined for the final step of BARSSPET construction. These dimensions and final definitions are found in Appendix H as a part of the instrument used for the step in which students allocated behavioral statements to dimensions.

Allocation and Value Assignment - Alternative Procedures

The final step in identifying anchors for the BARSSPET was accomplished by using students to allocate behavioral statements to dimensions and to assign values to each behavioral statement. The revised instrument design and the procedures used for the second phase of behavioral statement allocation and value assignment were based upon problems that seemed to arise from a lack of consensus in the initial procedures of allocation.

Instrument design. Editing changes were made to simplify the instrument and to increase clarity and readability. First, reductions were made in the number of words and the complexity of words in the instructions. The investigator was aided in this task by a North Carolina Public Schools guidance counselor who edited the first draft of the instrument. The second change was made by placing an example immediately following the directions. Third, a graphic rating scale was placed below each behavioral statement. (See Appendix H).

Procedural changes. Two procedural changes were incorporated to promote adequate consideration for each item. First, the instrument was designed to minimize the turning of pages. This was accomplished by providing the directions and criteria for making decisions on the first page. The second page contained the example, demographic data, and the first of the behavioral statements. The successive pages contained behavioral statements. Students turned a page only after reading the last behavioral statement on the page.

The second procedural change required students to evaluate only 33 or 34 statements, depending upon the version of the instrument to which he/she was responding. Also, different students were assigned to a single task: (a) Either to allocate behavioral statements to dimensions or (b) to assign values to the behavioral statements. Although not

precisely timed, all students were able to finish well within the 55 minute class period.

Subjects. Approval to use subjects in Robeson County was granted by the Robeson County Board of Education. The letter regarding this request is found in Appendix B. Subjects allocating behaviors to dimensions and assigning values were secondary English students from West Robeson High School in Robeson County, N.C. Intact classes were selected by guidance personnel from the Advanced and General English classes in order to account for all academic levels.

The method of subject selection violated the cluster sampling procedures described in earlier sections. Selection of classes according to English department assignment rather than the physical eudcation classes increased assurance that the subjects would be able to complete their task as the reading levels of these groups is generally average or above average. Although the subjects were not necessarily current participants in the physical education program, they had participated in physical education classes in the ninth grade. Students from the basic English curriculum were excluded from this phase of data collection.

A total of 71 students allocated behavioral statements to dimensions. Seventy-two students assigned numerical values to the behavioral statements. A description of each subject group is provided in Tables 8 and 9. Unlike the racial distribution in Cumberland County, the American Indian

race represented the majority in each subject group -- 85 percent and 83 percent, respectively.

Table 8

Subjects Allocating
Statements to Dimensions

Race	N	%	Sex	N	%	Class	N	%
Indian	60	84.5	Female	42	59.2	Sophomore	27	38.0
Black	8	11.3	Male	29	40.8	Junior	24	33.8
White	3	4.2	----	--	--	Senior	20	28.2

Table 9

Subjects Allocating Values
to Behavioral Statements

Race	N	%	Sex	N	%	Class	N	%
Indian	60	83.3	Female	41	57.0	Sophomore	27	37.5
Black	12	16.7	Male	31	43.0	Junior	44	61.1
White	--	--	----	--	--	Senior	1	1.4

The distribution of racial groups in both subject subgroups -- subjects allocating statements and subjects assigning values -- did not approximate the population statistics for Robeson County, N.C. The racial breakdown for Whites, American Indians, and Blacks was 39.5, 35.0, and

25.2 percent, respectively (Census of Population and Housing, 1980). The percentages for the same racial groups, as illustrated in Tables 8 and 9, were overwhelmingly represented by the American Indian group.

Dimensions. The data showed that 73 behavioral statements were allocated to seven dimensions using a 50 percent agreement criterion. All dimensions received more than the minimum of four statements to be considered for further analysis. The dimensions receiving the least number of allocated behavioral statements were Communication and Enthusiasm. Each were allocated six statements. The panel of judges had also allocated the least number of behavioral statements to the Communication dimension by agreeing upon nine statements. Teacher sensitivity and Instruction each received 20 and 15 statements, respectively. In contrast, Teacher sensitivity and Instruction was allocated 31 and 14 behavioral statements, respectively, by the panel of judges.

The students assigned 10 behavioral statements to Discipline while the panel of judges assigned 12 behavioral statements. In the Grading and Managerial dimensions, students assigned eight behaviors to each while the panel of judges assigned eight and 17 behavioral statements, respectively.

Value assignments. A second sample of students (See Table 9) assigned values to each of the behavioral statements. These values represented the perceived

effectiveness of the behavior found in the statement. Behaviors were rated on a seven-point scale with "one" representing an extremely effective behaviors while "seven" represented an extremely ineffective behavior.

Prior to the value assignments, there were 73 behavioral statements on which there was at least 50 percent agreement placement into a dimension. Analysis of the value assignments revealed 15 behavior statements which did not meet the 2.00 standard deviation criterion for retention. Fifty-eight behavioral statements remained after application of the standard deviation criterion. A comprehensive presentation of the dimension allocations and value assignments are presented in Tables 10 through 17.

Both the panel of judges (from the alternative procedures of dimension identification) and the Robeson County student group allocated six behavioral statements to the Enthusiasm dimension. The standard deviation values ranged from 1.59 to 2.27. Four of the six behavioral statements allocated were in excess of the 2.00 standard deviation criterion and were eliminated. Consequently, the dimension of Enthusiasm and its two remaining behavioral statements were eliminated and received no further consideration in the construction of the BARSSPET. Additional information on the Enthusiasm dimension is found in Tables 10 and 12.

In the alternative procedures of dimension identification, the panel of judges allocated nine behavioral

statements to the Communication dimension. Communication was allocated six behavioral statements by the student group. The standard deviation values ranged from .82 to 2.40. Following the application of te standard deviation criterion, four behavioral statements remained for the Chi-square analysis. Additional information on the Communication dimension is found in Tables 10 and 13.

Table 10

Allocation of Behavioral Statements and Value Assignment by Student Samples

Dimension	Statements Allocated	Statements Surviving SD	Statements Surviving X^2
T. Sensitivity	20	18	17
Instruction	15	13	11
Discipline	10	6	6
Managerial	8	7	7
Grading	8	8	7
*Communication	6	4	2
*Enthusiasm	6	2	2
Totals	73	58	52

*Dimensions eliminated as a result of having less than four behavioral statements.

The panel of judges and the student group allocated 17 and 8 behavioral statements, respectively, to the Managerial

dimension. The standard deviation values ranged from 1.32 to 2.09. The Managerial dimension lost one behavioral statement resulting from application of the standard deviation criterion. A total of seven statements remained in the Managerial dimension.

The Grading dimension received eight behavioral statements from the judging panel and the student group. The standard deviation values ranged from 1.06 to 1.86. Consequently, all eight allocated behavioral statements received additional analysis.

The panel of judges, in the initital procedures of dimension identification, allocated 12 behavioral statements to the Discipline dimension. Nine behavioral statements were allocated by the student group. Standard deviation values ranged from 1.32 to 2.14. Four behavioral statements were eliminated, leaving five in the Discipline dimension.

In the Instruction dimension the panel of judges and student group allocated 14 and 15 behavioral statements, respectively. Standard deviation values ranged from 1.12 to 2.12. Two behavioral statements were eliminated, leaving 13 behavioral statements for additional analysis.

In the Teacher sensitivity dimension, the panel of judges allocated 31 behavioral statements while the student group allocated 20 behavioral statements. Standard deviation values ranged from 1.07 to 2.45. Two behavioral statements were eliminated, leaving 18 statements for further analysis.

Several researchers (DeCotiis, 1975; Knoppers, 1978; and Landy and Guion, 1970) suggested that the lack of anchors depicting the midrange of performance effectiveness is a major problem with BARS. The results found in BARSSPET development for this concern are presented here.

Examination of the mean scores did not include those behavioral statements which failed to survive the standard deviation criterion. So, the investigator examined mean scores from the value assignments scores which ranged from 3.50 to 4.49. The Communication dimension contained one midrange behavioral statement with a mean score equal to 3.62. The Managerial dimension contained one such behavioral statement ($\bar{X} = 4.16$). Discipline contained two behavioral statements which represented midrange levels of performance (3.65 and 4.75). The Instruction dimension contained one behavioral statement with a mean score of 4.20. The Teacher sensitivity dimension contained one behavior statement with mean rating of 3.50. Thus, the data indicate that there may be adequate representation of anchors which depict the midrange levels of performance in the BARSSPET.

Similar to the rating tendencies described in the first phase of the value assignments, there was a lack of behaviors representative of unsatisfactory levels of performance. This phenomena has not been reported in the BARS literature. In the initial procedure of the value assignments, there were values assigned which depicted unsatisfactory levels of

performance. However, lack of agreement, as evidenced by excessive standard deviation values, eliminated those behavioral statements which were found in earlier iterations. Data from allocation of values showed that there were no behavioral statements with mean values in the 6.00 to 7.00 range (ineffective). Tables 11, 12, 13, 14, 15, and 16 illustrate the summary statistics for behaviors allocated to each dimension.

The Chi-square analysis and the Fisher's Exact Probability Test were used to determine the agreement between the two major racial groups (Siegel, 1956). In the first phase of value assignments, the White population was compared to the Black population. In Cumberland County, the White and Black racial groups represent the largest segments of the subject groups. However, in Robeson County, where value assignments were made, the American Indian and Black races were the largest racial groups. Consequently, the Chi-square analysis was used to examine the expected frequencies of the Indian and Black racial groups.

With the Chi-square analysis, significance of the null hypothesis was ascertained at the .05 level. There were 58 behavioral statements retained after the application of the standard deviation criterion. Fifty-four of these behaviors were subjected to the Chi-Square analysis. There was only one behavioral statement with a Chi-square value which exceeded the critical value.

In six cases, Fisher's Exact Probability Test was used instead of the Chi-square analysis. The test of exact probability was used when the two independent samples were small (Siegel, 1956). For the analysis of the six cases, a one tailed test was used. In order for each of the five behavioral statements to be retained, the probability that there is no differences between the groups must be equal or greater than p = .9500. Five of the behavioral statements were eliminated, based upon the criteria identified above.

Following the application of Chi-square analysis and Fisher's Exact Probability Test, six behavioral statements were eliminated. One statement was eliminated from each of the Teacher sensitivity and Grading dimensions. Two behavioral statements were eliminated from each of the Instruction and the Communication dimensions.

Summary. Following the application of the standard deviation criteria and the Chi-square criteria, two dimensions retained less than four behavioral statements each. A minimum of four behavioral statements were necessary for a dimension to be retained (Knoppers, 1978). Prior to the final selection of behavioral statements and dimensions five of the seven dimensions and a total of 52 items remained (See Table 10). Tables 11, 12, 13, 14, 15, 16, and 17 illustrate the mean ratings, standard deviation values and Chi-square values of behavioral statements allocated to dimensions.

Table 11

Grading Dimension
Statistics of Criteria
for Retention of Behavioral Statements

Statement	$\bar{X}$	SD	X^2	Df	%
15	3.52	1.63	1.4000	2	60
57	3.18	1.86	5.8803	2	50
68	3.93	1.61	.0273	2	60
69	5.17	1.66	2.9513	2	50
75	3.17	1.55	3.2798	2	60
100*	2.00	1.06	.7058	--	60
108	5.94	1.76	.3025	2	60
109	2.27	1.80	.3025	2	60

Note * Denotes Fishers Exact Probability Test.
X^2 significance was determined at $p < .05$.
Dotted lines indicate statements exceeding SD criterion.

Table 12

Enthusiasm Dimension
Statistics of Criteria
for Retention of Behavioral Statements

Statement	$\bar{X}$	SD	X^2	Df	%
3	3.35	1.81	.5996	2	60
60	3.06	2.14	------	--	60
64	2.50	1.59	4.8915	2	60
105	4.00	2.25	------	--	50
111	2.77	2.14	------	--	60
129	3.11	2.27	------	--	50

Note X^2 significance was determined at $p < .05$.
Dotted lines indicates statements exceeding SD criterion.

Table 13

Communication Dimension
Statistics of Criteria
for Retention of Behavioral Statements

Statement	$\bar{X}$	SD	X^2	Df	%
22*	1.55	.82	.9000	--	50
43	3.10	2.40	------	--	50
45	3.62	1.81	2.3794	2	50
85*	2.00	1.45	.4859	--	60
89	1.70	.98	.4700	2	60
120	2.61	2.25	------	--	50

Note * Denotes Fishers Exact Probability Test.
X^2 significance was determined at $p < .05$.
Dotted lines indicates statements exceeding SD criterion.

Table 14

Managerial Dimension
Statistics of Criteria
for Retention of Behavioral Statements

Statement	$\bar{X}$	SD	X^2	Df	%
7	2.85	1.69	1.4814	2	60
21	2.80	1.32	.8547	2	50
39	5.12	1.36	1.8871	2	50
55	2.31	1.70	1.2307	2	50
71	3.05	1.95	2.0683	2	60
88	5.00	1.54	.4700	2	60
96	3.05	1.78	1.2655	2	50
130	4.16	2.09	------	--	50

Note X^2 significance was determined at $p < .05$.
Dotted lines indicates statements exceeding SD criterion.

Table 15

Discipline Dimension
Statistics of Criteria
for Retention of Behavioral Statements

Statement	$\bar{X}$	SD	X^2	Df	%
26	2.20	1.47	2.2222	2	60
43	5.75	1.39	5.6084	2	60
44	3.65	2.08	------	--	60
48	4.75	2.01	------	--	60
50	5.87	1.70	1.2307	2	60
54	2.93	2.14	------	--	50
73	5.52	1.32	2.9513	2	60
82	5.70	2.05	------	--	60
102	3.47	1.73	.5163	2	50

Note X^2 significance was determined at $p < .05$.
Dotted lines indicates statements exceeding SD criterion.

Table 16

Instruction Dimension
Statistics of Criteria
for Retention of Behavioral Statements

Statement	$\bar{X}$	SD	X^2	Df	%
1	4.20	1.88	2.1428	2	60
8	2.70	1.12	.9523	2	60
9	3.00	1.86	1.4102	2	60
17	2.35	1.22	1.2963	2	60
26	2.20	1.47	2.2222	2	60
29	3.10	1.20	1.4818	2	60
44	2.75	1.87	1.9697	2	50
52	1.93	1.52	.5274	2	60
67*	2.12	1.66	.3500	--	60
78	3.14	1.22	.0269	2	60
92	1.82	1.28	.8297	2	60
97	2.29	1.31	3.6615	2	60
107	5.05	2.12	------	--	60
110	2.44	1.78	8.4705	2	50
112	3.94	2.12	------	--	60

Note * Denotes Fishers Exact Probability Test.
X^2 significance was determined at $p < .05$.
Dotted lines indicates statements exceeding SD criterion.

Table 17

Teacher Sensitivity Dimension
Statistics of Criteria
for Retention of Behavioral Statements

Statement	$\bar{X}$	SD	X^2	Df	%
2	2.70	1.49	.5555	2	50
5	2.50	1.84	2.1759	2	60
6	2.95	1.95	4.1269	2	60
20	3.00	1.91	2.4074	2	60
24	3.50	1.53	2.1428	2	60
30	3.40	1.50	1.4814	2	60
32	2.52	1.07	.4191	2	60
35	2.68	1.66	.5743	2	60
36	3.50	2.19	------	--	60
59	2.31	1.40	4.4472	2	60
65	1.87	1.20	4.7472	2	60
70	2.58	1.62	1.8931	2	60
72*	1.88	1.31	.2941	--	50
86*	1.82	1.13	.4852	--	60
93	4.76	1.75	3.6739	2	60
94	2.00	1.32	5.6464	2	60
103	3.05	1.69	1.0588	2	50
104	2.66	1.84	3.7058	2	60
119	2.50	2.45	------	--	60
120	5.00	1.94	.4072	2	60

Note * Denotes Fishers Exact Probability Test.
X^2 significance was determined at $p < .05$.
Dotted lines indicates statements exceeding SD criterion.

Construction of the BARSSPET

Final selection of items and dimensions. The final selection of anchors to be used in the BARSSPET was based upon (a) selection of behavior statements representing degrees of effectiveness ranging from extremely effective to neutral to extremely ineffective (b) the rejection of behavioral statements which have similar mean values with greater standard deviation values and (c) rejection of behavioral statements which contain any duplicate behaviors.

No further dimensions were eliminated in the final selection of behavioral statements. One behavioral statement, having a standard deviation value of 2.12 (X=5.00) was added to the Instruction dimension. The addition of the aforementioned behavioral statement was necessary to fully represent the ineffective range. In the Teacher sensitivity dimension, there were two behavioral statements with the same behavior in a similar context. Teacher allowance of time for cooling down (X=3.00) was the context of the first statement. The second statement (X=2.50) was concerned with allowance of time for resting. The second statement was eliminated. Finally, twenty-eight behavioral statements were selected as anchors in a total of five subscales (dimensions). These scales are found in Appendix I.

Statistical Analysis

A new group of students in the Cumberland County School system now responded to the BARSSPET. Student data are presented in Table 18. The rating data were analyzed using the Digital VAXX 11/780 system at the University of North Carolina at Greensboro. The SPSSX (1983) statistical package was used in the data analysis.

Table 18

BARSSPET Respondents
Statistical Evaluation Phase

Race	N	%	Sex	N	%	Class	N	%
				School 1				
Indian	2	1.7	Female	47	40.9	Sophomore	96	83.5
Black	34	29.6	Male	66	57.4	Junior	12	10.4
White	64	55.6	M.V.	2	1.7	Senior	7	6.1
Hispanic	11	9.6	----	--	----	------	--	----
Asian	4	3.5	----	--	----	------	--	----
				School 2				
Indian	1	1.7	Female	34	56.7	Sophomore	49	80.4
Black	27	45.0	Male	26	43.3	Junior	10	16.4
White	29	48.3	----	--	----	Senior	1	1.6
Hispanic	1	1.7	----	--	----	M.V.	1	1.6
Asian	3	3.3	----	--	----	------	--	----

Note M.V. = Missing value

Mean and standard deviation scores were generated from student responses. Mean and standard deviation scores by dimensions were computed for overall ratings and teachers. Pearson Product Moment correlations were generated to determine interdimension correlations and test-retest reliability. Means, standard deviations, and ranges are presented in Tables 19 and 20. Interdimension correlations are presented in Table 21.

Psychometric Analysis

Leniency bias. Bernardin (1977) stated that leniency bias is defined as a shift in the mean ratings in the positive direction. DeCotiis (1977) suggested that leniency bias is characterized by extremely high or low ratings with little variability.

The indicators of leniency were discussed by comparing data from the raters at School One and School Two for each dimension. Students at School One rated Teachers One through Four while students at School Two rated Teachers Five through Seven. Individual teacher data is presented in Table 19. Also, the overall rating data (See Table 20) was used to draw conclusions regarding leniency in the BARSSPET.

The mean scores for School One data on Teacher sensitivity ranged from 3.67 to 4.67. The standard deviation scores ranged from 1.47 to 1.86. All of the scale points were used by the raters of School One. The teachers rated by

School Two raters received lower mean scores ($\bar{X}$ range from 1.89 to 4.00). In each case, the standard deviation scores were lower, ranging from .79 to 1.36. Students in School Two tended to avoid the use of scale points representing ineffective teaching performance.

Table 19

BARSSPET Ratings of Teachers

	Teacher 1			Teacher 2		
Dim	$\bar{X}$	SD	Range	$\bar{X}$	SD	Range
Teas	3.67	1.73	1 - 7	3.90	1.47	1 - 7
Mana	3.15	1.79	1 - 7	3.75	1.46	1 - 7
Grad	3.44	1.87	1 - 7	3.41	1.60	1 - 7
Disc	3.93	1.10	1 - 6	3.78	1.01	2 - 7
Inst	1.70	0.78	1 - 3	1.97	1.03	1 - 7
	Teacher 3			Teacher 4		
Dim	$\bar{X}$	SD	Range	$\bar{X}$	SD	Range
Teas	4.67	1.77	1 - 7	4.16	1.86	1 - 7
Mana	3.53	1.33	1 - 6	3.36	1.47	1 - 6
Grad	4.70	1.58	2 - 7	4.52	1.36	2 - 7
Disc	4.20	1.10	3 - 7	3.56	0.77	2 - 6
Inst	4.00	1.41	1 - 6	3.04	1.27	2 - 7

Table 19 Continued

	Teacher 5			Teacher 6		
Dim	$\bar{X}$	SD	Range	$\bar{X}$	SD	Range
Teas	2.54	.79	2 - 4	4.00	1.36	2 - 6
Mana	2.93	1.18	1 - 5	3.09	1.04	2 - 5
Grad	2.28	.94	1 - 5	1.87	.55	1 - 3
Disc	2.89	1.00	1 - 4	3.13	.87	2 - 6
Inst	1.85	.52	1 - 3	1.87	.63	1 - 4

	Teacher 7		
Dim	$\bar{X}$	SD	Range
Teas	1.89	.93	1 - 4
Mana	2.56	1.06	2 - 3
Grad	1.89	.33	1 - 2
Disc	3.22	1.71	1 - 7
Inst	2.00	.50	2 - 3

The mean scores assigned by School One for the Managerial dimension ranged from 3.15 to 3.75. The standard deviation scores ranged from 1.33 to 1.79. In School Two, the mean scores ranged from 2.56 to 3.09 while the standard deviation scores ranged from 1.04 to 1.18. The raters of School Two, when compared to the raters of School One were were very restrictive in their use of all scale points representive of ineffective performance.

In the Grading dimension, teachers of School One received mean and standard deviation scores ranging from 3.44 to 4.70 and 1.36 to 1.80, respectively. In cases where all of the scale points were not used, the effective range was avoided. In School Two, the mean and standard deviation scores ranged from 1.87 to 1.89 and .33 to .55, respectively. In contrast to the rating tendencies of School One raters, School Two avoided the use of scale points representing neutral and ineffective teaching performances.

The mean scores in the Discipline dimension for the teachers in School One ranged from 3.56 to 4.20. The standard deviation scores ranged from .77 to 1.10. Students at School One who did not use all of the scale points tended to avoid the use of scale points representing effective performance. Following the trend established in discussion of the first three dimensions, the mean scores of Group Two teachers were more effective. The mean scores ranged from 2.89 to 3.22. In contrast, the magnitude of the standard deviation scores were higher. The standard deviation scores ranged from .87 to 1.71. Again, the raters at School Two avoided the use of scale points representing ineffective teaching performance.

In both School One and School Two, the mean cores for the Instruction dimension were judged more effective than mean scores for the other four dimensions. The range of mean scores were from 1.70 to 4.00 and 1.85 to 2.00, respectively.

The standard deviation scores ranged from .78 to 1.41 and from .50 to .63, respectively. In the use of scale points, the raters from both schools avoided the use of scale points representing ineffective performance. Knoppers obtained mean and standard deviation values ranging from 5.14 to 5.71 and 1.29 to 1.66, respectively. Similarly, Bernardin (1977) obtained values ranging from 4.13 to 5.27 and .91 to 1.34.

Table 20

Overall Ratings
of Teaching Behaviors

	School 1			School 2		
Dim	$\bar{X}$	SD	Range	$\bar{X}$	SD	Range
Teas	4.12	1.72	1 - 7	2.98	1.32	1 - 6
Mana	3.47	1.51	1 - 7	2.93	1.06	1 - 5
Grad	4.01	1.70	2 - 7	2.07	.76	1 - 5
Disc	3.89	1.02	1 - 7	3.03	1.07	1 - 7
Inst	2.67	1.46	1 - 7	1.88	.56	1 - 4

Examination of the overall ratings for teachers in Group One (see Table 20) showed three of the five scales with mean ratings below the midpoint value (4.00). The range of standard deviation values and the use of scale points indicate substantial variance in the ratings. The mean values and the indicators of variance suggest that leniency bias was minimal. Leniency bias was greatest in the

Instruction dimension. The raters at School Two, for each dimension, illustrated shifts in all three indicators of leniency. The positive shift in mean scores, representing almost one scale point, indicate leniency as defined by Bernardin (1977). Mean scores and standard deviation values confirm the presence of greater leniency bias in the ratings of School Two.

Halo bias. DeCotiis (1977) defined halo bias as the unwillingness of a rater to discriminate between dimensions of job performance. The magnitude of intercorrelations between the dimensions provide indications of halo bias (Borman and Motowildo, 1977). Table 21 illustrates the range among interdimension correlation coefficients and the median coefficient for each dimension. The interdimension correlations were obtained from the correlation matrix which illustrated the relationship between each dimension and all other dimensions. The median correlation for the ratings produced by the raters of each school was the mean of the sum of the five median interdimension correlations. The raters at School One produced a median correlation coefficient of .33. The median correlation coeffecient obtained from School Two was .07.

Knoppers (1978) obtained a coefficient of .50 for her instrument. Other studies produced overall median coefficients ranging from .32 (Keaveny and McGann, 1975) to .73 (DeCotiis, 1977). Judgment of halo bias in BARSSPET (for

School One) ratings was termed low. Halo bias for School Two was judged to be very low (Morehouse and Stull, 1975).

Table 21

Medians and Ranges of Interdimension Correlations of BARSSPET Ratings

Dimensions	Median Correlation	Range
	School 1	
Teacher Sensitivity	.35	.27 - .48
Managerial	.34	.24 - .45
Grading	.40	.36 - .43
Discipline	.28	.13 - .36
Instruction	.26	.13 - .43
	School 2	
Teacher Sensitivity	.05	-.17 - .26
Managerial	.07	-.17 - .36
Grading	-.17	-.18 - .02
Discipline	.04	.02 - .26
Instruction	.10	-.14 - .36

Note The median correlation for each dimension was the mean of the two middle correlations.

Although the median correlations and the range of interdimensional correlations were very low for the raters at each school, one difference between the groups was noted. In School Two, four of the BARSSPET dimensions reflected

negative correlations in the ranges. The negative correlations indicated an inverse relationship between some dimensions. That is, in some cases the raters at School Two would rate the teacher as effective on one dimension and then assign a negative rating on the next dimension.

Central tendency bias. Isaac and Michael (1980) defined central tendency bias as ratings of subjects toward the middle of the scale. DeCotiis (1977) stated that standard deviation scores provide indications of central tendency bias. Means and standard deviation scores are found in Table 20.

The overall mean scores for School One ranged from 2.67 (Instruction) to 4.12 (Teacher Sensitivity). With the exception of the mean score for Instruction, all other scores were within .53 of the scale midpoint of 4.0. The standard deviation scores ranged from 1.02 (Discipline) to 1.72 (Teacher Sensitivity). The magnitude of the standard deviation scores indicated considerable variability around the mean scores.

The overall means scores for School Two ranged from 1.88 (Instruction) to 3.03 (Discipline). The mean score (Discipline) nearest the scale midpoint was 3.03. Standard deviation values ranged from .55 (Instruction) to 1.32 (Teacher Sensitivity). The responses of School Two did not vary to the degree of School One.

In School One, the ratings of teachers produced 12 mean scores within .50 of the scale midpoint. In contrast, the

ratings of the teachers from School Two produced only one mean score within .50 of the scale midpoint. In School Two, the Grading and Instruction dimensions appeared less affected by central tendency than the remaining dimensions. In the School Two data, no dimensions appeared to be affected by central tendency.

Isaac and Michael (1980) stated that central tendency bias is likely to occur when raters are unfamiliar with or uncertain about what is to be rated. Upon completion of the directions for responding to the BARSSPET, some students at School One were still confused about their task. Although additional time was provided for explanations of directions, the investigator could not be sure whether student questions were satisfactorily answered. Indications of central tendency bias was low for the BARSSPET ratings (Group One). It is possible that some central tendency bias was due to unresolved questions. In the ratings obtained from Group Two, central tendency bias was judged to be negligible.

Scale reliability. Test-retest procedures were used to assess scale reliability for the BARSSPET. Data collected from students at School Two was used for this procedure. Sixty-one students constituted the sample for the first administration. The BARSSPET was administered a second time after a ten day interval. Fifty students responded to the BARSSPET on the second administration. The absence of 11 students might be attributed to an avoidance factor.

According to the the teacher, the absent students were poor readers.

Correlation coefficients ranged from .21 to .56 for the BARSSPET. Fogli, Hulin, and Blood (1971) constructed a BARS to appraise the performance of grocery store checkers. The lowest reported coefficient for the nine scales was .97. In a BARS developed to ascertain performance levels of nurses, Smith and Kendall (1963) obtained scale reliabilities ranging above .97. Knoppers (1978) developed a BARS to appraise coaching performance. Correlational coefficients ranged from .83 to 1.00. These comparisons should be viewed in the context of the different types of investigations including the age of subjects where stability of opinions over time may be an influence.

Table 22

BARSSPET Scale Reliability

Dimension	Scale r
Teacher Sensitivity	.47
Managerial	.56
Grading	.54
Discipline	.43
Instruction	.21

Intervening events may be a contributing factor for the low correlation coefficients. First, the second administration of the BARSSPET was postponed due to an unanticipated school-wide event. Circumstances produced a time interval which included two weekends. Secondly, it is possible that students were preoccupied with their yearbooks. Yearbooks were passed out on the original date specified for the second administration of the BARSSPET. Another plausable explanation for the low test-retest might possibly be in changes in teaching behaviors that occur at the end of the school year. A final explanation may be that in spite of careful, iterative procedures, student maturity levels may interfere with temporal stability in this age group.

Discriminant validity. A type of discriminant validity may be ascertained by correlating ratings of student satisfaction to BARSSPET ratings. Evidence of BARSSET independence may be obtained when satisfaction scores do not correlate with BARSSPET ratings. For example, if the scales are independent, one would not find all highly satisfied students giving effective ratings or dissatisfied students giving ineffective ratings. Kerlinger (1973) stated:

> Discriminability means that one can empirically differentiate the construct from other constructs that may be similar, and that one can point out what is unrelated to the construct. We point out, in other words, what other variables are correlated with the construct and how they are so correlated (p.462).

The satisfaction scale by Kneer (1972), Your Feelings About This Class, contains six subscales. The six subscales are entitled: (a) My self in this class, (b) my fun in this class, (c) my learning in this class, (d) my classmates, (e) my new experiences in this class, and (f) my freedom in this class. In each subscale, students responded to a seven point continuum anchored with bi-polar adjectives. There were ten sets of bi-polar adjectives for each subscale. Subscale scores were obtained by summing the 10 responses on each subscale (M.E. Kneer, personal communication, April 17, 1985). In order to compare more easily the subscale scores to the BARSSPET scores, mean scores were computed for each subscale. Similarly, a total score was computed for the entire scale.

After responding to the BARSSPET, students at School One responded to the satisfaction scales. Student data are found in Table 18. Descriptive statistics for the satisfaction subscales are found in Table 23. Mean scores ranged from 4.42 to 5.18. Standard deviation scores ranged from .87 to 1.13. The mean scores and the standard deviation scores indicated that central tendency bias was present in the satisfaction scores.

The Pearson Product Moment correlation technique was used to ascertain the relationship between each of the BARSSPET scales and the six satisfaction subscales. A total of 30 correlation coefficients were produced. The

correlation coefficients are found in Table 24. Twenty-six of the 30 coefficients were significant at $p < .01$, meaning only that these correlations were unlikely to have been found by random chance alone.

The correlation coefficients between the satisfaction subscales and the BARSSPET dimensions ranged from .058 (Instruction dimension and Self subscale) to to .450 (Grading dimension and Experience subscale). The lowest correlation coefficients were associated with the Self subscale. The Experience subscale appeared to show the greatest relationship with the BARSSPET. The correlation coefficents ranged from .281 to .450.

Table 23

Satisfaction Scores
Kneer Inventory (1972)

Scale	$\bar{X}$	SD	Range
Self	5.17	.87	3 - 7
Fun	4.80	1.10	2 - 7
Learning	4.85	1.09	2 - 7
Classmates	4.42	1.13	1 - 7
New Experiences	4.65	1.12	1 - 7
Freedom	4.62	1.05	1 - 7

Nearly all of the 30 correlation coefficients between the Kneer Satisfaction subscales and the BARSSPET dimensions

indicate a low but slight relationship. These data suggest that the BARSSPET ratings were not highly related to student satisfaction. Thus, a lack of association with the student satisfaction constucts might be claimed.

Table 24

BARSSPET Independence
from Student Satisfaction

	Satisfaction Subscales						Totals
Dim	Self	Fun	Learn	C. Mates	Exper	Freedom	
Teas	.314	.364	.442	.261	.348	.298	.413
Mana	.154*	.305	.383	.160	.281	.333	.333
Grad	.223*	.426	.354	.317	.450	.302	.427
Disc	.134*	.396	.308	.300	.434	.362	.402
Inst	.058*	.163	.284	.211	.334	.254	.272

All coefficients are significant at $p < .01$. except for those with an asterisk

Content validity. Brown (1976) suggested that the basic question in content validity is whether evaluative items constitute a representative sample of the content domain of concern. Kerlinger (1978) stated...

> ..."competent" judges should judge the content of the items. The universe of content must, if possible, be clearly defined; that is, the judges must be furnished with specific directions for making judgements (p.459).

Also, Isaac and Michael (1980) stated that content validity is demonstrated by showing how well the content of the test

samples the subject matter about which the conclusions are made.

In this investigation, the universe set of behavioral statements was determined by the first sample of students and teachers who identified 735 examples of teaching behaviors. The first sample of students and teachers followed explicit criteria for identifying examples of teaching behaviors. The behavioral statements were judged and edited with student assistance to insure that student raters would recognize and understand the content and terminology of the statements. The criteria for retention of behaviors after rating of the statements and allocation and value assignment represented additional criteria by which content validity was established.

Since the behavioral statements surviving the selection criteria were judged to be representative of the initial domain of behavioral statements, the anchors in the BARSSPET were judged to possess content validity.

Construct validity. Brown (1976) stated that construct validity is important whenever a test is designed to measure some attribute or quality that people are presumed to possess. It is assumed by the investigator that the BARSSPET measured two separate but related domains. These domains include dimensions of teaching behaviors and levels of effectiveness. Essentially, the BARSSPET respondent, for each scale, identified a behavioral anchor which was most

like his/her teacher. The behavior identified by the respondent aided in the determination of the level of effectiveness for the teacher on that scale. However, this investigation did not address BARSSPET discrimination of effectiveness.

The claims of construct validity focus upon the dimensions found in the BARSSPET. The first group of students generated behavioral statements representing specified levels of teaching effectiveness. Students were instructed to follow carefully uniform guidelines in determining the peformance level of the behaviors. A second group of students rated the examples of behaviors for accuracy in describing effective teaching behaviors. Cronbach's alpha analysis of the ratings examined the internal cohesiveness of the ratings. The objective of this analysis was to retain behavioral examples in which there were agreements as to levels of effectiveness.

From the remaining statements, the investigator formed clusters of statements representing dimensions of teaching behavior. In the next step of dimension development expert judges assigned each of the behavioral statements to one of seven dimensions. A final group of students assigned the same set of behavioral statements to the dimensions. Thus, some claim for construct validity was based upon the process of having independent groups identify and select behavioral statements and the iterative process of identifying and

defining BARSSPET dimensions from the selected behavioral statements.

Profile Analysis

To aid in analysis of Group one and Group Two data, a discussion of teacher rating profiles follows. The descriptive statistics for the two rating groups are found in Table 20. Global ratings and Z scores are found in Appendix K. The profiles will be used to examine trends rather than individual ratings.

For use in the profile analysis, the data were converted to standard Z scores. By this conversion, positive Z scores are indicative of ineffective ratings while negative Z scores indicate effective ratings. Profiles of teachers rated by students from School One are represented in Figures 1 through 4. Profiles of teachers rated by students of School Two are represented in Figures 5 through 7. Each profile illustrates BARSSPET ratings and global ratings.

BARSSPET ratings. BARSSPET ratings appeared to exhibit considerable variability between dimensions. Scores appeared above and below the scale midpoint. For each dimension, except for "Instruction," no score exceeded one standard deviation unit.

Instruction scores in five of the seven profiles were below the Z scale midpoint. In two profiles, the Instruction scores exceeded one standard deviation unit. The factor

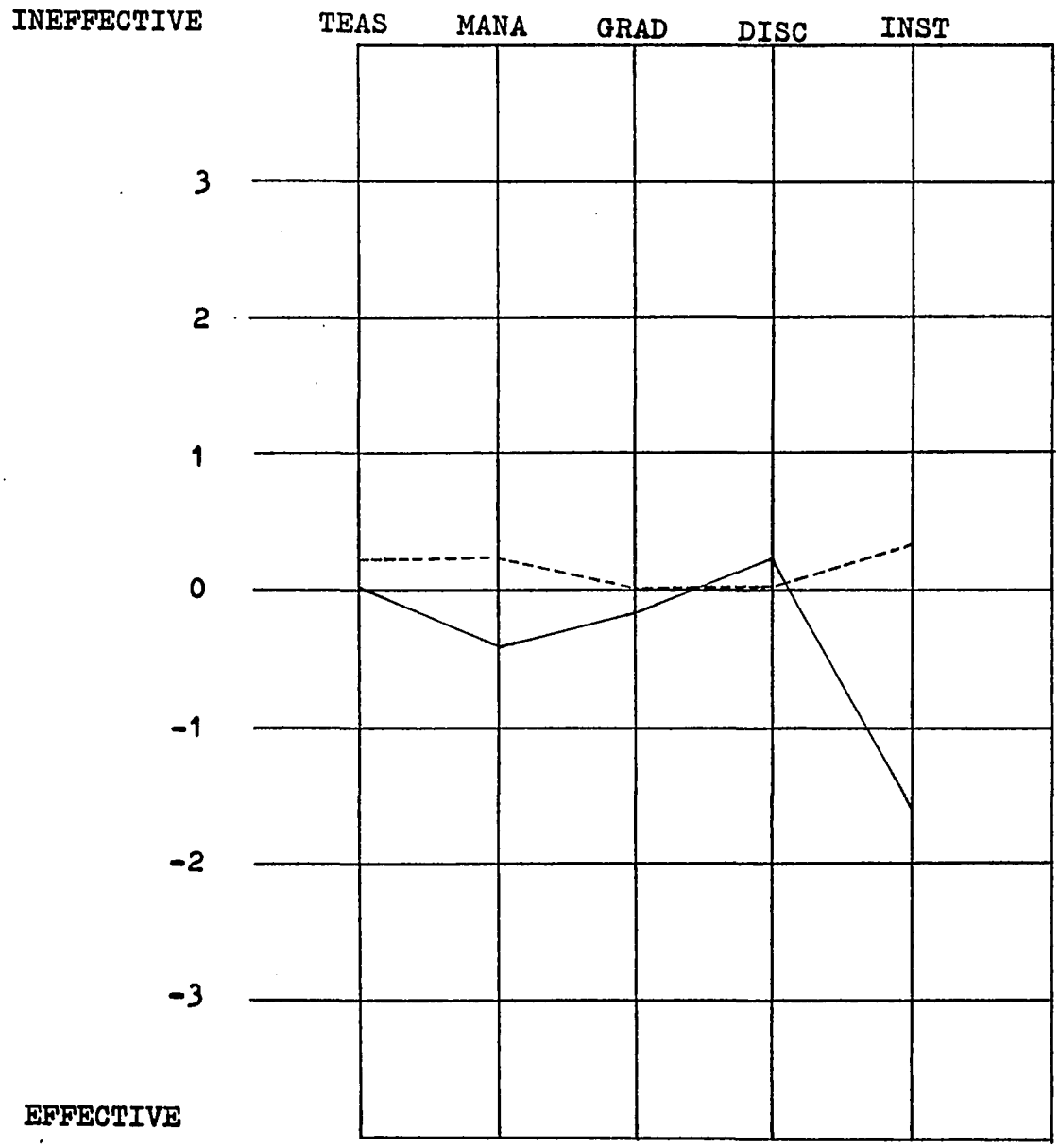
promoting negative Z scores in Instruction scores may be associated with anchors on the scale. Finley, et al. (1977) could find no differences between the effects of general and specific anchors. The first two anchors on the BARSSPET instruction scale appear to be less specific than the remaining anchors. It may be that the latitude of general anchors could predispose students to rate instructional behaviors as extremely effective.

The other trend was associated with Grading and Instruction scores from the raters of School Two. All standard scores for Grading and Instruction were below the Z scale midpoint, indicating effective ratings. A likely explanation is found in the rating procedures used for the School Two raters. During student response to the BARSSPET, at least one physical education faculty member was present. For some students, the mere presence of a teacher in the room may have represented an intimidation factor. It is possible that the presence of the teacher introduced bias into the ratings, especially those associated with Grading and Instruction.

Global ratings using Likert scales. Before each administration of the BARSSPET, students responded to a set of global scales representing the BARSSPET dimensions. The use of the global ratings was intended to reduce leniency and halo bias in the BARSSPET ratings (Lawler, 1967 and Zedeck, et al. 1976).

Two trends were noted in the profiles of the global scores. When the global ratings were compared to the BARSSPET ratings, the global scores were closer in proximity to the scale midpoint and showed less variability between dimensions. Figures One through Seven illustrate the differences in variability along the scale points and between dimensions for the global and BARSSPET ratings.

Explanations for the differing results may be found with the characteristics associated with each of the two types of rating scale. The anchors of the BARSSPET force the respondent to relate to specific behaviors. The bi-polar adjectives and the numbers represent the only guide for responding to the global scales. Consequently, the raters using global scales must rely upon general impressions of behaviors to make rating decisions. Thus, the behavioral anchors of the BARSSPET may have aided the raters in discriminating between dimensions and levels of effectiveness on each dimension. This is, of course, illustrative of the theoretical advantages of the BARS over Likert-scale ratings.

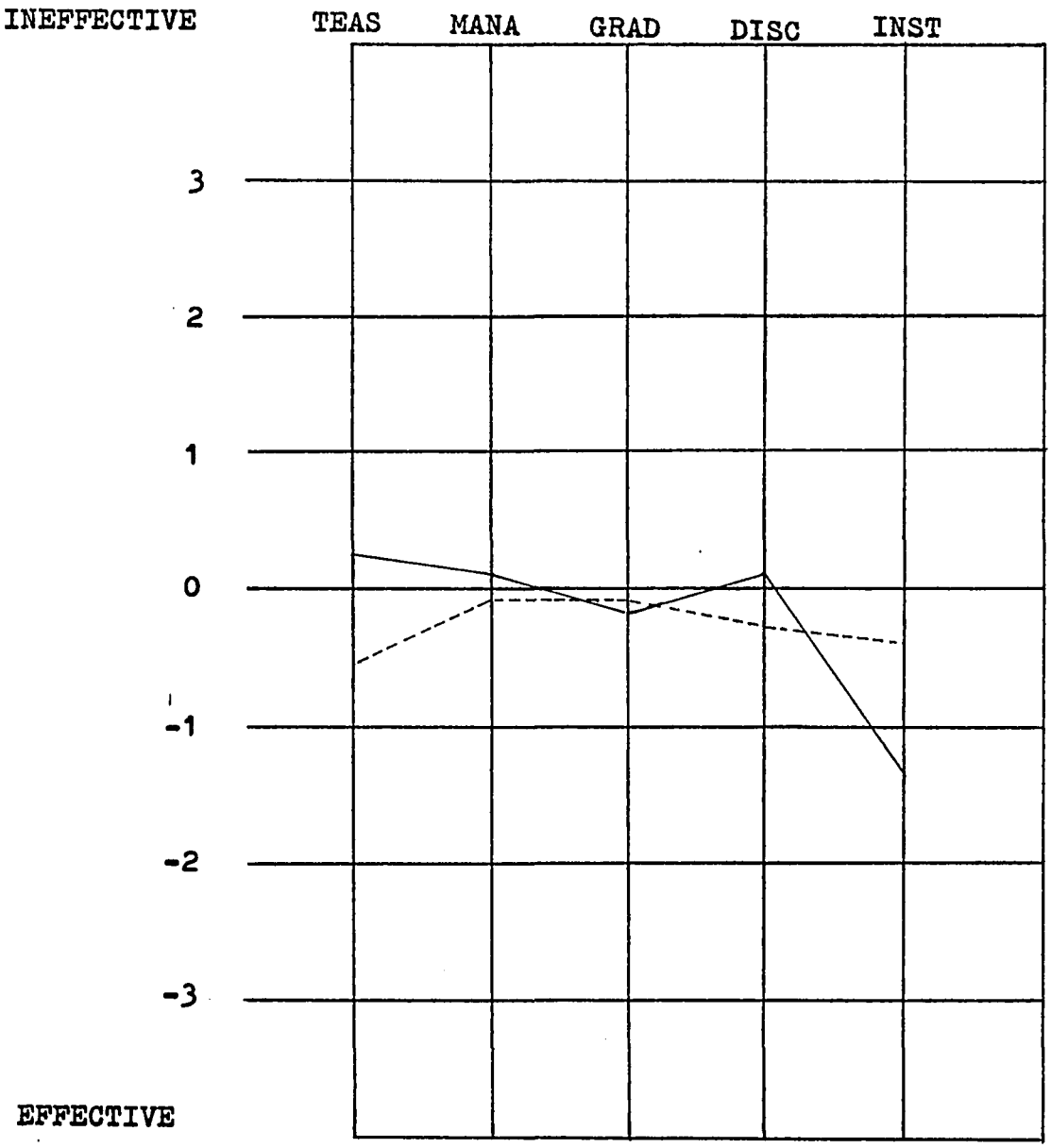


Note 1. Solid line denotes BARSSPET ratings.
2. Broken line denotes global ratings.

Figure 1

Profile Analysis of
Teacher 1 Ratings

Standardized Z Scores

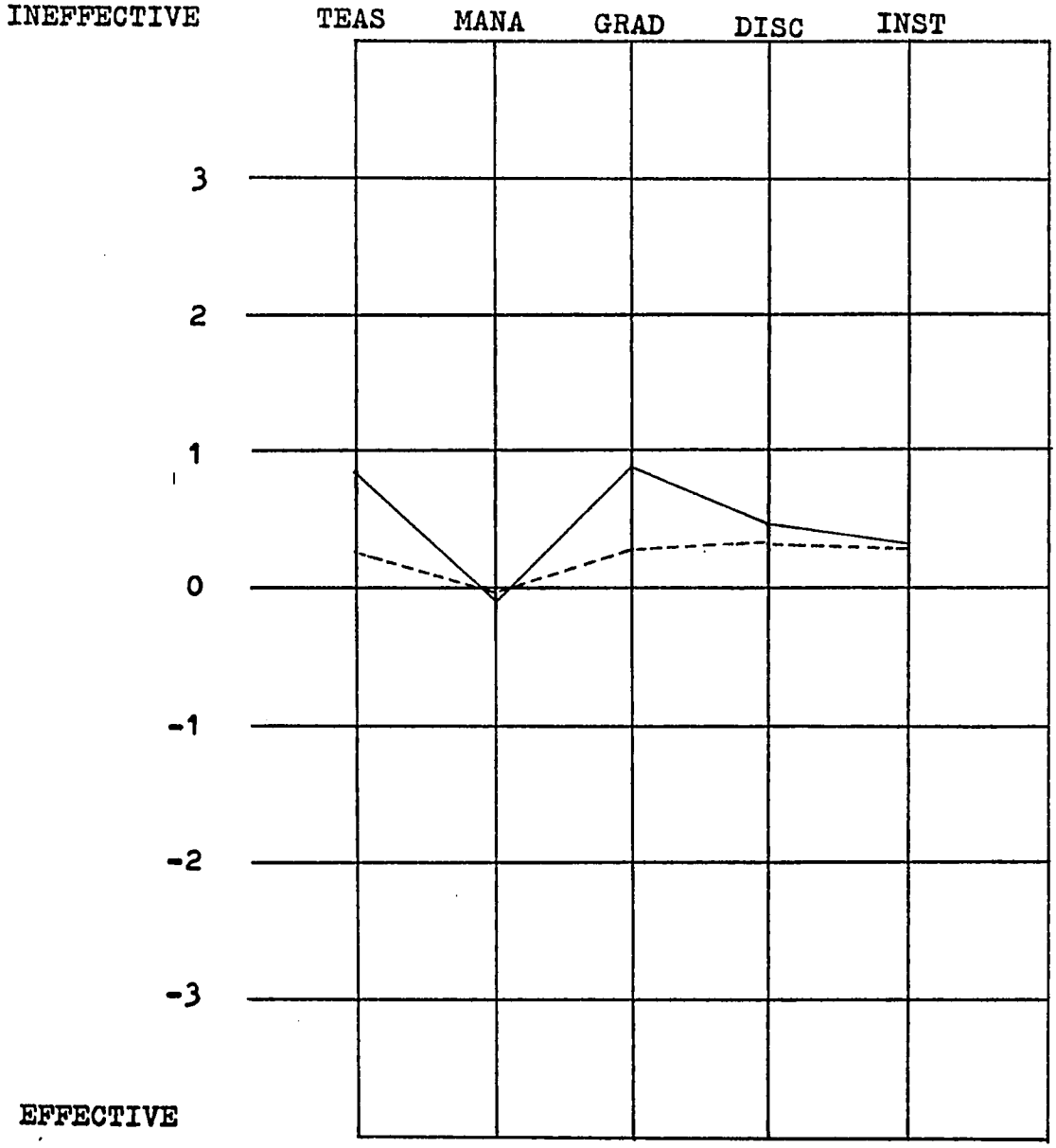


Note 1. Solid line denotes BARSSPET ratings.
2. Broken line denotes global ratings.

Figure 2

Profile Analysis of
Teacher 2 Ratings

Standardized Z Scores

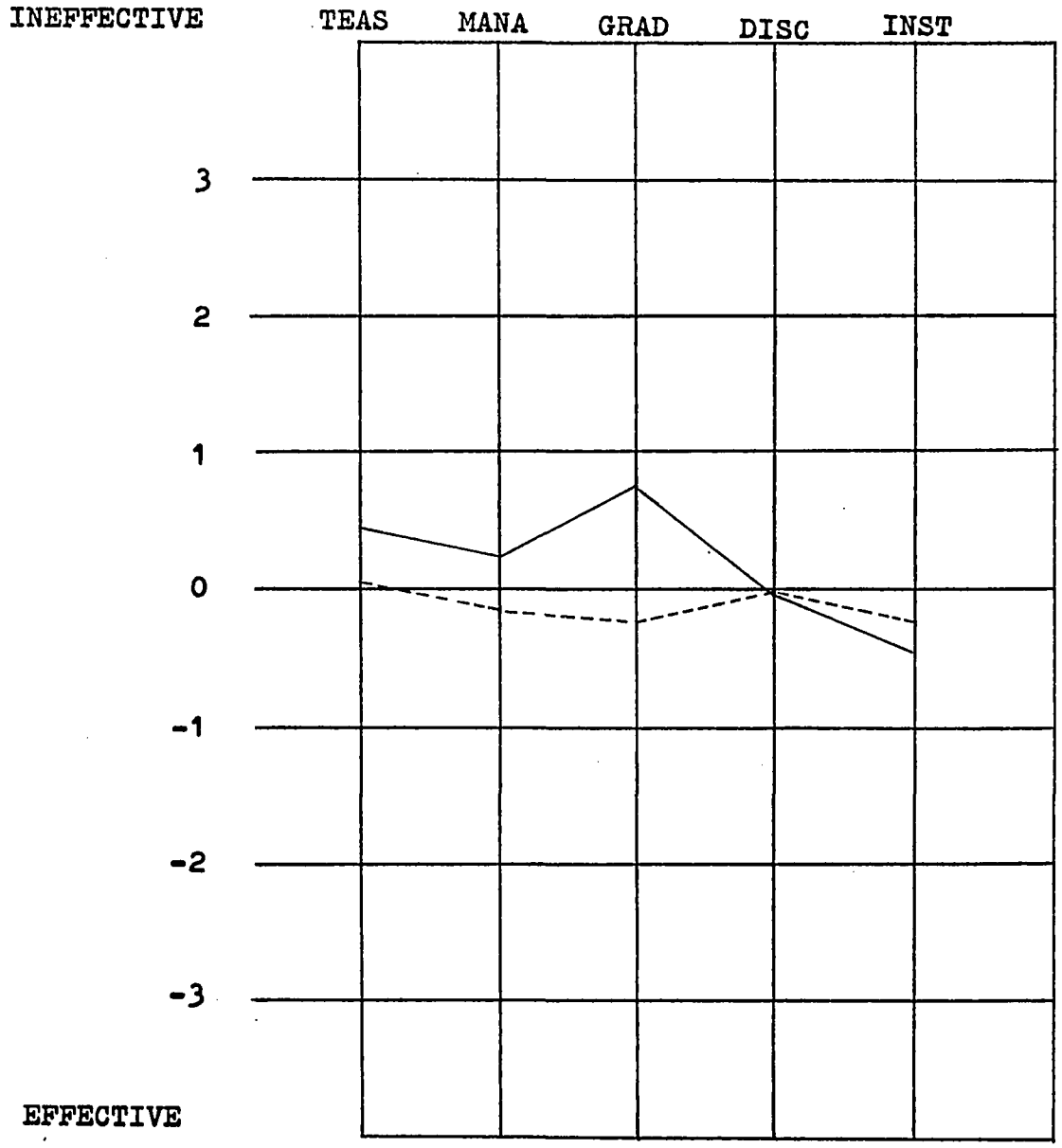


Note 1. Solid line denotes BARSSPET ratings.
2. Broken line denotes global ratings.

Figure 3

Profile Analysis of
Teacher 3 Ratings

Standardized Z Scores

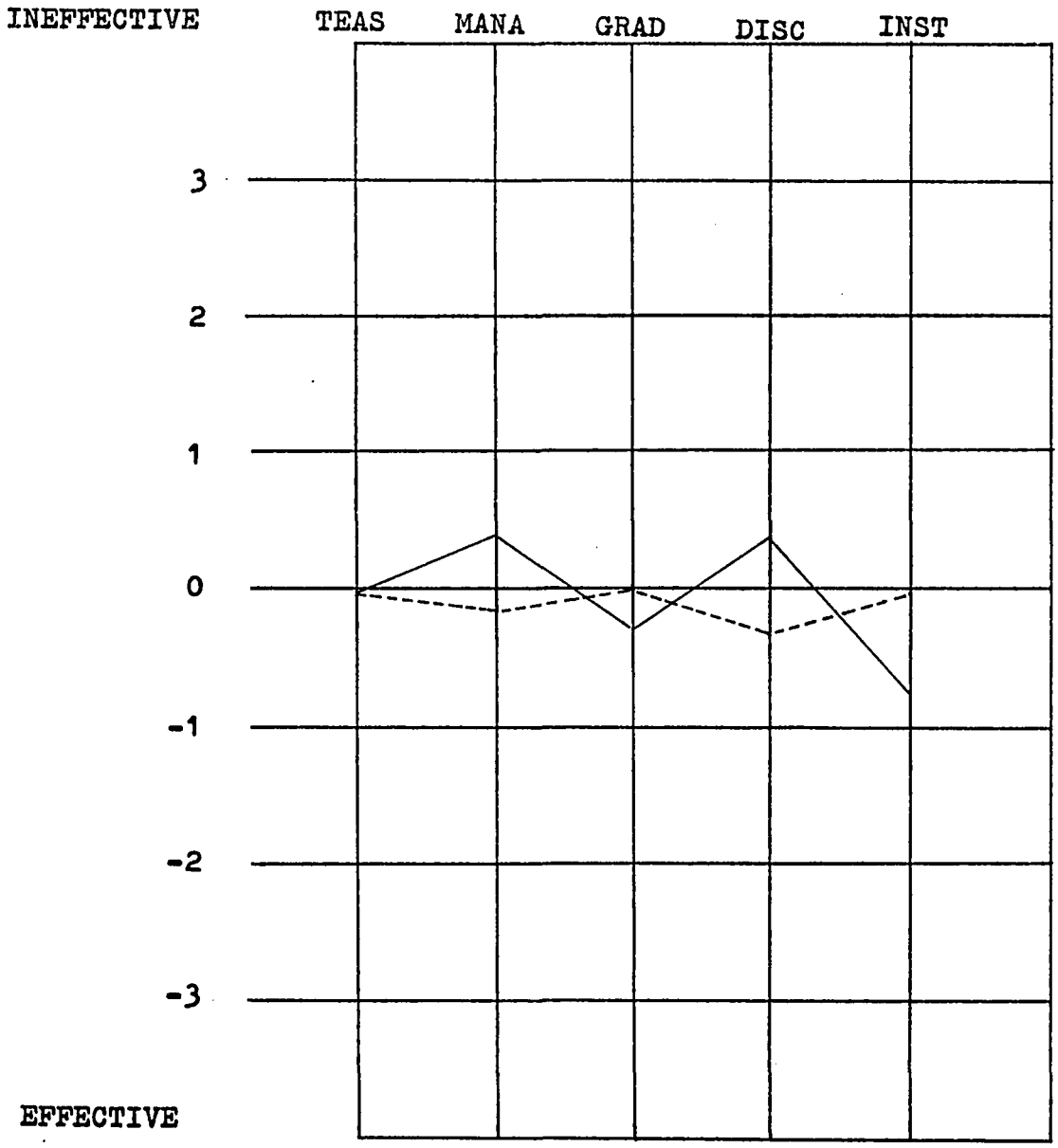


Note 1. Solid line denotes BARSSPET ratings.
2. Broken line denotes global ratings.

Figure 4

Profile Analysis of
Teacher 4 Ratings

Standardized Z Scores

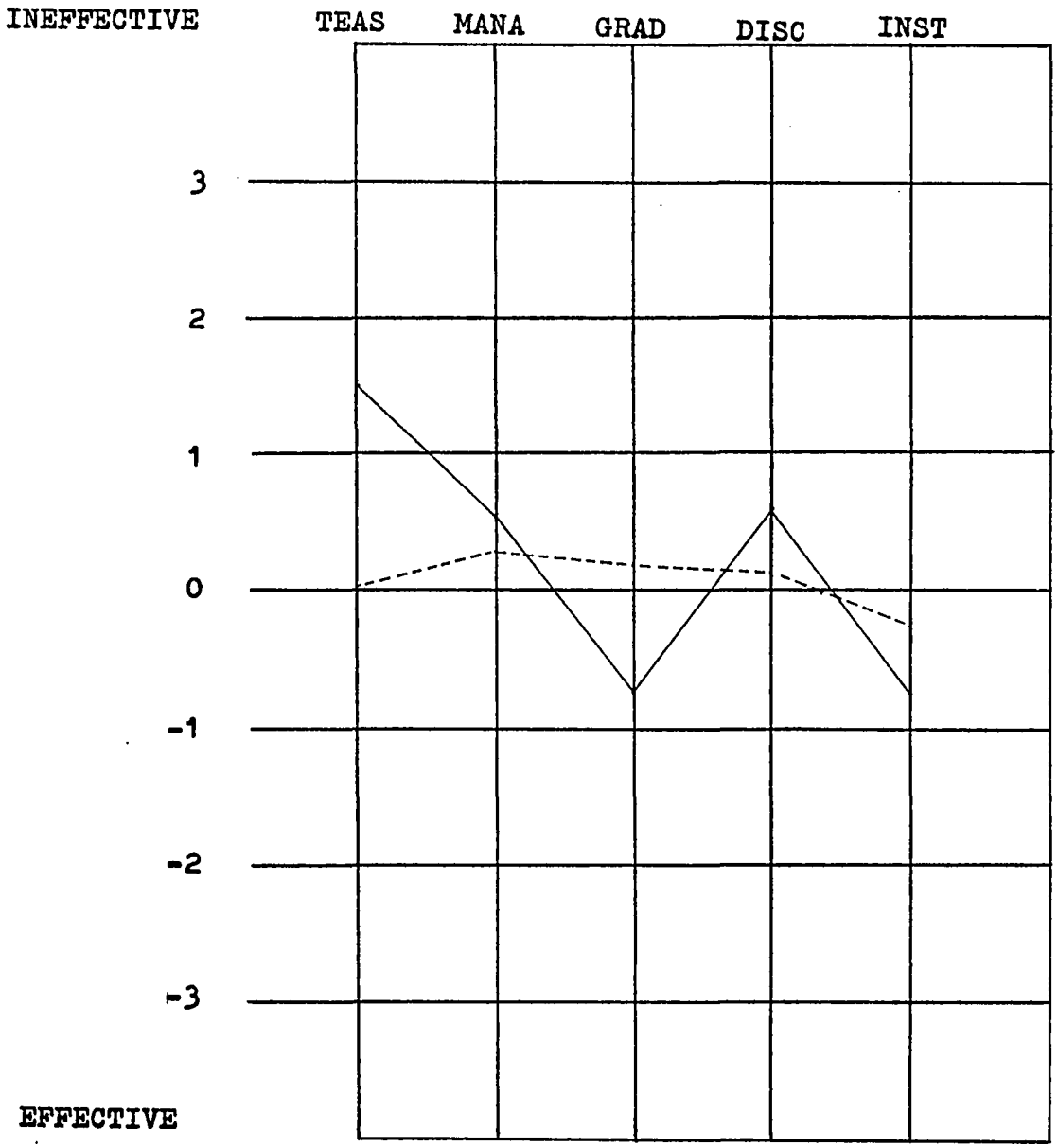


Note 1. Solid line denotes BARSSPET ratings.
2. Broken line denotes global ratings.

Figure 5

Profile Analysis of
Teacher 5 Ratings

Standardized Z Scores

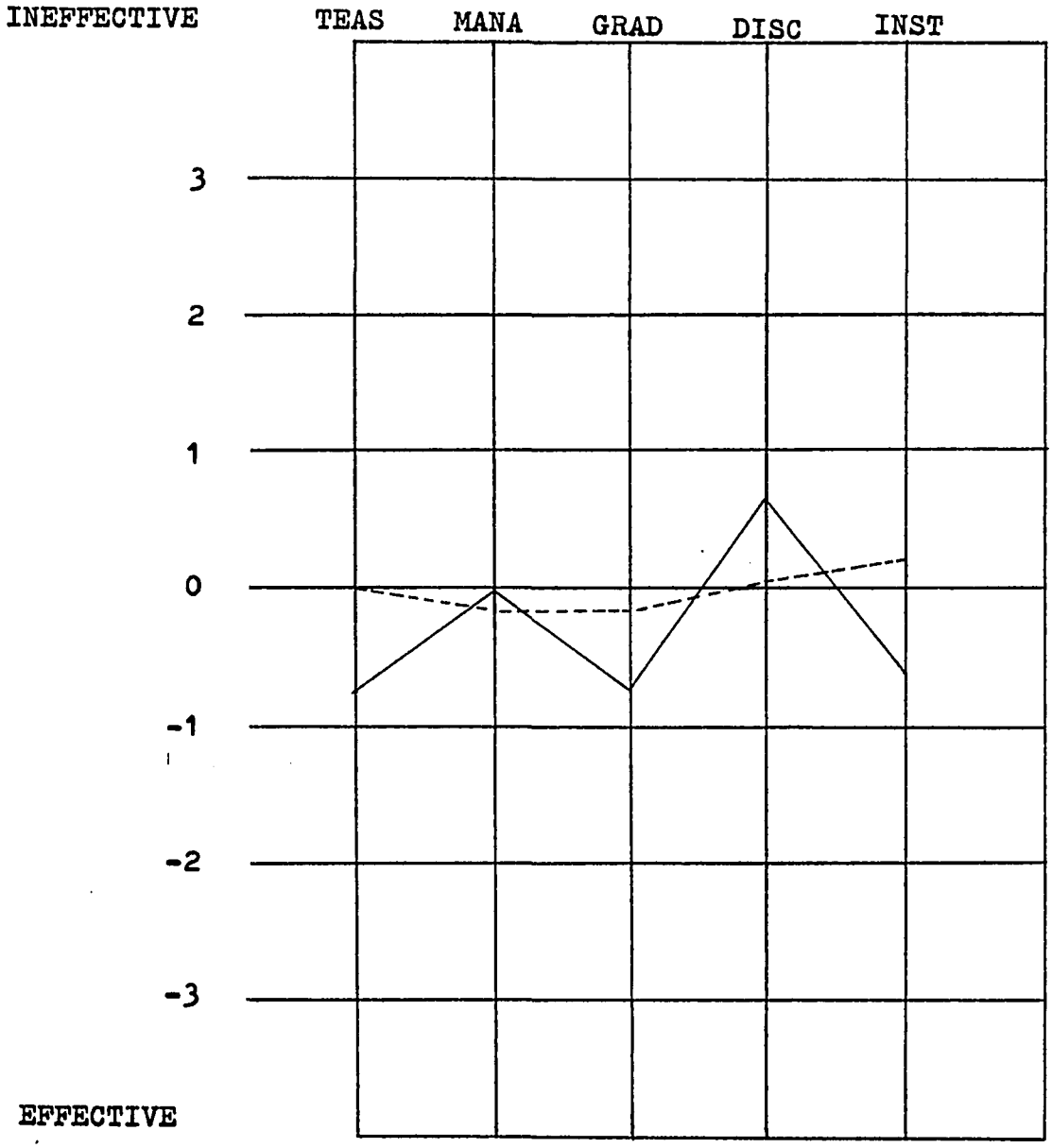


Note 1. Solid line denotes BARSSPET ratings.
2. Broken line denotes global ratings.

Figure 6

Profile Analysis of
Teacher 6 Ratings

Standardized Z Scores



Note 1. Solid line denotes BARSSPET ratings.
2. Broken line denotes global ratings.

Figure 7

Profile Analysis of
Teacher 7 Ratings

Standardized Z Scores

SUMMARY AND CONCLUSIONS

In this chapter, the results of the development and evaluation of the BARSSPET are summarized. The feasibility of obtaining student feedback of teaching performance is discussed. Included in this section are recommendations for the use of the BARSSPET. Finally, suggestions are presented for obtaining more data about psychometric properties of the BARSSPET.

Summary

The purpose of this investigation was to construct an instrument designed to provide student feedback on teaching performance in relation to program goals. The research was conducted during the 1984-1985 and 1985-1986 academic years.

BARSSPET development. The generation of behavioral statements consisted of collecting a pool of statements depicting effective, ineffective, and neutral physical education teaching performance. One hundred and twenty-eight students and teachers generated 735 behavioral statements.

The investigator and a group of five secondary students edited the statements. The investigator eliminated statements which were expressions of opinions or attitudes.

Redundant statements were also eliminated. The student group identified and eliminated words which might not be easily understood by sophomore students or particular gender and ethnic groups. Finally, specificity was added to selected general statements.

In the initial procedures of dimension identification, literature was reviewed for possible BARSSPET scale dimensions. Labels and definitions were identified for Grading, Rapport, Communication, Feedback, Management, Organization, Discipline, Enthusiasm, Instruction, Empathy, and Aggression. One hundred and forty-five students rated each category on its importance to teaching physical education. Ten dimensions were selected to be used in assigning behavioral statements. The dimension, Aggression, was not selected.

A sample of high school students who rated the teaching categories also rated the behavioral statements. Students rated the behavioral statements for accuracy in describing effective teaching behaviors. The students rated a total of 385 behavioral statements. Cronbach's alpha was used to measure the internal consistency of the ratings. Behavioral statements which depressed the alpha coefficient were eliminated. Two hundred and fifty statements were eliminated leaving 135 statements for further analysis.

A total of 80 secondary students participated in the initial procedures of allocating behavioral statements to

nine of the dimensions and assigning values to the behavioral statements. Only three behavioral statements were allocated to dimensions under the 60 percent agreement criteria. Thirty-five behavioral statements were allocated to dimensions using a 40 percent criteria. Eleven behavioral statements survived the standard deviation and chi-square criteria. Several factors were identified as contributing to unsuccessful scale development. It was surmised that students did not have adequate time to allocate statements to dimensions and also to assign values. Also, the dimensions identified in the literature and rated by students may have been a contributing factor to unsuccessful allocation as these labels might not have been meaningful constructs for students. Dimension identification and allocation and value assignment steps were repeated using new procedures.

In the alternative procedures of dimension identification, the investigator presorted the behavioral statements and assigned tentative labels and definitions using ideas found in the statements. Then a judging panel of five professional educators placed the behavioral statements into categories and critiqued the category labels and definitions. One hundred and seven of the statements, previously gained from the high school students, were placed into dimensions with a minimum of 60 percent agreement. The dimensions included Instruction, Teacher sensitivity, Enthusiasm, Communication, Managerial, Grading, and Discipline.

A total of 144 students participated in the second round of statement allocation and value assignment. This time all categories received more than the minimum to be retained for further analysis. Fifty-eight of the 107 allocated statements survived the 2.00 standard deviation criterion. Fifty-two behavioral statements survived the criterion of the Chi-square analysis. Five dimensions and 52 statements were retained for placement into the BARSSPET.

BARSSPET. The final BARSSPET consisted of three parts. The first part consisted of an orientation detailing the importance of teacher evaluation and the merits of student feedback and behaviorally anchored rating scales. The second portion of the BARSSPET consisted of five Likert-type scales corresponding to each BARSSPET scale. The Likert scales were global ratings included to help reduce scale rating errors. The final portion of the BARSSPET included five behaviorally anchored rating scales. These scales included the following categories: Teacher sensitivity, Managerial, Grading, Discipline, and Instruction.

Statistical analysis. A total of 175 students, from two schools, responded to the BARSSPET. The data generated from their responses provided the basis for the psychometric analysis. The ratings from each school were independently analyzed. Overall means and standard deviations for the raters at School One ranged from 2.67 to 4.12 and from 1.02 to 1.72 respectively. Means and standard deviations ranged

from 1.88 to 3.03 and from .56 to 1.32, respectively, for School Two. Dimension inter-correlations ranged from .26 to .40 for group 1 and from -.17 to .10 for School Two.

Psychometric analysis. Leniency bias was minimal for School One while school Two exhibited moderate leniency. In the results from School One, central tendency was judged to be moderate while, for School Two, central tendency was minimal. Halo bias was minimal for raters of both Schools.

Correlation coefficients on a test-retest procedure indicated that scale reliability ranged from low to moderate. Scale reliabilities (interrater reliability) reported in the literature were much higher (Fogli, Hulin, and Blood, 1979; Knoppers, 1978: and Smith and Kendall, 1963).

Discriminant validity was ascertained by correlation of BARSSPET scores with a measure of satisfaction (Kneer, 1976). The correlation coefficients indicated that little relationship existed between the BARSSPET dimensions and the Satisfaction subscales.

The nature of BARSSPET construction contributed to a claim of a high degree of content validity. Similarly, the sorting processes by judges, ratings of statements and internal consistency analysis, and allocation and value assignments by students led the investigator to conclude that the BARSSPET also possesses a degree of construct validity.

Conclusions

Within the limitations of this investigation and from the data analysis, the following conclusions can be drawn:

1. The anchors found on the scales of Part III of the BARSSPET provide a description of teaching behaviors of physical education teachers in the Cumberland County School system.

2. Unlike the results reported in the literature describing other behaviorally anchored rating scales (BARS), the anchors in each BARSSPET dimension presented a neutral level of performance as well as the ineffective and effective levels of performance.

3. The dimensions which appear in the BARSSPET would seem to be relevant to the teaching behaviors perceived by students in the processes of daily high school physical education.

4. The anchors for the dimensions contained in Part III of the BARSSPET represent student perceptions of teaching performance in physical education activity.

Implications for Teacher Evaluation

The implications for the use of the BARSSPET are directed toward use in the Cumberland County and Robeson County School systems.

Feedback. In the Cumberland County School system, there are no formal methods for obtaining student feedback with

regard to teaching performance. Feedback is usually obtained via class discussions, individual conferences, or student discussions. The reluctance of students to express their true feelings or to directly criticize teachers may be inhibiting factors. The present methods may not represent the best way to obtain reliable and valid information. The BARSSPET may represent one better way to obtain feedback from students about teaching performance in physical education activity.

Student perspectives. The behavioral statements represented as anchors in the BARSSPET are observed instances of behavior about which students in Cumberland County agree. These anchors clearly represent a student oriented perspective. This point becomes evident when one relates the developmental status of the adolescent tothe dimension definitions, anchors, and values assigned to those anchors. Consequently, teachers should not equate the rating data with the yearly evaluations conducted by administrative or supervisory personnel, but rather, view the student rater results as a way to better understand the unique perspective of the high school student.

Knowledge of adolescent development and psychology aid in understanding the values assigned to the anchors found in the BARSSPET. For example the anchor, "...gives us an "A" if we dress out every day and participate," was viewed as an effective teaching behavior by students while the anchor,

"...makes us take a skills test for our grade was viewed as an ineffective teaching behavior. Adolescent feelings about their changing physique, body functions, and skills adds credibility to student perceptions of these anchors. Contrarily, many educators would perceive these same anchors in a different light.

Student satisfaction and ratings. The relationship of the ratings by students at School One and the measures of satisfaction appeared to be independent. However, the correlation coefficients did not indicate a complete absence of a relationship between the BARSSPET and measures of satisfaction. Consequently, teachers should be aware of the potential for feelings of satisfaction contaminating the BARSSPET ratings. In fact, there might be some practical value in using the two schedules together deliberately so as to encourage students to analyze their own feelings of satisfaction while encouraging honest feedback to the teacher.

Recommendations for the BARSSPET

Orientation. Part I of the BARSSPET should be retained. The investigator felt that the orientation was necessary if students were not frequently involved in the evaluative process or were not familiar with rating scales. The information about student involvement in the evaluative processes may help to communicate its importance. However,

ways to reduce the length of the orientation should be investigated.

Global scales. Part II of the BARSSPET should be retained. Students usually required no more than five minutes to complete this section. The effect of the global scales should be ascertained through further investigation. An experimental design may be necessary to illustrate its effects legitimately. Once the effects of the global scales are known, a decision regarding the retention of Part II can be made.

Dimensions and anchors. No dimensions should be deleted from the BARSSPET. However additional statements should be added to the original pool of statements. Students should have another opportunity to allocate statements to dimensions to see if the Communication and Enthusiasm dimensions would be eliminated a second time.

The test-retest procedures yielded indications of low stability for each BARSSPET scale. However, in addition to student maturity levels, events associated with the end of the school year may have been contributing factors in the low scale stability. Consequently, additional data are necessary to fully ascertain the stability of the BARSSPET scales.

Additional investigation is warranted with a focus on the Instruction dimension. The ratings teachers received on the Instruction dimension were consistently indicated to be more effective than other dimension scores. Two anchors,

"...demonstrates all activities" and "...goes over the basic fundamentals of the sport," appeared to be more general than the remaining anchors on the scale. These two anchors were on the effective end of the scale. It may be that the two general anchors allow the rater more latitude in relating examples of the teacher's behaviors to the effective end of the scale.

Rater training. Prior to administration of the BARSSPET, the investigator discussed with the students the problem of rating errors. Future versions of the BARSSPET Rater's Guide should contain a section which describes rating errors and cautions for avoiding these errors.

Protocol for administration of the BARSSPET. Based on experiences in administering the BARSSPET, a protocol for its administration has been developed. The administration protocol appears in Appendix L.

Suggestions for Further Study

Assignment of values to behavioral statements. Students, in the first part of this study assigned values to the pool of behavioral statements. The mean scores for the value assignments were used to determine the scale position of the anchors (behavioral statements). Perceptions of performance effectiveness may be indicated by comparison of the mean ratings assigned to statements by students from different geographical locations. This information may be

useful in determining the usefulness of the BARSSPET outside of Cumberland County, N.C.

Convergent validity. The present study did not address the question of agreement in ratings of teachers by different components of the student population. Data on the extent of ratings agreement, or disagreement, would provide the user of the BARSSPET ratings a more comprehensive feedback profile. Consequently, future investigations using the BARSSPET should incorporate procedures to ascertain the extent of agreement on ratings by racial, gender, or grade level groups.

Data from Part One and Two of this investigation were obtained from students in required physical education classes. Ratings of teachers instructing in both elective and required physical education would be useful in determining the desirability of administering the BARSSPET in elective physical education classes.

REFERENCES

REFERENCES

A framework for physical education: Kindergarten through grade 12. Raleigh, N.C.: North Carolina Department of Public Instruction, 1979.

Aleamoni, L.M. Student ratings of instruction. in Millman, J. (Ed.) Handbook of teacher evaluation. Beverly Hills: Sage Publications, 1981.

Amidon, E.J. and Flanders, N.A. The role of the teacher in the classroom. St. Paul, Minn.: Association for Productive Teaching, 1971.

Beebee, R.J. The use of the behaviorally anchored rating scale in evaluating teacher performance. A paper presented at the Annual Conference of the Evaluation Network. September 30 - October 2, 1980 (ERIC Document Reproduction Service No. ED 194-567).

Beggs, D.L. and Lewis, E.L. Measurement and evaluation in the schools. Atlanta: Houghton Mifflen Company, 1975.

Berliner, D.C. The beginning teacher evaluation study: Overview and selected findings. A paper presented for the Conference on Research on Teacher Effects. University of Texas, Austin, Texas, November 2-4, 1975 (ERIC Document Reproduction Service No. ED 128-339).

Bernardin, H.J. Behavioral expectation scales versus summated scales: a fairer comparison. Journal of Applied Psychology. 1978, 62(4), 422-427.

Bernardin, H.J. Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology. 1978, 63(3), 301-308.

Bernardin, H.J. and Walter, C.S. Effects of rater training and diary keeping on psychometric error in ratings. Journal of Applied Psychology. 1977, 62(1), 301-308.

Bernardin, H.J. et al. A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology. 1976, 61(5), 564-570.

Bernardin, H.J., et al. Behavioral expectation scales: effects of developmental procedures and formats. Journal of Applied Psychology. 1976, 61(1), 75-79.

Bickwell, J.L. Nominated samples from public schools and statistical bias. American Educational Research Journal. 1974, 11, 333-341.

Biggs, J. and Chopra, P. Pupil evaluation of teachers. The Australian Journal of Education. 1979 23(1), 45-47.

Blanz, F. and Ghiselli, E.E. The mixed standard scale: A new rating system. Personnel Psychology, 1972, 25, 185-189.

Blood, M.R. Spinoffs from behavioral expectation scale procedures. Journal of Applied Psychology. 1974, 59(4), 513-515.

Bloom, B.S. The new direction in educational research: Alterable variables. Phi Delta Kappan. 1980, 61(6), 382-385.

Bolton, D.C. Selection and evaluation of teachers. Berkley, Calif.: McCutchan Publishing Co., 1973.

Bookhout, E.C. Teaching behavior in relation to the social -emotional climate of physical education classes. The Research Quarterly. 1967, 38(3),336-347.

Borman, W.C. Exploring upper limits of reliability and validity in performance ratings. Journal of Applied Psychology, 1978, 63, 561-565.

Borman, W.C. and Dunette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology. 1975, 60(5), 561-565.

Borman, W.C. and Vallon, W.R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. Journal of Applied Psychology. 1974, 59(2), 197-201.

Brophy, J.E. Teacher behavior and student learning. Educational Leadership. 1979, 37(1), 33-37.

Brown, F. G. Principals of educational and psychological testing. New York: Holt Rinehart and Winston, 1976.

Campbell, J.P., et al. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology. 1973, 57(1), 15-22.

Campion, J.F., et al. Work observation versus recall in developing behavioral examples for rating scales. Journal of Applied Psychology. 1973, 58(2), 286-288.

Carmines, E.G. and Zeller, R.A. Reliability and validity assessment. Beverly Hills, Calif.: Sage Publications, 1979.

Caruso, V.M. Teacher enthusiasm behaviors reported by teachers and students. A paper presented at the Annual Meeting of the American Educational Research Association. New York, N.Y., March 22, 1982 (ERIC Document Reproduction No. ED 217-038).

Census of Population and Housing. North Carolina State Data Center, Office of State Budget and Management. April, 1983

Cheffers, et al., Interaction Analysis: An application to Nonverbal Activity. St. Paul Minn.: P.S. Amidon, 1980.

Church, W.C. A catalog of core competencies for teachers physical education based on a theoretical model of pupil outcomes. Unpublished doctoral dissertation, Florida State University, 1974.

Colvin, W.W. and Roundy, E.S. An instrument for the student evaluation of teaching effectiveness in physical education. The Research Quarterly. 1977, 47(2), 296-298.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika. 1951, 16, 297-334.

Cronbach, L.J. Essentials of psychological testing (3rd ed.). New York: Harper and Rowe, Publishers, 1970.

Crowe, P. An observational study of teachers expectancy effects and their mediating mechanisms in physical education activity classes. Unpublished doctoral dissertation. The University of North Carolina at Greensboro, 1979.

Cruickshank, D.R. Synthesis of selected recent research on teacher effects. Journal of Teacher Education. 1976, 27(1), 61-64

Cummings, L.L. and Schwab, D.P. Performance in organizations: Determinants and appraisal. Glendview, Illinois: Scott, Foresman and Company, 1973.

Dalton, E.L. Pupil selection of teachers. Educational Leadership. 1971, 28(5), 476-479.

DeCotiis, T.A. A critique and suggested revision of behaviorally anchored rating scales developmental procedures. Educational and Psychological Measurement. 1978, 38, 681-690.

DeCotiis, T.A. An analysis of the external validity and applied revelance of three rating formats. Organizational Behavior and Human Performance. 1977, 19, 247-266.

Denton, J.J., et al. Pupil perceptions of a student teacher's competencies. Journal of Educational Research. 1976, 70(4), 180-5.

Dickinson, T.L. and Zellinger, P.M. A Comparison of the behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology. 1980, 65(2), 147-154.

Doyle, K.O. Evaluating teaching. Lexington, Mass.: D.C. Heath and Company, 1983.

Eastridge, H.E. Student evaluation and teacher performance. NASSP Bulletin. 1976, 60(401), 48-54.

Flanagan, J.C. The critical incident technique. Psychological Bulletin. 1954, 51, 327-355.

Fogli, L., et al. Development of first level behavior job criteria. Journal of Applied Psychology. 1971,55(1), 3-8.

Finley, D.M., et al. Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology. 1977, 30, 659-669.

Frey, P.W., et al. Student ratings of teaching: validation research. American Educational Research Journal. 1975, 12(4), 435-447.

Goodale, J.G. and Burke, R.J. Behaviorally based rating scales need not be job specific. Journal of Applied Psychology. 1975, 60(3), 389-391.

Graham, G. and Heimerer, C. Research on teacher effectiveness: A summary with implications for teaching. Quest. 1981, 33(1), 14-25.

Grasha, A.F. Assessing and developing faculty performance. Cincinatti, Ohio: Communication and Education Association, 1977.

Green, S.B., et al. Shortcut methods for deriving behaviorally anchored rating scales. Educational and Psychological Measurement. 1981, 41(3), 761-775.

Guidelines for secondary physical education. A position paper by the National Association for Sport and Physical Education. Reston, Va.: American Alliance for Health, Physical Education, Recreation, and Dance, 1979.

Guilford, J.P. Psychometric methods. New York: McGraw-Hill, 1967.

Hall, G.E. (ed.) Research and development agenda in teacher education. Austin, Texas: The Research and Development Center for Teacher Education, 1979.

Harari, O. and Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology. 1973, 58(2),261-265.

Isaac, S. and Michael, W.B. Handbook in Research and evaluation. San Diego, Calif.: Edits Publishers, 1980.

Isaacson, R.L., et al. Dimensions of student evaluations of faculty teaching. Journal of Applied Psychology. 1964, 55(6), 344-351.

Jewett, A.E. and Mullan, M.R. Curriculum design: Purposes and processes in physical education teaching-learning. Reston, Va.: American Alliance for Health, Physical Education, Recreation, and Dance, 1977.

Jordan, J.B. The development and evaluation of behaviorally anchored rating scales for the assessment of department chairman performance. Unpublished doctoral dissertation, University of Kentucky, 1976.

Kafry, D., et al. Discriminability in multidimensional performance evaluations. Applied Psychological Measurement. 1979, 3(2), 187-192.

Kaufman, B.J. and Madden, J.M. The development of behaviorally anchored rating scales for student evaluation of college teaching effectiveness. A paper presented at the Annual Meeting of the Eastern Psychological Association, Hartford, Ct., April 9-12, 1980 (ERIC Document Reproduction Service No. ED 189-995).

Kearney, W.J. Behaviorally anchored rating scales - MBO's missing ingredient. Personnel Journal. 1979, 58(1), 20-25.

Keaveny, T.J. and McGann, A.F. A comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology. 1975, 60(6),695-703.

Kepler, K.B. Descriptive feedback: Increasing teacher awareness, adapting research techniques. A paper presented at the Annual Meeting of the American Educational Research Association, New York, April 4-8, 1977 (ERIC Document Reproduction Service No. ED 139-738).

Kerlinger, F.M. Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart, and Winston, Inc., 1973.

Kneer, M.W. Influence of selected factors and techniques on student satisfaction with a physical education experience. Unpublished doctoral dissertation, University of Michigan, 1972.

Kneer, M.E. The role of student satisfaction in developing play skills and attitudes. Quest. 1976, 26, 102-106.

Knoppers, A. A behaviorally based appraisal of coaching performance in women's athletics at Michigan State University. Unpublished doctoral dissertation, University of North Carolina-Greensboro, 1978.

Knouse, S.B. and Rodgers, D.T. An analysis of the resident-assistant position based on the behaviorally anchored rating-scales technique. Journal of College Student Personnel. 1981, 22(5), 396-400.

Landy, F.J. and Guion, R.M. Development of scales for the measurement of work motivation. Organizational Behavior and Human Performance. 1970, 5, 93-103.

Landy, F.J., et al. Behaviorally anchored scales for rating the performance of police officers. Journal of Applied Psychology. 1976, 61(6),750-758.

Larson, J.R. The limited utility of factor analytic techniques for the study of implicit theories of student ratings of teacher behavior. American Educational Research Journal. 1979, 16(2), 201-211.

Latham, G.P. and Wexley, K.M. Behavioral observation scales for performance appraisal purposes. Personnel Psychology. 1977, 30, 255-268.

Lawler, F.E. The multitrait-multirater aproach to measuring managerial job performance. Journal of Applied Psychology. 1967, 51(2), 177-183.

Lombardo, B.J. and Cheffers, J.T. Variability in teaching behavior and interaction in the gymnasium. Journal of Teaching in Physical Education. 1983, 2(2), 33-47.

McDonald, F.J., et al. Beginning teacher education study, phase II: Final report. Princeton, N.J.: Educational Testing Service, 1975.

McKeachie, W.J., et al. Student ratings of teacher effectveness: Valitity studies. American Educational Research Journal. 1971, 8(3), 435-445.

McKethan, J.F. Student attitudes toward instructional processes in secondary physical education. Unpublished doctoral dissertation, University of North Carolina-Greensboro, 1979.

Masters, J.R. and Weaver, W.G. The development of a student observation of teachers instrument for use in high schools. A paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, N.Y., April 5-7, 1977 (ERIC Document Reproduction Service No. ED 141-409).

Martinek, T., et al. Pygmalion in the gym: Causes and effects of expectations in teaching and coaching. West Point, N.Y.: Leisure Press, 1982.

Martinek, T. and Johnson, S. Teacher expectations on dyadic interactions and self concept in elementary age children. The Research Quarterly. 1979,50, 60-70.

Martinek, T. and Mancini, V. CAFIAS: Observing dyadic interactions between teacher and students. Journal of Classroom Interaction. 1979, 14, 18-23.

Motowildo, S.J. and Borman, W.C. Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology. 1977, 62(2), 177-183.

Noble, L. and Cox, R.H. Development of a form to survey student reactions on instructional effectiveness of lifetime sports classes. Research Quarterly for Exercise and Sport. 1983, 54(3), 247-254.

Norris, W.R. Seven premises for improving teacher evaluation. NASSP Bulletin. 1980, 64(434), 30-35.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Oliver, B. Improving evaluation of physical education teachers. The Clearing House. 1980, 54(3), 82-84.

Owen, S.A. The validity of student ratings: A critique. A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April 19-23, 1976 (ERIC Document Reproduction Service No. ED 129-902).

Patrick, O.L. Ethnic students' perceptions of effective teachers. Educational Research Quarterly. 1978, 3(2), 67-73.

Peters, D.L. and McCormick, E.J. Comparative reliability of numerically anchored versus job-task anchored rating scales. Journal of Applied Psychology. 1966, 60(1), 92-96.

Phillips, D.A. and Carlisle, C. A comparison of physical education teachers categorized as most effective and least effective. Journal of Teaching in Physical Education. 1983, 2(3), 247-254.

Pieron, M. and Hacourt, H. Teaching behaviors at different levels of physical education teaching. FIEP Bulletin. 1979, 49(2), 33-41.

Popham, W.J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1975.

Price, M. The development of behaviorally anchored rating scales for the performance evaluation of special education teachers. Unpublished doctoral dissertation, Temple University, 1979.

Purkey, W.W. Inviting school success. Belmont, Calif.: Wadsworth Publishing Company, Inc., 1978.

Rosenshine, B. Recent research on teaching behaviors and student acheivement. Journal of Teacher Education. 1976, 27(1), 61-64.

Saklofske, et al. A psychometric study of a measure of teachers' directiveness, student perception of teacher style. Perceptual and Motor Skills. 1980, 51(1), 192-194.

Schoolmen split on student evaluation of teachers. Nation's Schools. October, 1970, 53.

Schwab, D.P., et al. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology. 1975, 28, 549-562.

Self Study Report of Cape Fear Senior High School. Cumberland County Schools, Fayetteville, N.C., 1982.

Self Study Report of Douglas Byrd Senior High School. Cumberland County Schools, Fayetteville, N.C., 1974.

Self Study Report of Seventy-First Senior High School. Cumberland County Schools, Fayetteville, N.C. 1972.

Shaw, J.S. Students evaluate teachers and it works. Nation's Schools. 1973, 91(4), 49-53.

Siedentop, D. and Hughley, C. O.S.U. Teacher behavior rating scale. Journal of Physical Education and Recreation. 1975, 46(2), 45.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw Hill Book Co., 1956.

Sizemore, R.W. Do black and white students look for the same characteristics in teachers. Journal of Negro Education. 1981, 50(1), 48-53.

Smith, J.P. and Brown, T.J. Relationship among secondary students' evaluations of their courses and teachers. A paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 19-23, 1976 (ERIC Document Reproduction Service No. ED 129-867).

Smith, P.C. and Kendall, L.M. Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology. 1963, 47(2), 149-155.

SPSS-X user's guide. (1983). New York: McGraw-Hill Book Company.

Steele, J.M., et al. An instrument for assessing instructional climate through low-inference student judgements. American Educational Research Journal. 1971, 8(3), 447-467.

Sullivan-Kowaski, J.P. Evaluating teacher performance. Arlington, Va.: Educational Research Service, 1978.

Thompson, B.L. Secondary school pupils' attitude to school and teachers. Educational Research. 1975, 18(1), 62-66.

Thompson, J.E. Student evaluation of teachers. NASSP Bulletin. 1974, 58(384), 25-30.

Traugh, C.E. and Duell, O.K. Secondary students' attitudes toward and experience with evaluating teachers. High School Journal. 1980, 64(3), 97-107.

Tuckman, B.W. A technique for the assessment of teacher directiveness. Journal of Educational Research. 1970, 63(9), 395-398.

Whitney, S.E. and Doyle, K.D. Implicit theories in student ratings. American Educational Research Journal. 1976, 13(4), 241-253.

Wiersna, W. Research methods in education (2nd ed.). Itasch, Ill.: F.E. Peacock Publishers, Inc., 1975.

Wilkinson, K.L. Effective and ineffective teacher behaviors as viewed by students in secondary business law classes. Delta Phi Epsilon Journal. 1979, 21(1),1-11.

Zakrajsek, D.B. Student evaluations of teaching performance. Journal of Physical Education and Recreation. 1978, 49, 64-65.

Zedeck, S. and Baker, H.T. Nursing performance as measured by expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance. 1972, 7, 457-466.

Zedeck, S., et al. Behavioral expectations: development of parallel forms and analysis of scale assumptions. Journal of Applied Psychology. 1976, 61(1), 112-115.

Zedeck, S., et al. Development of behaviorally anchored rating scales as a function of organizational level. Journal of Applied Psychology. 1974, 59(2), 249-252.

Zedeck, S., et al. Format and scoring variations in behavioral expectation scales. Organizational Behavior and Human Performance. 1976, 17, 171-184.

APPENDIX A

Orientation
Informed Consent

ORIENTATION

Introduction

I am a student at the University of North Carolina at Greensboro. As a part of my graduation requirements, I must develop and complete a dissertation. The topic that I have selected requires me to develop and evaluate a behaviorally anchored rating scale for the appraisal of secondary physical education teachers.

Rating Scales

The set of rating scales which I will develop is not as common as other types of rating scales. Most rating scales use numbers or simple traits to help the rater make judgements about qualities. The scale that I am developing will have, in addition to numbers, observed behaviors to assist the rater in making judgements.

The Raters

The people who will use the rating scale will be students rather than teachers or administrators. There are a number of reasons for using students instead of teachers or administrators. First, students see the teacher more than the principal or other administrators. Unfortunately, students are rarely called upon to help evaluate teachers. It only seems logical that teachers should be aware of how their clients, students, view their performance. Secondly, student evaluations should be a part of a teacher's overall evaluation. Finally, students are accurate and reliable with their perceptions of teaching behavior.

I am going to solicit from students observations of teacher behavior which range from very effective to very ineffective. A second group of students will determine how accurately these behaviors describe effective teaching. The second group will also identify categories of teaching behaviors. A third group of students will be asked to assign a value to each behavior and place each behavior into a category. Using the information obtained from each of the three groups, I will construct the rating scale. When the rating scale is complete, another group of students will rate their physical education teachers.

Importance

For me, the effort that I am about to undertake is a very serious one. Its importnace is in, not only the graduation requirement, but also in making a worthwhile contribution to my profession. It is also a very serious responsibility for those students who agree to take part in the effort. With my best efforts, the finished rating scale can only be as good as the information provided by you, the student. For this reason, students must be accurate and honest when making statments, ratings, or assignments of items to categories.

INFORMED CONSENT

The purpose of this investigation is to develop a rating instrument for the evaluation of secondary physical education teachers. Students, such as yourself, will provide the information necessary to construct the rating scale. Another purpose of this research is to see if different groups of students rate physical education teachers differently.

Please understand that your participation in this research is completely voluntary and that you are not being pressured to participate. You have the right to withdraw your consent to participate at any time.

Any information that you provide will remain completely anonymous. At the completion of this research, a written summary of the results will be made avaliable to you at your request.

Please check one of the following:

____ I agree to participate in this investigation.

____ I am not willing to participate in this investigation.

_____
(Parent signature)

_____
(Student signature)

_____
(Date)

APPENDIX B

Correspondence

THE UNIVERSITY OF ILLINOIS AT CHICAGO

Department of Physical Education
College of Health, Physical Education, and Recreation
Box 4348, Chicago, Illinois 60680
(312) 996-4600

October 19, 1983

Mr. Robert N. McKethan
346 Fillyaw Road
Fayetteville, N.C. 28303

Dear Robert:

I'm delighted that you are interested in using my satisfaction inventory. I have enclosed a copy. Additional information can be secured through the Dissertation Abstracts (late 1972 or 1973) and an article in Quest, #26, 1976, pg. 102-108.

If you decide you use it, you have my permission to do so. Let me know if I can be of further help.

Sincerely,

Marian E. Kneer

Marian E. Kneer
Professor

March 1, 1984

Dear Coach,

I am collecting data for the development of a rating scale for the evaluation of secondary physical education teaching performance. A part of data collection consists of generating examples of physical education teaching behaviors. Groups of students at Douglas Byrd, Cape Fear, and Seventy-First have already contributed a sample of behavioral statements. My purpose in asking teachers to generate examles of teaching behaviors is to insure that I have a representative sample of teaching behaviors from which to develop the rating scale.

I am asking for your assistance in this important step of data collection. Enclosed are an orientation to my project, directions for generating the behavioral statements, the instrument on which to record responses, and a stamped addressed envelope.

I am asking that the instruments, on which you record your responses, be returned to me by March 15. Use the stamped envelope to return your responses. If you are unable to respond to this request, please return the instrument using the stamped envelope.

Thank you for your attention, your time, and your consideration. I am appreciative of your efforts.

Respectfully,

Robert N. McKethan

Robert N. McKethan

April 26, 1984

Dear Coach Taliani,

I am beginning the second part of the data collection for the development of a rating scale for secondary physical education teachers. The second phase of data collection will consist of ratings of behavioral statements supplied by students and physical education teachers. Groups of students at Douglas Byrd, Cape Fear, and Seventy-First have responded to the same set of behavioral statements. By having teachers rate the same set of statements, I will be able to determine the differences that exist between student and teacher perceptions of same teacher behaviors.

I am asking for your assistance in this important step of data collection. Your response to the enclosed instrument would be greatly appreciated. Please note that part one of the instrument requires a rating of behavioral statements while part two requires a rating of categories of teacher behavior. If, after responding to the instrument, you note that there are some important behaviors missing; please include them on the page entitled, ADDITIONAL BEHAVIORS.

I am asking that the rating instruments, with your responses, be returned to me by May 11. Use the stamped return envelop to return your responses. If you are unable to respond to the rating, please return the rating instrument using the return envelop.

Thank you for your attention, your time, and your consideration. I am appreciative of your efforts.

Respectfully,

Robert N. McKethan

November 6, 1984
346 Fillyaw Road
Fayetteville, N.C.
28303

Mr. Purnell Swett,
Superintendent
Robeson County Schools
Box 1328,
Lumberton, N.C. 28358

Dear Mr. Swett,

I am appreciative of your willingness to listen to my request for including the Robeson County high school students in my dissertation work. Enclosed, you will find a prospectus detailing the involvement of the Robeson County students.

Again, thank you for including my proposal on the agenda for the next meeting of the Robeson County School Board. I look forward to your response.

Respectfully,
Robert N. McKethan
Robert N. McKethan

PROSPECTUS

Robert N. McKethan

DISSERTATION TOPIC

The development of a Behavioraly Anchored Rating Scale for Secondary Physical Education Teachers.

PURPOSE

The purpose of this investigation is the development of an evaluation instrument to provide feedback on student perceptions of physical education teaching performance.

SUBJECTS

The subjects for this investigation are secondary students from Cumberland and Robeson Counties.

CONSENT FORMS

Participating students must sign a consent form. Students will have the option to withdraw from participation at any time. SEE ATTACHMENT.

TASK

Robeson County students will perform two tasks. Students will sort 135 behavioral statements (collected from Cumberland County students) into seven categories of teaching performance. Another group of students will decide the level of effectiveness of the behavior found in the statements.

TIME REQUIREMENT

Each student will work with a maximum of 34 behavioral statements. Students will either sort the behavioral statements into categories or decide the effectiveness of the behavioral statements. Students will not be required to complete both tasks.

Assuming there is good participation on the part of the students, six class periods will be more than adequate. Three class periods will sort the behavioral statements into seven categories. The other three periods will determine the effectiveness of the behavioral statements.

Within a class period, there will be four groups of students performing the task. for example, Group a will work

with statements 1-34, Group B will work with statements 35-68, Group C will work with statements 69-102, and Group D will work with statements 103-135. The organizational procedure, appearing in the preceeding lines, is necessary to insure that students are able to complete their task within a single class period.

NUMBER OF SUBJECTS

A minimum of 40 students will be needed to sort the behavioral statements into categories. Also, 40 students will be needed to decide the effectiveness of the behavioral statements.

SCHOOL LOCATION

There is no preference for a particular high school. There is, however, a need to have students who are motivated to complete the task as well as possible.

TARGET DATE

The target date for data collection is the week of November 26-30. However, I will work with the principal or School Board on deciding a mutually agreeable date.

PROSPECTUS

Robert N. McKethan

DISSERTATION TOPIC

The development of a Behavioraly Anchored Rating Scale for Secondary Physical Education Teachers.

PURPOSE

The purpose of this investigation is the development of an evaluation instrument to provide feedback on student perceptions of physical education teaching performance. The purpose of this phase of the investigation is to provide data for reliability and validity assessment of the rating scales.

SUBJECTS

The subjects for this phase of the investigation are secondary physical education students from Westover Senior High School. A minimum of 150 students are needed. The first and second period physical education classes will meet this need.

CONSENT FORMS

Participating students must sign a consent form. Students will have the option to withdraw from participation at any time. SEE ATTACHMENT.

TASK

Westover students will perform two tasks. First, students will respond to a rating scale constructed from data obtained from students and teachers in Cumberland and Robeson Counties. The purpose of this rating scale is to provide student assessment of teacher behaviors in several categories. SEE ATTACHMENT #2.

The second task requires that students respond to a set of scales designed to measure student satisfaction of the physical education experience. The purpose of this task is to see is measured satisfaction is independent of student ratings of teacher behavior. SEE ATTACHMENT #3.

TIME REQUIREMENT

Students should be able to complete both tasks within a

single class period. Students should report to the testing location (cafeteria) as soon as possible. Both tasks will require approximately 40-45 minutes.

TEACHER RESPONSIBILITIES

Teachers will not be burdened with excessive work to prepare for the data collection. However, teachers should encourage as many students as possible to participate. Teachers will be asked to have students sign the consent forms on a day prior to the data collection. Also, teachers should get students to the cafteria as rapidly as possible. During the administration of both rating scales, teachers should not be present.

TARGET DATE

March 7 and March 21. On March 21 students will resond only to the teacher behavior rating scales. The purpose of the second adiministration is to provide data on scale reliability.

RESULTS

Results will be available at the request of the physical education teachers.

APPENDIX C

INSTRUMENT FOR DATA COLLECTION

Collection of behavioral statements

DIRECTIONS

Collection of statements

Introduction

There are five objectives for your physical education program. These objectives show how the physical education program is supposed to help you. These objectives are listed below.

1. To provide experiences that will promote physical fitness.
2. To provide opportunities for skill development in life-time sports.
3. To provide experiences that will promote group interaction and communication.
4. To provide experiences that will promote a positive self image.
5. To provide experiences that promote emotional stability.

Think of times when you saw a physical education teacher do something that was very effective or very ineffective. You may think of things that happened before you were in this class (when you were in another class). These incidents or things that happened should "stand out" in your mind in order to be considered. These behaviors that you saw do not necessarily have to involve you, but should be occurrences of behaviors that you saw. The emphasis is not on feelings or attitudes about the teacher, but on behaviors that you saw happen.

Directions

1. In each space on page 3, write down a sentence that tells about a behavior that you saw. These sentences telling about a behavior that you saw may have happened in your class, another class, this year, or a previous year.

2. Look at the objectives that are listed above. If the actions helped to accomplish an objective, then it is an effective behavior. If the actions interfered with accomplishing of an objective, then it is an ineffective behavior. If the actions did not really help or interfere in the accomplishment of an objective, then it is a neutral behavior.

3. In the box next to the space where you put your sentence, circle an "E", an "N", or an "I".
   E = Effective behavior
   N = Neutral behavior
   I = Ineffective behavior

4. List at least
   A. 2 Effective behaviors
   B. 3 Neutral behaviors
   C. 2 Ineffective behaviors

You may list more behaviors than which is being asked.

5. Put only one statement in each space.

6. Please do not use the names of people.

Examples

Effective behavior	The teacher allows time for questions after giving directions.
Neutral behavior	The teacher leads the warm-ups instead of the students.
Ineffective behavior	The teacher yells at a student for making a mistake.

Self study Reports: Cape Fear Senior High School, 1982; Douglas Byrd Senior High Senior high School, 1974; Seventy-First Senior High School, 1972.

BEHAVIORAL STATEMENTS

Complete the information found below by circling the response that best applies to you.

A. Your high school class: Sophomore Junior Senior
B. Your ethnic background: White Black Indian Hispanic Asian
C. Your Sex: Male Female
D. Your P.E. grade (1st semester) A B C D E

Put the sentence describing the behavior here. Circle "E", "N", or "I" in the space to the right.	E N I
	E N I
	E N I
	E N I
	E N I
	E N I
	E N I
	E N I
	E N I
	E N I

APPENDIX D

INSTRUMENT FOR DATA COLLECTION

Rating of Behavioral Statements

Rating of Behavioral Statements

Student Information

Please circle one response for each of the questions found below.

A. What is your class? Sophomore Junior Senior
B. What is your sex? Male Female
C. What is your ethnic background? White Black Indian Hispanic Asian
D. What was your first semester P.E. grade? A B C D E
E. What is your third nine weeks P.E. grade? A B C D E

Introduction

The behavioral statements found on this page and the following pages were recorded by students and P.E. teachers from three high schools in Cumberland County. These statements are behaviors that P.E. teachers show in their P.E. classes. The purpose of this rating is for you to decide how accurately these behaviors describe effective physical education teaching behaviors.

Directions

You are to rate the behavioral statements listed below and on the following pages. You must decide how well each statement shows effective teaching behaviors.

Use the scale appearing below each behavior statement to help you rate the statement. If the statement describes an effective teacher behavior most accurately, put a check on the line above the 1. Example A shows how a most accurate behavioral statement, which all students see, is marked.

A. The teacher checks the roll.

most accurate ___: ___: ___: ___: ___:___: ___: most inaccurate
1 2 3 4 5 6 7

If you decide that a behavioral statement describes an effective teaching behavior least accurately, check the line above the 7. Example B shows how a most inaccurate behavior, about which all students know, is marked.

B. The teacher wears a shirt and tie.

most accurate ___: ___: ___: ___: ___:___: ___: most inaccurate
1 2 3 4 5 6 7

Some of the statements won't really describe effective teaching behavior most accurately or most inaccurately. This kind of behavioral statement describes teacher behavior

somewhere in between most accurate and most inaccurate. You will put your check on a number between 1 and 7 that shows where the statement belongs. Example C shows an example of a behavior that is neither most accurate or most inaccurate.

C. The P.E. teacher corrects your grammar.

most accurate ___1___: ___2___: ___3___: ✓4: ___5___:___6___: ___7___: most inaccurate

APPENDIX E

INSTRUMENT FOR DATA COLLECTION

Dimension Identification
Initial Procedures

RATING OF DIMENSIONS

Introduction

Dimensions are groups or categories. A category of behaviors includes groups of behaviors that are similar in someway. For example, "communication" is a dimension (category) of teacher behavior. Behaviors that fall into the communication category include behaviors in which the teacher is trying to get across an idea to the class or where the teacher is attempting to understand the student.

The behaviors that you have just rated may be sorted into categories. However, before the statements can be sorted, students must decide which dimensions (categories) are most important to the teaching of physical education.

Directions

For the categories listed below and on the following page, you must decide how important each is to teaching of physical education. When deciding whether a category is most important or least important, you will place your check on the scale in the same manner as you did for rating the behavioral statements.

If the dimension (category) is most important to P.E. teaching, then place your check on the line above the 1. Place your check on the line above the 7 if you decide that the dimension (category) is least important to P.E. teaching. If the category is not really most important or least important, you will put your check on a number that shows where the category belongs.

Categories

1. FORCEFULNESS - Behaviors in this category include teacher actions that (a) criticize others publicly, (b) make fun of others, (c) tell off others when when met with disagreement, and (d) show anger.

most important $\underset{1}{\_\_}$: $\underset{2}{\_\_}$: $\underset{3}{\_\_}$: $\underset{4}{\_\_}$: $\underset{5}{\_\_}$: $\underset{6}{\_\_}$: $\underset{7}{\_\_}$: least important

2. EMPATHY - Behaviors in this category include teacher actions that show (a) an understanding of how others feel, (b) putting oneself in another person's place, and (c) a judgement of people by what they do rather than what they are.

most important $\underset{1}{\_\_}$: $\underset{2}{\_\_}$: $\underset{3}{\_\_}$: $\underset{4}{\_\_}$: $\underset{5}{\_\_}$: $\underset{6}{\_\_}$: $\underset{7}{\_\_}$: least important

3. GRADING - Behaviors in this category include teacher actions that show the ways that student grades are assigned.

most important __1__: __2__: __3__: __4__: __5__: __6__: __7__: least important

4. RAPPORT - Behaviors in this category are teacher actions that show (a) how the teacher deals with students, (b) the teacher's interest in student learning, and (c) the teacher's interest in skill development.

most important __1__: __2__: __3__: __4__: __5__: __6__: __7__: least important

5. MANAGERIAL - Behaviors in this category are teacher actions that show (a) how the teacher directs the class to change activities and (b) roll taking, marking down performance scores, and other forms of record keeping.

most important __1__: __2__: __3__: __4__: __5__: __6__: __7__: least important

6. INSTRUCTIONAL - Behaviors in this category are teacher actions that show (a) how skills are described and demonstrated, (b) abilities to explain game rules, and (c) how game strategies are explained and used in activities.

most important __1__: __2__: __3__: __4__: __5__: __6__: __7__: least important

7. COMMUNICATION - Behaviors in this category are teacher actions that show (a) abilities and security in expressing themselves to students and (b) the freedom and security students have in expressing themselves to other students and the teacher.

most important __1__: __2__: __3__: __4__: __5__: __6__: __7__: least important

8. DISCIPLINE - Behaviors in this category are teacher actions that show (a) how the teacher maintains class control, (b) dealing with inappropriate behavior, and (c) the assignment of work as a response to inappropriate behavior.

most important __1__: __2__: __3__: __4__: __5__: __6__: __7__: least important

9. ENTHUSIASM - Behaviors in this category are teacher actions that show (a) how excited the teacher is about his/her work, (b) what the teacher does to encourage students in participation, and (c) what the teacher does to get students to feel good about class.

most important ___1: ___2: ___3: ___4: ___5: ___6: ___7: least important

10. ORGANIZATION - Behaviors in this category are teacher actions that show (a) how the teacher prepares for class and (b) how well class time is used for skills activities, and warm-ups.

most important ___1: ___2: ___3: ___4: ___5: ___6: ___7: least important

11. FEEDBACK - Behaviors in this category are teacher actions that show how students are informed or told of their progress in skill development, knowledge of game rules, sportsmanship, and class behaviors. Feedback behaviors do not include grading.

most important ___1: ___2: ___3: ___4: ___5: ___6: ___7: least important

APPENDIX F

INSTRUMENT FOR DATA COLLECTION

Allocation of Statements and Assignment of Values
Initial Procedures

ASSIGNMENT OF BEHAVIORAL STATEMENTS

ASSIGNMENT OF NUMERICAL VALUES

Introduction

On page 2 are listed 9 categories which are related to the teaching performance of physical education teachers. These categories were identified by students from several Cumberland County high schools. Each category is identified by a letter of the alphabet (A-I).

Beginning on page 4 are 135 statements which describe examples of teaching behaviors. These behaviors were identified and recorded by a different group of students from the three Cumberland County high schools.

You will be asked to place each of the 135 statements into a category. Additionally, you will be required to to decide how effective or ineffective the behavior is by assigning a value to each behavior. This task will require you to make 270 decisions before the end of the period.

Your task is a very important responsibility. Concentration and effort on your part is needed to insure that your responses are most accurate. The information that you provide will be used to construct a rating scale to evaluate physical education teaching performance.

The following suggestions will aid you in your task.

1. By reducing the amount of talking with your classmates, you will have more time for your task.

2. If you have any questions please raise your hand for help.

3. When you start your task, place the directions at the top of your desk, the statements at the bottom on one side and the answer sheet on the other side.

ASSIGNMENT TO CATEGORIES

Categories of behaviors include groups of behaviors that are similar in some way. The categories into which you will be putting behavior statements are identified and defined below. Notice that each category is identified by a letter of the alphabet.

A. GRADING - Behaviors in this category include teacher actions that show the ways that student grades are assigned.

B. EMPATHY - Behaviors in this category include teacher actions that show (a) an understanding of how others feel, (b) the putting of oneself in another person's place, and (c) a judgement of people by why they act rather than what they do.

C. RAPPORT - Behaviors in this category are teacher actions that show (a) how the teacher deals with students, (b) the teacher's interest in student learning, and (c) the teacher's interest in skill development.

D. COMMUNICATION - Behaviors in this category are teacher actions that show (a) abilities and security in expressing themselves to students and (b) the freedom and security students have in expressing themselves to the teacher.

E. FEEDBACK - Behaviors in this category are teacher actions that show how students are informed or told of their progress in skill development, knowledge of game rules, sportsmanship, and class behaviors. Feedback behaviors do not include grading.

F. MANAGERIAL - Behaviors in this category are teacher actions that show (a) how the teacher directs the class to change activities and (b) roll taking, marking down performance scores, and other forms of record keeping.

G. ORGANIZATION - Behaviors in this category are teacher actions that show (a) how the teacher prepares for class and (b) how well class time is used for skills, activities, and warm-ups.

H. DISCIPLINE - Behaviors in this category are teacher actions that show (a) how the teacher maintains class control, (b) dealing with inappropriate behavior, and (c) the assignment of work as a response to inappropriate behavior.

I. ENTHUSIASM - Behaviors in this category are teacher actions that show (a) how excited the teacher is about his/her work, (b) what the teacher does to encourage students in participation, and (c) what the teacher does to get students to feel good about class.

After reading the statement, please write the letter (A-I) for the category in which the statement should be placed. You may choose <u>one</u> category per statement. You <u>must choose one</u> for each statement.

<u>example</u>

1. The teacher calls us by our first name.

Statement no.	Dimension A-I	Rating 1-7	Statement no.	Dimension A-I	Rating 1-7
1.	D	___	2.	___	___

The letter "D" was selected because the behavior best fits in the "Communication" category.

ASSIGNMENT OF VALUES

After assigning a statement to a category, rate the statement in terms of its effectiveness as a teaching behavior by assigning it a number from 1-7. A value of 1 represents a very effective teaching behavior; 4 represents neutral behaviors and 7 represents an extremely ineffective behavior.

Use the following physical education program objectives to help you decide whether a behavior is extremely effective, neutral, or extremely ineffective.

1. To provide experiences that will promote physical fitness.
2. To provide opportunities for skill development in life-time sport.
3. To provide experiences that will promote group interaction and communication.
4. To provide experiences that will promote a positive self image.
5. To provide experiences that promote emotional stability.

Look at the objectives that are listed above. If the actions helped to accomplish an objective, then it is an effective behavior. if the actions interfered with accomplishing of an objective, then it is an ineffective behavior. If the actions do not really help or interfere in the accomplishment of an objective, then it is a neutral behavior.

In the column Rating 1-7, put the number which describes the effectiveness of the behavior. A "1" represents a very effective behavior. The value, "4", represents neutral behaviors while "7" is representative of very ineffective behaviors.

example

1. The teacher calls us by our first name.

Statement no.	Dimension A-I	Rating 1-7	Statement no.	Dimension A-I	Rating 1-7
1.	___	1	2.	___	___

The number "1" was selected because the behavior is an extremely effective behavior. The behavior provides an experience that promotes positive feelings about oneself or promotes communication.

APPENDIX G

INSTRUMENT FOR DATA COLLECTION

Dimension Identification
Alternative Procedures

October 8, 1984

Dear Panel Member,

Thank you for agreeing to assist in the critique of the dimension labels and their definitions. To provide some background information, let me summarize the results of my investigation.

A. The purpose of my investigation is to develop a behaviorally anchored rating scale for the assessment of teaching behaviors in secondary physical education.

B. In the first step of the investigation, classes of secondary physical education students supplied statements of their observations of teaching behaviors.

C. The statements were edited, refined, or eliminated by the investigator and a panel of five secondary students.

D. Another independent group of students rated the remaining statements for accuracy in describing effective teaching behaviors. Cronbach's alpha analysis was used to analyze the responses of the students. The statements which depressed the alpha score were eliminated (135 statements retained).

E. The students who rated the statements also rated a set of categories obtained from the literature. ANOVA was used to identify sources of variance and eliminate categories where the variance was significant (10 categories retained).

F. Another independent group of secondary students assigned each of the 135 behavioral statements to one of ten categories (obtained

in step E). However, the students were not able to agree upon the allocation of behavioral statements to categories.

Since the students were not able to agree upon the placement of the behavioral statements into categories, I am using a different process by which to identify categories. The directions for identifying the categories are attached to this cover letter. When you complete your task, would you please return the results to Dr. Sarah Robinson (School of HPERD) by Monday, October 22.

Thank you for your time and consideration. I look forward to your responses and critique.

Respectfully,

Robert N. McKethan

SORTING INSTRUCTIONS

Introduction

From the cover letter (step F), you should understand that the students were unable to agree on the placement of the enclosed behavioral statements into categories. After discussions with my committee, it was ascertained that there were problems with overlapping categories and category labels.

Your task is to sort each of the 135 behavioral statements into categories. Understand that I have presorted these statements into cateogies. My purpose in having you sort these behaviors into categories is to expose and eliminate any biases to which I may be predisposed (as a result of my familiarity with student responses [step F in cover latter]). This procedure represents an alternative way of identifying and defining categories.

Instructions

A. You will be working with the statements and worksheets found these instructions.

B. Sort the statemenets into categories. All statements which are placed into a particular group should possess some common aspect of teacher behavior.

C. Once you have placed the statements into groups, try to fit each group of statements to one of the 7 categories identified by my presorting process (there is one worksheet for each category).

1. After matching your cluster of statements to a category, use the worksheet and record the statement numbers in the appropriate spaces.
2. Examine the category label to see if it properly, in your estimation, reflects the statements.
3. Critique the category label in the appropriate space.
4. Examine the category definition to see if it comprehansively depicts the category.
5. Critique the category definition in the appropriate space.

D. If one you your clusters of statements does not fit the categories presented, please identify its label and provide a new definition. Do this on the worksheet entitled "New Category".

E. Upon completion of this task, please insert the worksheets in the enclosed envelope and return to Dr. Sarah Robinson (School of HPERD) no later than October 22.

WORK SHEET

COMMUNICATION

Definition

A. Teacher actions which give information to students about the schedule of physical education activities.

B. Teacher actions which show a response or lack of response to student questions.

C. Actions which show the manner in which the teacher speaks to the class.

List Statement Numbers

___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___

Category Label Criticism

Category Definition Criticism

WORK SHEET

MANAGERIAL

Definition

A. Teacher actions which show how time is used in the physical education class.

B. Teacher actions which show the organization of class activities.

C. Teacher decisions about the roles of students during class.

List Statement Numbers

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

Category Label Criticism

Category Definition Criticism

WORK SHEET

GRADING

Definition

A. Teacher actions which show how students will be graded in physical education.

B. Teacher actions which show the procedures used in assigning grades.

List Statement Numbers

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

Category Label Criticism

Category Definition Criticism

WORK SHEET

DISCIPLINE

Definition

A. Any teacher action directed toward settling disagreements among students.

B. Any teacher action directed toward inappropriate student behavior.

C. Teacher actions that have the purpose of promoting conformity to class rules.

List Statement Numbers

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

Category Label Criticism

Category Definition Criticism

WORK SHEET

INSTRUCTION

Definition

A. Teacher or student demonstrations of game rules, skills, strategies, and warmups.

B. Teacher explanations of game rules, skills, strategies, and warm-ups.

C. The oral presentation of notes by the teacher.

List Statement Numbers

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

Category Label Criticism

Category Definition Criticism

WORK SHEET

TEACHER SENSITIVITY

Definition

A. Teacher actions that show (or doesn't show) recognition of individual differences in students.

B. Teacher actions that show (or doesn't show) recognition and acceptance of student emotions and feelings.

C. Teacher actions that show concern (or lack of) for student welfare.

D. Teacher actions that show attempts in relating to students.

List Statement Numbers

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

Category Label Criticism

Category Definition Criticism

WORK SHEET

ENTHUSIASM

Definition

A. Actions which communicate the teacher's level of excitement about the physical educaton program.

B. Actions that show the level of involvement in the physical education program by the teacher.

C. Actions which might encourage or discourage students in participating in class activities.

List Statement Numbers

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

___ ___ ___ ___ ___ ___ ___ ___ ___

Category Label Criticism

Category Definition Criticism

APPENDIX H

INSTRUMENT FOR DATA COLLECTION

Allocation of Statements and Assignment of Values Alternative Procedures

ASSIGNMENT OF NUMERICAL VALUES

Introduction

The purpose of this assignment is for you to decide whether or not the listed behaviors are effective, neutral, or ineffective. Please understand that you will not be rating your teacher's behaviors.

Below are 5 general P.E. program objectives. Please read them and keep them in mind when rating the following 34(33) behaviors.

PROGRAM OBJECTIVES

Introduction

Program objectives are statements which show how the physical education program will benefit students. Ideally, the teaching behaviors of physical education teachers should help these objectives come about.

You will use the following physical education program objectives to help you decide whether a behavior is very effective, neutral, or very ineffective.

Objectives

A. To provide experiences that will promote physical fitness.

B. To provide opportunities for skill development in life-time sport.

C. To provide experiences that will promote group interaction and communication.

D. To provide experiences that will promote a positive self image.

E. To provide experiences that promote emotional stability.

Instructions

After you read the statement, study the program objectives. If the actions found in the statement help to accomplish one of the objectives, then the behavior is effective. If the actions interfered with accomplishing of

an objective, the behavior is ineffective. If the actions do not really help or interfere in the accomplishing of an objective, the behavior is neutral.

A "1" is for very effective behaviors. The numbers, "2" and "3" are for effective behaviors. Four (4) is for neutral behaviors. The numbers "5" and "6" are for ineffective behaviors. Seven (7) is for very ineffective behaviors. After deciding the effectiveness of the behavior, place a check in the space above the number of your choice.

Example

1. The teacher calls us by our first names.

very effective neutral very ineffective

____:____:____:____:____:____:____
1 2 3 4 5 6 7

The number "1" was selected because the behavior is an extremely effective behavior. The behavior provides an experience that promotes positive feelings about oneself (Objective D) or promotes communication (Objective C).

Complete the informaton found below by circling the response that best applies to you.

A. Your class: Sophomore Junior Senior
B. Your ethnic background: White Indian Hispanic Asian Black
C. Your sex: Male Female
D. Your P.E. Grade (1st nine weeks) A B C D E

STATEMENTS

1. The teacher has the class take notes on the rules of tennis.

very effective neutral very ineffective

____:____:____:____:____:____:____

1 2 3 4 5 6 7

2. The teacher turns a negative response by a student in to a positive situation by using positive comments.

very effective neutral very ineffective

____:____:____:____:____:____:____

1 2 3 4 5 6 7

3. The teacher lifts weights with the students during class.

very effective neutral very ineffective

____:____:____:____:____:____:____

1 2 3 4 5 6 7

4. The teacher provides free time during part of the class.

very effective neutral very ineffective

____:____:____:____:____:____:____

1 2 3 4 5 6 7

5. The teacher usually gives us time if we need to talk about a problem.

very effective neutral very ineffective

____:____:____:____:____:____:____

1 2 3 4 5 6 7

6. The teacher is always patient with us if we are doing something wrong.

very effective neutral very ineffective

____:____:____:____:____:____:____

1 2 3 4 5 6 7

Note The remaining statements are found in Appendix J.

ASSIGNMENT OF BEHAVIORAL STATEMENTS

Introduction

The purpose of your task is for you to sort statements describing teaching bheaviors of P.E. teachers into categories. Please understand that you are not rating your physical education teacher. Below are 7 categories, A through G. Please read each thoroughly.

CATEGORIES

Introduction

Categories of behaviors include groups of behaviors that are similar in some way. The categories into which you will be putting behavior statements are identified and defined below.

A. Grading Teacher grading are behaviors that teachers use in grading students. Behaviors in this category include teacher actions which show how students will be graded in physical education and teacher actions which show the procedures used in assigning grades.

B. Managerial Teacher management are behaviors which guide and direct the physical education program. Behaviors in this category include teacher actions which (a) show how time is used in the class, (b) show the organizing of class activities, and (c) show decisions about the roles of students during class.

C. Communication Teacher communication behaviors are verbal and nonverbal interactions with students. Behaviors in this category include teacher actions (a) which give information to students about the schedule of physical education activities, (b) that show a response of lack of response to student questions, and (c) which show the ways in which the teacher speaks to the class.

D. Enthusiasm Teacher enthusiasm are behaviors which show interest in students and the P.E. program. Behaviors in this category include teacher actions which (a) communicate the teacher's level of excitement about the physical education program, (b) shows the level of involvement in the physical education program by the teacher, and (c) might encourage or discourage students in participating in class activities.

E. Teacher Sensitivity Teacher sensitivity behaviors show the teacher's ability to feel or sense student feelings, emotions, and attitudes. Behaviors in this category include teacher actions that show (or doesn't show) (a) recognition of individual differences in students, (b) recognition and acceptance of student emotions and feelings, (c) concern for student welfare, and (d) attempts in relating to students.

F. Instruction Teacher instruction are behaviors which inform students about skills, games, and warm-ups. Behaviors in this category include teacher actions that show (a) explanations and demonstrations of game rules, skills, straegies, and warm-ups, (b) the oral prsentation of notes by the teacher, and (c) use of students in demonstrations.

G. Discipline Teacher discipline are behaviors that attempt to keep a class atmosphere which promotes good use of time, safety, and sportsmanship. Behaviors in this category include teacher actions (a) directed toward settling disagreements among students, (b) directed toward inappropriate student behavior, and (c) that have the purpose in promoting conformity to class rules.

Instructions

Notice that there is a letter of the alphabet which goes with each category. After you read the statement, study the teacher behavior categories. Decide the category to which the statement belongs and put a check above the letter of your choice.

Example

1. The teacher calls us by our first names.

A	B	C	D	E	F	G
		✓				

The letter "C" was selected because the behavior best fits the communication category. This behavior shows the way the teacher speaks to the class.

Complete the informaton found below by circling the response that best applies to you.

A. Your class: Sophomore Junior Senior
B. Your ethnic background: White Indian Hispanic Asian Black
C. Your sex: Male Female
D. Your P.E. Grade (1st nine weeks) A B C D E

STATEMENTS

1. The teacher has the class take notes on the rules of tennis.

A	B	C	D	E	F	G

2. The teacher turns a negative response by a student in to a positive situation by using positive comments.

A	B	C	D	E	F	G

Note The remaining statements are found in Appendix J.

APPENDIX I

INSTRUMENT FOR DATA COLLECTION

Behaviorally Anchored Rating Scale for The Evaluation of Secondary Physical Education Teachers

1. Question: Who are the raters?

   Answer: Secondary P.E. students will be the raters.

2. Question: Who will you be rating.

   Answer: You will be rating your physical education teacher (The one to whom you are currently assigned for class activity). You will be rating categories of your teacher's actions that are important to teaching physical education.

3. Question: Why should secondary students rate their teachers?

   Answer: In the North Carolina public schools, students are not usually called upon to rate their teachers. However, students are an important source of information about teachers. First, students see their teachers more than principals or supervisors. Secondly, including students in the rating of teaching performance increases the points of view in the iformation used to evaluate teachers. Finally, students can provide insightful information that cannot be provided by others.

   Also, the information that your class and other classes provide about your teacher will be used to determine the usefulness of these scales.

4. Question: If I respond honestly to these scales, will I get into trouble?

   Answer: No! Your responses will be kept anonymous. There is no place on the scales to sign your name.

5. Question: Will my responses to this rating scale effect how my teacher teaches?

   Answer: Teachers are sensitive to the needs and perceptions of students. Research indicates that teachers do respond positively to student ratings.

6. Question: What is a rating scale?

   Answer: A rating scale is a measuring instrument that requires the rater (student) to assign the rated object (teacher behavior) to a number on a continua. Rating scales use numbers and/or adjectives to help the rater make assignments. You will respond to two kinds of rating scales. One is like the kind just described and the other is a set of behaviorally anchored rating scales.

7. Question: What is a behaviorally anchored rating scale?

   Answer: A behaviorally anchored rating scale is a rating scale that has examples of behaviors in addition to numbers or adjectives. These behaviors identified by students from two different school systems.

8. Question: What are the advantages of using a behaviorally anchored rating scale?

   Answer: The examples of behaviors help the rater to make decisions about the object being rated (teacher behaviors). The behaviors found on the scales clearly show examples of effective, neutral, and ineffective behaviors. Thus, the behavioral examples on the scales will help you decide the effectiveness of your teacher's behaviors.

Part II

OVERALL RATINGS OF CATEGORIES

Introduction

The rating scales appearing below are overall rating scales. With each rating scale is a category of teaching behaviors and its definiton.

Your task is to decide how effective your teacher's behavior is for each category. On each scale place a check over the number of your choice to rate your teacher's usual behaviors.

Overall rating scales

Grading

Teacher grading actions are behaviors that teachers use in grading students.

My teacher is:

extremely effective neutral extremely ineffective

1 : 2 : 3 : 4 : 5 : 6 : 7

Managerial

Teacher management activities are behaviors which guide and direct the physical education program.

My teacher is:

extremely ineffective neutral extremely effective

1 : 2 : 3 : 4 : 5 : 6 : 7

Teacher Sensitivity

Teacher sensitivity behaviors show the teacher's ability to feel or sense student feelings, emotions, and attitudes.

My teacher is:

extremely ineffective neutral extremely effective

7 : 6 : 5 : 4 : 3 : 2 : 1

Instruction

Teacher instructional activities are behaviors which inform students about skills, games, and warm-ups.

My teacher is:

extremely effective neutral extremely ineffective

1 : 2 : 3 : 4 : 5 : 6 : 7

Discipline

Teacher discipline actions are behaviors that attempt to keep a class atmosphere which helps to bring about good use of time, safety, and sportsmanship.

My teacher is:

extremely effective neutral extremely ineffective

7 : 6 : 5 : 4 : 3 : 2 : 1

Part III
BARSSPET

Student Information

Answer the questions found below by circling the response that best applies to you.

1. Your high school class?. Sophomore Junior Senior
2. Your ethnic background? White Black Indian Hispanic Asian
3. Your sex? Male Female
4. Your first Semester grade? A B C D E
5. Your teacher's name? A C M P

Directions

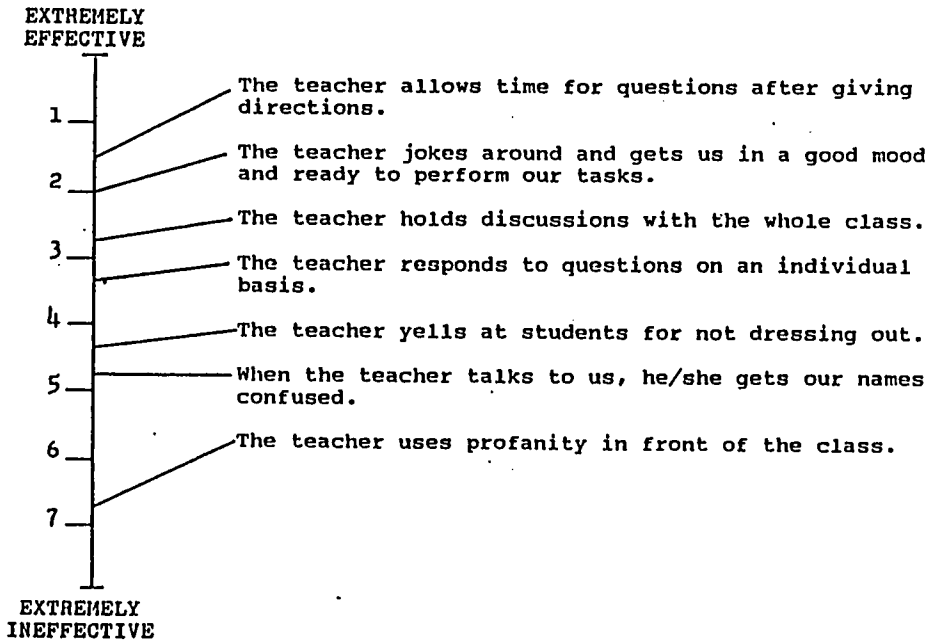
Here are the directions for completing the rating scales found on the next five pages. Please read each step carefully and study the example.

1. Review the directions at the top of each page.
2. Study the definition that appears above each scale.
3. Read the behaviors that are found on one side of the scale. Remember that you may or may not have seen these particular behaviors in your teacher's actions. Since you see your teacher each school day, you should be very familiar with your teacher's instructional (in class) behaviors. You are to decide which of the behaviors on each scale that you think would be most like examples of your teacher's actions.
4. It is important to remember that your teacher's behaviors may be more effective or less effective than the behaviors on the scales
5. After you study each behavior on the scale and decide which one would be most like examples of your teacher's actions, circle the number on the scale nearest the behavior you selected. If you think that your teacher's behavior is more or less effective than the behaviors on the scale, then circle the number above or below the most effective or least effective behavior on the scale.

Example

COMMUNICATION

Definition Teacher communication includes verbal and nonverbal interactions with students.

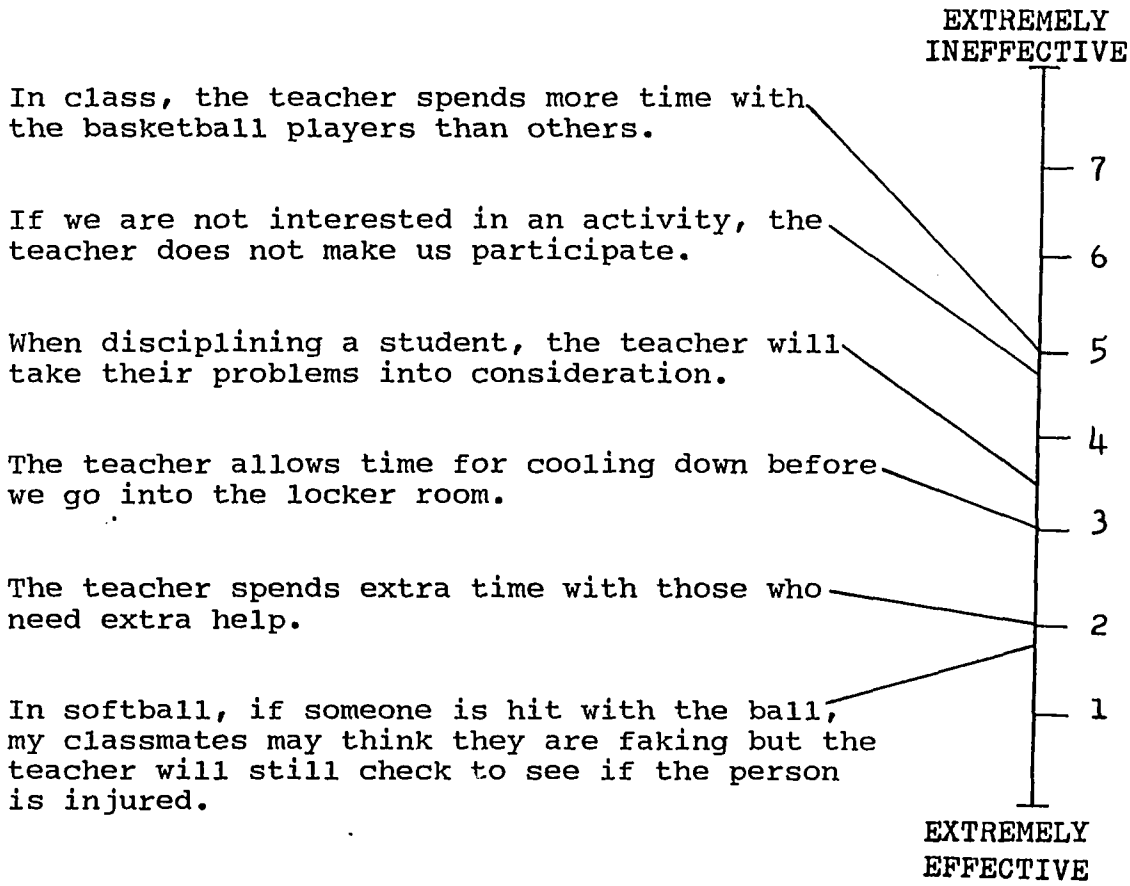


TEACHER SENSITIVITY

Directions

1. Read the definition of the Teacher Sensitivity category.
2. Read all statements appearing on the scale below the definition.
3. Decide which statement is most like examples of your teacher's behaviors. Remember, examples of your teacher's behaviors may be more effective or less effective than the behaviors on the scale.
4. Consider which number on the scale best represents examples of your teacher's behaviors for this category.
5. Circle the number of your choice. Circle only one number.

Definition Teacher sensitivity behaviors show the teacher's ability to feel or sense student feelings, emotions, and attitudes. Behaviors in this category include teacher actions that show [or doesn't show] (a) recognition of individual differences in students, (b) recognition and acceptance of student emotions and feelings, (c) concern for student welfare, and (d) attempts in relating to students.

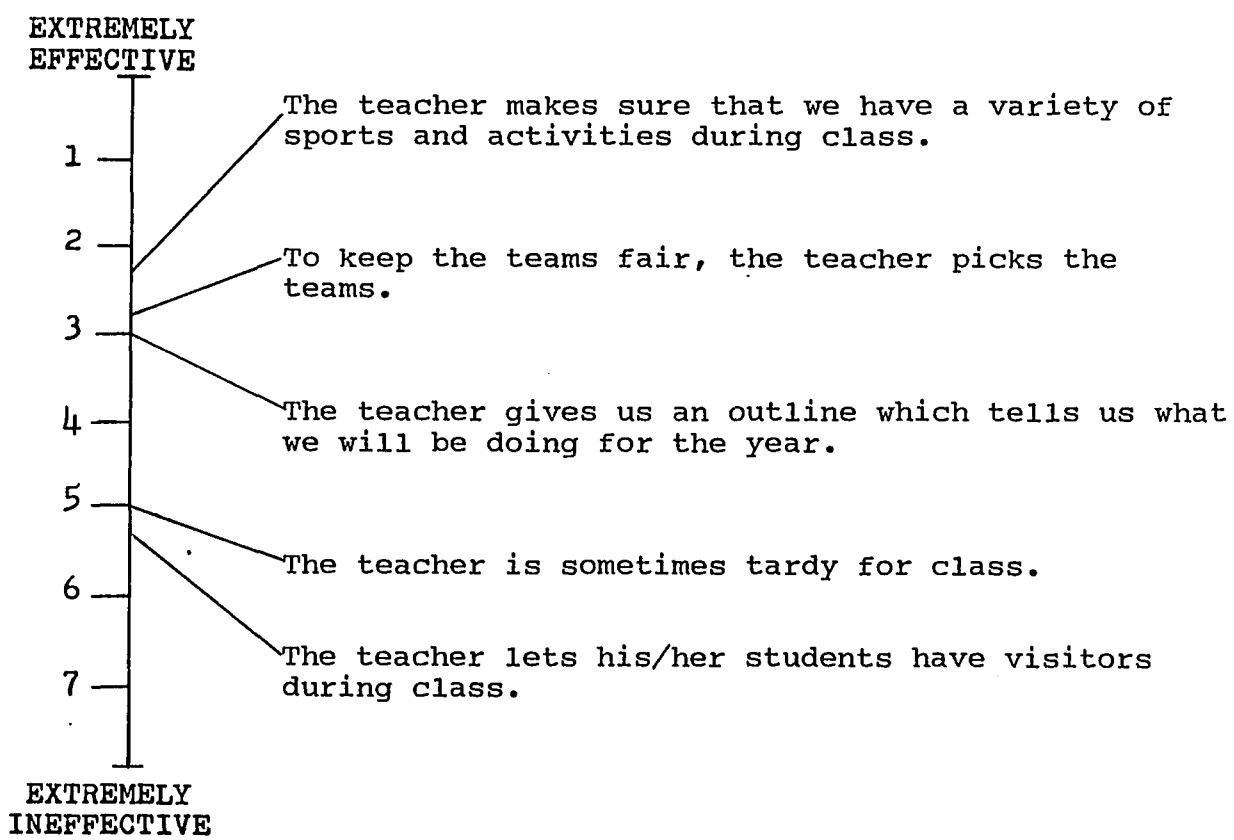


MANAGERIAL

Directions

1. Read the definition of the Managerial category.
2. Read all statements appearing on the scale below the definition.
3. Decide which statement is most like examples of your teacher's behaviors. Remember, examples of your teacher's behaviors may be more effective or less effective than the behaviors on the scale.
4. Consider which number on the scale best represents examples of your teacher's behaviors for this category.
5. Circle the number of your choice. Circle only one number.

Managerial Teacher management actions are behaviors which guide and direct the physical education program. Behaviors in this category include teacher actions which (a) show how time is used in the class, (b) show the organizing of class activities, and (c) show decisions about the roles of students during class.

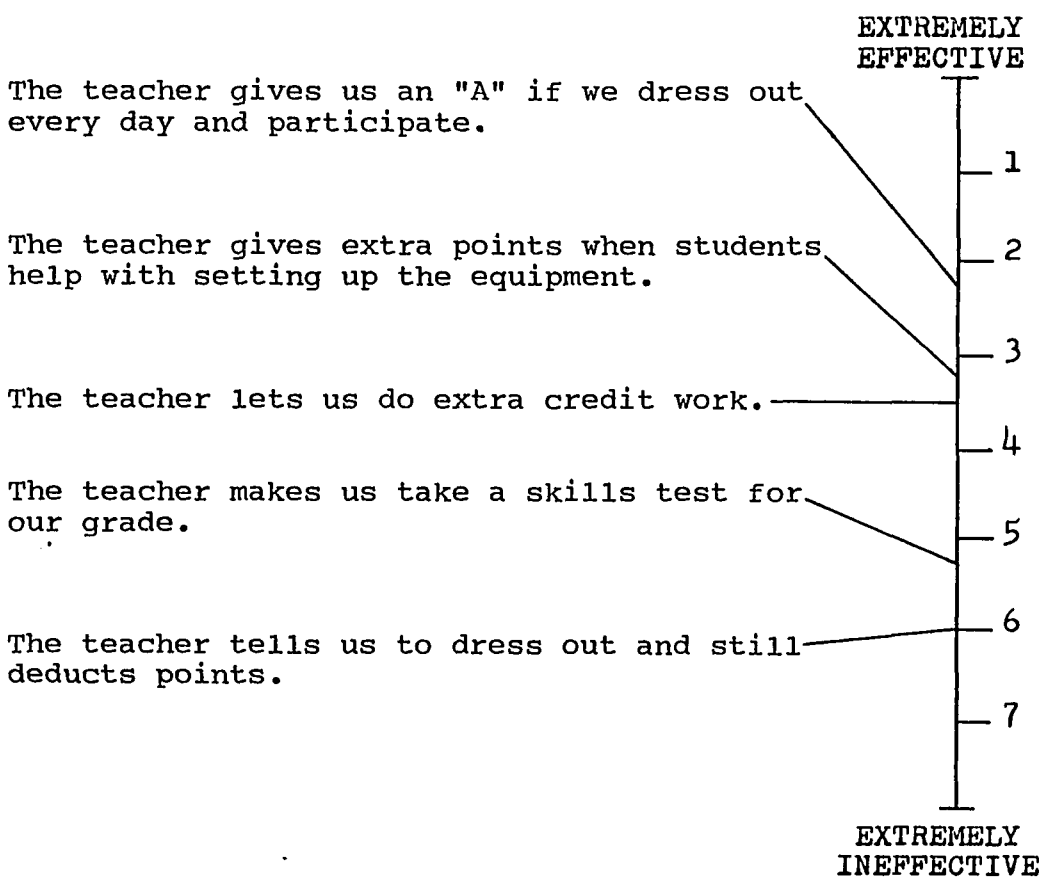


GRADING

Directions

1. Read the definition of the Grading category.
2. Read all statements appearing on the scale below the definition.
3. Decide which statement is most like examples of your teacher's behaviors. Remember, examples of your teacher's behaviors may be more effective or less effective than the behaviors on the scale.
4. Consider which number on the scale best represents examples of your teacher's behaviors for this category.
5. Circle the number of your choice. Circle only one number.

Definition Teacher grading actions are behaviors that teachers use in grading students. Behaviors in this category include teacher actions which show how students will be graded in physical education and teacher actions which show the procedures used in assigning grades.

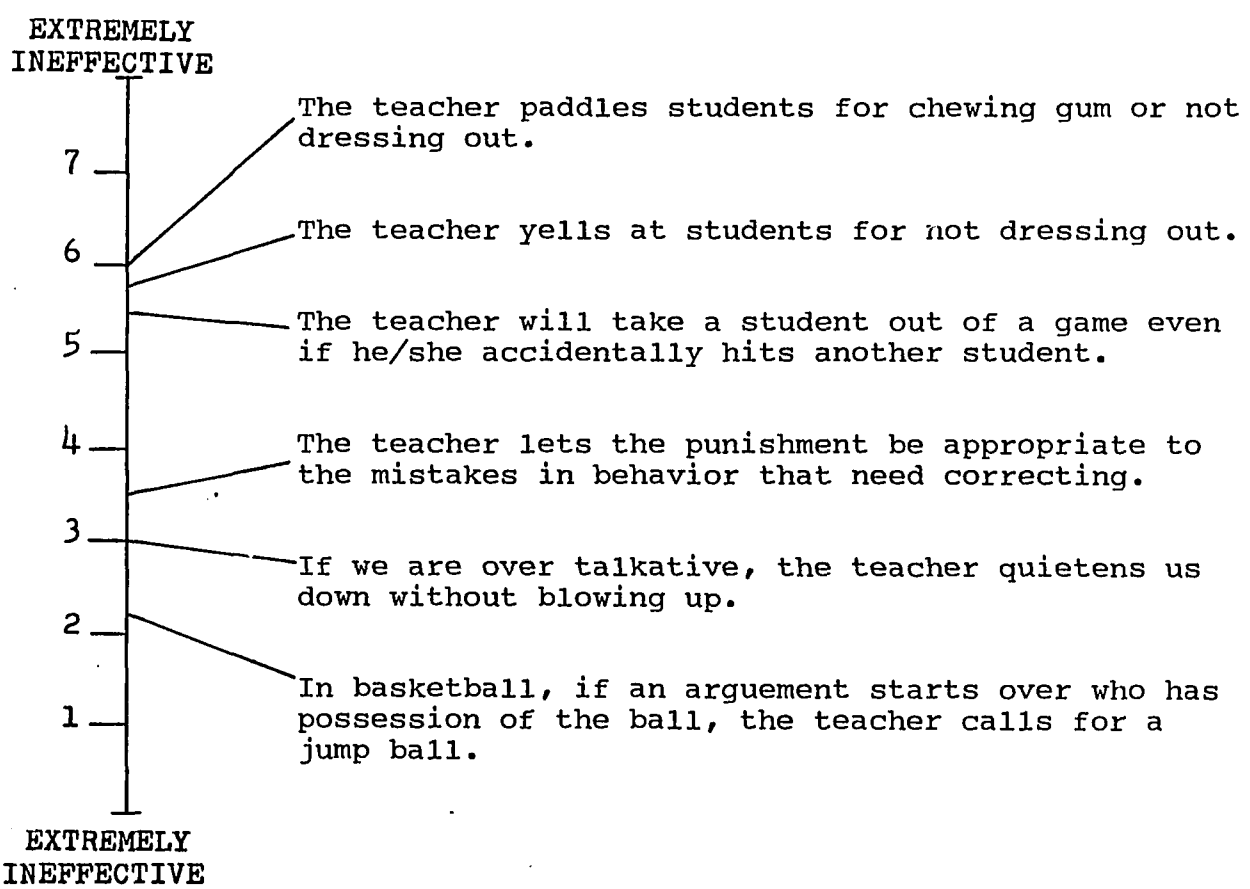


DISCIPLINE

Directions

1. Read the definition of the Discipline category.
2. Read all statements appearing on the scale below the definition.
3. Decide which statement is most like examples of your teacher's behaviors. Remember, examples of your teacher's behaviors may be more effective or less effective than the behaviors on the scale.
4. Consider which number on the scale best represents examples of your teacher's behaviors for this category.
5. Circle the number of your choice. Circle only one number.

Definition Teacher discipline actions are behaviors that attempt to keep a class atmosphere which brings about good use of time, safety, and sportsmanship. Behaviors in this category include teacher actions (a) directed toward settling disagreements among students, (b) directed toward inappropriate student behavior, and (c) that have the purpose in promoting conformity to class rules.

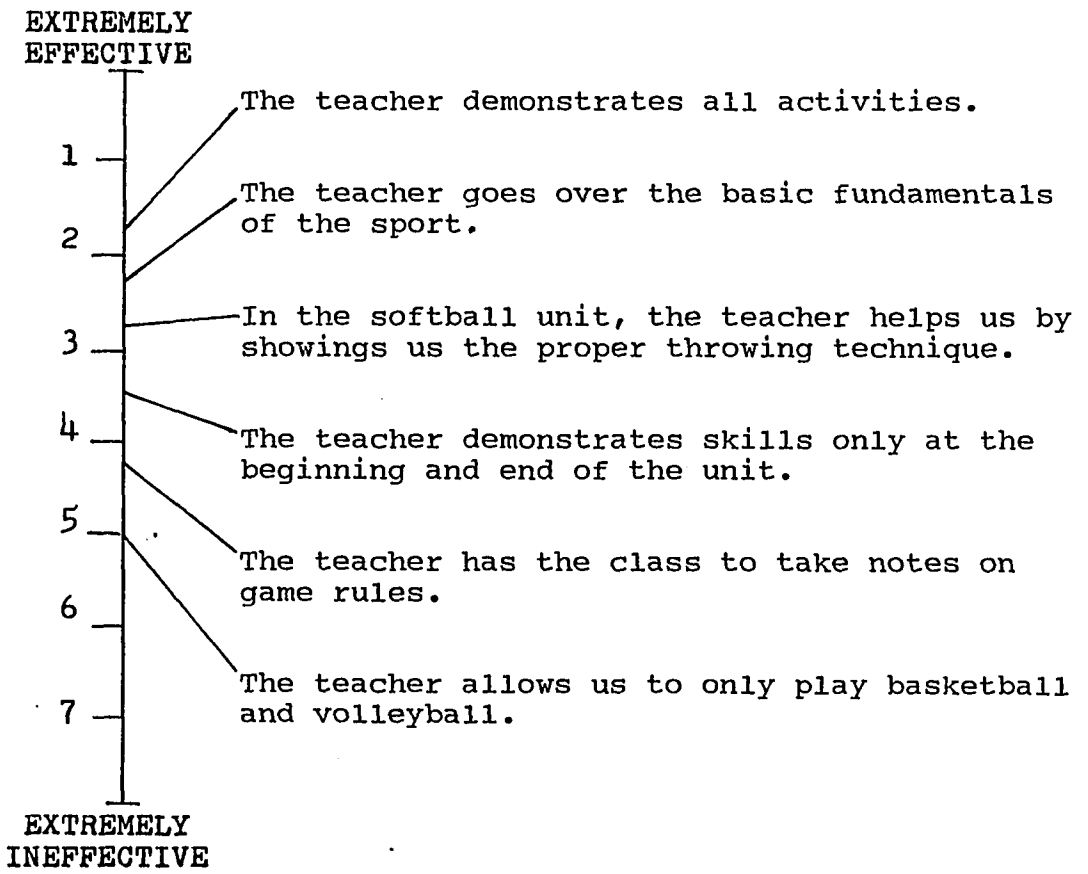


INSTRUCTION

Directions

1. Read the definition of the Instruction category.
2. Read all statements appearing on the scale below the definition.
3. Decide which statement is most like examples of your teacher's behaviors. Remember, examples of your teacher's behaviors may be more effective or less effective than the behaviors on the scale.
4. Consider which number on the scale best represents examples of your teacher's behaviors for this category.
5. Circle the number of your choice. Circle only one number.

Definition Teacher instructional activities are behaviors which inform students about skills, games, and warm-ups. Behaviors in this category include teacher actions that show (a) explanations and demonstrations of game rules, skills, strategies, and warm-ups, (b) the oral presentation of notes by the teacher, and (c) use of students in demonstrations.



APPENDIX J

BEHAVIORAL STATEMENTS

Statements Retained for Allocation and Value Assignment

STATEMENTS

1. The teacher has the class take notes on the rules of tennis.

2. The teacher turns a negative response by a student in to a positive situation by using positive comments.

3. The teacher lifts weights with the students during class.

4. The teacher provides free time during part of the class.

5. The teacher usually gives us time if we need to talk about a problem.

6. The teacher is always patient with us if we are doing something wrong.

7. To keep the teams fair, the teacher picks the teams.

8. In the softball unit, the teacher helped us by showing us the proper throwing technique.

9. The teacher explains the different muscle groups involved in exercises and the reasons for the exercises.

10. The teacher will give examples of what we will be doing on a certain day.

11. The teacher gives us advice to help us.

12. The teacher helps to learn independently by showing us how to help each other.

13. Sometimes the teacher will have us sit down in a group and just talk.

14. The teacher talks with small groups or individuals about discipline.

15. The teacher lets us do extra credit work.

16. At the end of class, the teacher will tell us what we will be doing tomorrow.

17. The teacher shows us the proper way to do our warm-ups.

18. The teacher gives a good explanation of what is to be done in the class for the day.

19. The teacher sets up the equipment before the beginning of class.

20. When we are injured at the end of the semester before we change teachers, our present teacher will tell our next teacher about our injury.

21. In some of our units, the teacher will give us a choice of activities.

22. The teacher allows time for questions after giving directions.

23. The teacher goes over the information that will be on the test and then asks us for questions.

24. The teacher helps those in a bad mood to smile.

25. If we get discouraged and want to quit the activity, the teacher will say, "Oh, come on! You might surprise yourself!"

26. In basketball, if an arguement starts over who has possession of the ball, the teacher calls for a jump ball.

27. When starting a new workout, the teacher gives an explanation and asks for questions.

28. In wrestling, if we get tired or fatigued, the teacher will tell us to get off the mat and call us a "fish". (A "fish" is someone who doesn't know his wrestling moves)

29. After the warm-ups, the teacher gives us time to stretch individually.

30. When disciplining a student, the teacher will take their problems into consideration.

31. The teacher explains what he/she expects for the entire nine weeks.

32 After a hard game, the teacher gives us a chance to rest.

33. The teacher allows us time after school to make up any work missed.

34. In the track unit, our teacher always says, "In track you've got to hurt."

35. The teacher admits to having faults similar to our own.

36. In assigning our daily grade, the teacher takes into consideration if we are sick.

37. The teacher admits to being out of shape.

38. The teacher stays in his/her office and does not explain what to do.

39. The teacher lets his/her students have visitors during class.

40. The teacher lets volunteers demonstrate skills and activities.

41. When a student can not perform a certain routine in tumbling, the teacher works with the student until they are successful.

42. The teacher lets us choose our own teams which allows us to avoid the embarrassment of being placed on a certain team.

43. The teacher yells at a student for not dressing out.

44. The teacher has us to do a variety of new warm-ups.

45. The teacher responds to questions on an individual basis.

46. If a person gets hurt on a balance beam and doesn't want to get back on, the teacher will yell at the student for not wanting to get back on.

47. The teacher plays the radio while we do our warm-ups.

48. The teacher allows him/herself to get in an arguement with a student after the student refuses to follow directions.

49. The teacher knocks over a desk when he/she gets angry.

50. The teacher paddles students for chewing gum or not dressing out.

51. The teacher will tell a student that he/she is "useless".

52. The teacher tells us the exercises to do and then demonstrates them if they are new.

53. The teacher tells us ways we can make learning easier by telling us the most important facts or by showing us the most important parts of a skill.

54. The teacher will not allow students to go wild in class.

55. The teacher makes sure that we have a variety of sports and activities during class.

56. The teacher makes us run outside when it is raining.

57. The teacher will give us all of the information that will be found on the exam.

58. The teacher makes the students take turns leading warm-up exercises.

59. The teacher does not yell, but comforts both when one student accidentally hurts another student.

60. The teacher gives out first and second place certificates to promote total competition.

61. The teacher gives out an up-to-date study guide that includes a history on each sport.

62. The teacher works with each student according to his or her skill level.

63. The teacher puts peer pressure on students with performance problems.

64. During games in the basketball unit, the teacher will cheer a good play.

65. In softball if someone is hit with the ball, my classmates may think that they are faking, but the teacher will still check to see if the person hit is injured.

66. The teacher watches us play volleyball to see who cheats.

67. The teacher explains the techniques before demonstrating them.

68. The teacher requires us to bring current events for our grade.

69. The teacher makes us throw a football for our grade.

70. The teacher doesn't make us do as much as other students after recovering from an injury.

71. The teacher picks teams instead of allowing the students to pick the teams.

72. The teacher shows consideration of those who can not play a sport by offering to help them improve.

73. The teacher will take a student out of a game even if the student hits another student accidentaly.

74. In softball, the teacher tries to help student, who is having trouble batting, how to hold the bat.

75. The teacher gives extra points when students help with setting up the equipment.

76. The teacher lets students decide on what warm-ups to do.

77. The teacher lets the students decide who will lead the warm-ups.

78. The teacher demonstrates skills only at the beginning and end of the unit.

79. The teacher requires us to remove our jewelry for classes.

80. The teacher does not lock the locker room door which results in things being taken.

81. The teacher gives us 10 minutes to dress and undress.

82. Sometimes the teacher will paddle someone for something they did not do.

83. In softball the teacher will look at someone very cruelly if they miss fielding the ball.

84. In softball, the teacher has the boys to play the girls.

85. The teacher jokes around and gets us in a good mood and ready to perform our tasks.

86. When someone falls off the balance beam and begins to cry, the teacher will go over and try to comfort them.

87. The teacher participates with the aerobics class.

88. The teacher is sometimes tardy for class.

89. The teacher speaks loudly and clearly.

90. The teacher takes time out to help in things we don't understand.

91. The teacher leads warm-ups along with the students.

92. The teacher demonstrates all activities.

93. If we are not interested in an activity, the teacher doesn't make us participate.

94. The teacher spends extra time with those who need extra help.

95. The teacher and the students together choose what the class is going to do.

96. The teacher gives us an outline which tells us what we will be doing for the year.

97. The teacher goes over the basic fundamentals of the sport.

98. The teacher makes students run in extremely cold weather.

99. The teacher corrects students, in a respectful manner, in front of the class.

100. The teacher gives us a total grade on how well we try and participate when we are given the physical fitness test.

101. The teacher treats each student as an individual with a different personality and emotions.

102. The teacher lets punishment be appropriate to the mistakes in behavior that need correcting.

103. The teacher allows time for cooling down before we go into the locker room after class.

104. The teacher gives individual attention for students with learning disabilities.

105. When running laps, the teacher runs with the entire class instead of sitting in the bleachers.

106. If we are over talkative, the teacher quietens us down without blowing up.

107. The teacher only allows to play basketball and volleyball.

108. The teacher tells us to dress out and still deducts points.

109. The teacher gives us an "A" if we dress out every day and participate.

110. Sometimes when the teacher has not come out, he/she lets a student lead the warm-ups.

111. The teacher becomes part of the activity to encourage studemts to participate.

112. The teacher gives only a few notes to study by.

113. The teacher makes us do all the exercises until we give the signal,"hit-hit-one".

114. While we sit on the floor for roll to be checked, the teacher is in the bleachers talking to someone.

115. Even if students are not able to get along with one another and it doesn't mess up the class, the teacher tries to help students to get along with one another.

116. The teacher makes me feel better about myself and the class by saying that I can run the mile in the required amount of time.

117. The teacher lets students pick their own teams for the activity.

118. The teacher rushes students on tests.

119. The teacher listens to student comments.

120. The teacher holds discussions with the entire class.

121. The teacher lets the students check the roll.

122. In class, the teacher spends more time with the basketball players than others.

123. The teacher does not dress out for class.

124. The teacher doesn't provide enough equipment which results in part of the students sitting around while others play the game.

125. In volleyball or basketball, if a team gets upset over someone making a mistake, the teacher will take the person who make the mistake out of the activity with no explanation.

126. The teacher uses profanity in front of the class.

127. The teacher does not allow us to talk that much.

128. In softball, the teacher gets angry at students if they do something wrong.

129. The teacher participates with us in volleyball.

130. The teacher wastes class time by calling roll.

131. In softball, the teacher pitches for both teams.

132. The teacher checks to make sure students are taking showers.

133. The teacher lets us vote on what to play.

134. The teacher makes us run four laps; if we walk, he/she yells at us.

135. The teacher makes sure that all people get a chance to ask questions.

APPENDIX K

STATISTICS FROM GLOBAL RATINGS

Means, Standard Deviations, Score Ranges

Interdimension Correlations

Medians and Ranges of Inter-Dimesnion Correlations of Global Scales

Dimensions	Median Correlation	Range
	School 1	
Teacher Sensitivity	.49	.33 - .57
Managerial	.45	.36 - .56
Grading	.26	.15 - .39
Discipline	.39	.15 - .62
Instruction	.54	.19 - .62
	School 2	
Teacher Sensitivity	.12	-.11 - .27
Managerial	.30	.25 - .41
Grading	.24	.01 - .41
Discipline	.27	.20 - .40
Instruction	.26	-.11 - .40

Overall Ratings
Global Rating Scales

School 1				School 2			
Dim	$\bar{X}$	SD	Range	Dim	$\bar{X}$	SD	Range
Teas	3.45	1.43	1 - 7	Teas	2.72	1.06	1 - 6
Mana	3.29	1.39	1 - 7	Mana	2.93	1.48	1 - 7
Grad	3.75	1.78	2 - 7	Grad	3.45	1.83	1 - 7
Disc	2.95	1.57	1 - 7	Disc	2.47	1.48	1 - 7
Inst	3.18	1.80	1 - 7	Inst	2.27	1.38	1 - 7

APPENDIX L

PROTOCOL FOR ADMINISTRATION OF THE BARSSPET

PROTOCOL OF ADMINISTRATION OF THE BARSSPET

Introduction

The following directions for administration of the BARSSPET are based upon the procedures used in collecting data for construction of the BARSSPET and its administration. The directions include recommendations regarding class size, testing areas, and procedures. These recommendations are based upon its administration by a single individual.

Class Size

The size of any group responding to the BARSSPET should be no larger than 30-35 students. An optimum sized group permits the test administrator to:

1. Quickly settle the group prior to instructions
2. Adequately respond to directions in a manner that all participants can understand.
3. Monitor students as they are responding to the BARSSPET.

Testing Location

The testing area should be located in a regular classroom. Gym floors or bleachers do not provide an adequate surface for any activity where written responses are required. Acoustical qualities of the gymnasium are distracting and prohibitive.

Before Administration

Before students enter the class, have all test materials in place on the desks (tables). Instruct students to leave the pencils on the desk.

Administration Procedures

Part I. Instruct students to read Part I silently, as Part I is read orally. This procedure may aid in reducing apprehension on the part of poor readers. After orally reading Part I, allow students time to ask questions.

Part II. Part II is a global rating of five teacher behavior categories. The directions for Part II are listed below.

1. Identify and define the following vocabulary words. These words should include but not be limited to "response," "extremely," "effective," "neutral," and "ineffective".
2. Brief instruction on rating errors should now be given to the students. Content of the instruction should address the following errors.
   A. Central tendency error -- reluctance of students to use the whole scale.
   B. Halo error -- unwillingness to differentiate among job elements.
   C. Leniency error -- tendency to use only higher ratings.
3. Read the introduction orally as the students are reading silently. Prior to allowing students to respond to the global scales ask the students to:
   A. Raise their hands silently if they have questions.
   B. To place their pencil down and not to turn to the next page when finishing Part II.

Part III. Part III includes four parts: (a) Student information, (b) directions, (c) example scale, and (d) five behaviorally anchored rating scales. Directions are detailed below.

1. Ask students to respond to the student information by circling the appropriate response.
2. As you read the directions orally, ask the student to read along silently. After the directions are completed, allow time for student questions.
3. Orally review the example below the directions. It is recommended that an overhead projector be used in working through this example.
4. Allow additional time for questions.Before giving students the starting signal, tell students to raise their hands for help after they start. Tell students to remain seated and quiet until all students complete their responses.

Time Requirement

Even when using a careful orientation protocol, the BARSSPET can be completed comfortably by high school students in a 45 minute period.