THE DEVELOPMENT AND RELIABILITY TESTING OF A PROCEDURE TO OBSERVE AND RECORD TEACHER MOTIVATION OF PUPIL ON-TASK BEHAVIOR

CONRAD BURTON GLAS[redacted]

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

THE DEVELOPMENT AND RELIABILITY TESTING OF A PROCEDURE TO OBSERVE AND RECORD TEACHER MOTIVATION OF PUPIL ON-TASK BEHAVIOR

by

© Conrad Burton Glasgow, B.A., B.Ed.

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Education

Department of Educational Psychology
Memorial University of Newfoundland

August 1981

St. John's Newfoundland

ABSTRACT

The purpose of this study was the development and reliability testing of a classroom observation coding scale.

The observation scale focused simultaneously on the on- or off-task activity of the student and on the motivational aspect of teacher behavior. The various categories of student and teacher behavior were presented along with an outline of the training program developed for the preparation of classroom observers.

Two dimensions of the reliability problem were addressed in this study. First, eleven observers were trained in the use of the observation instrument, with the trainees given varying amounts of training time. Training was carried out using a video-tape package prepared for use with the instrument. The degree of agreement for each observer with the coding scheme for the instrument was determined using a video-taped criterion test containing samples of pupil-teacher behaviors.

The second aspect of the reliability study addressed the reliability of observations made using the coding scale. This involved an examination of the generalizability of the behavior categories used to actual classrooms. The two most highly trained observers were employed in live

observations of nine different classrooms on several occasions for each classroom. An analysis of variance of the data furnished by these observations produced generalizability coefficients for each of the categories of the observation scale.

The study concluded that the behavior categorizations employed permitted an acceptable level of criterion agreement. Generally, the pupil-focus categories produced higher agreement levels than did the teacher-focus categories. However, the data indicated that additional observer training might overcome any deficiencies in coding skills on either aspect of the instrument.

The coefficients of generalizability provided by the classroom observation data indicated that the categories of the scale are generalizable across teachers. Observations recorded using the observation instrument would therefore appear to be reliable.

ACKNOWLEDGEMENTS

The author wishes to gratefully acknowledge those persons involved in the preparation and completion of this report.

Thanks are first and foremost expressed to Dr. William Spain, supervisor of this thesis, for his advice and encouragement throughout.

Appreciation is also extended to Dr. David Watts for his contribution by way of feedback on the drafts of this report.

To the several others who assisted in this study--the teachers who permitted filming of their classrooms, the observer trainees who volunteered their time, the teachers who cooperated in classroom observation, the staffs of the Education Library and the Centre for Audio-visual Education--sincere thanks.

Finally, to a very patient family, whose support made this work possible, a very special 'thank you'.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Statement of Purpose

The purpose of this study was to develop a classroom observational coding scale which focused on the motivational aspect of teacher behavior. Coincident with the construction of this observational scale was the development of a training program to be used in the preparation of observers, and the implementation of a reliability study to determine the usability of the instrument.

Rationale for the Instrument

A model for motivation combining the theories of Maslow (1962) and Bandura and Walters (1964) provided the theoretical framework upon which the construction of this observational instrument was based. Motivation, throughout the course of this study, can be defined as any teacher behavior aimed at influencing the on-task behavior of students. The interaction taking place between teacher and student has, in recent years, been given great importance in educational research (Rosenshine, 1969;

Medley & Mitzel, 1963; Soar, 1972; Aspy, 1972). Numerous studies in this area have sought to relate the interactive process to the academic achievement of pupils, and recent research has indeed demonstrated differing teacher behaviors which can be associated with varying degrees of pupil achievement (Brophy & Good, 1970; Braun, 1976; Aspy, 1977).

Studies which have focused on the in-class behavior of students have consistently found that the level of on-task behavior is the pupil behavior most directly related to achievement. Cobb (1972), in a study of fourth grade students, found significant positive relationships between achievement and four task-oriented classroom behaviors, including attending. Several significant negative relationships were found for non-task-oriented behaviors such as non-attending. An earlier study by Perkins (1965) concluded that low achievers, as compared to high achievers, spent a significantly greater proportion of in-class time on non-task-oriented activity.

McKinney et al. (1975), using a set of twelve composite categories of classroom behaviors, studied their relationship to achievement. A composite achievement index was calculated for the 90 grade two students studied. The study concluded that children who were attentive in class and engaged in task-oriented interaction with peers were more likely to succeed academically than children who were distractible or passive in group activities.

A review by Rosenshine (1976) of research into the behavioral correlates of high and low achievers, found that most recent studies showed that both teacher-directed and peer-directed on-task behaviors were positively related to high achievement. In general, non-attending classroom behaviors were related to low achievement gain.

A further study by Fisher et al. (1978) examined the relationship between 'academic learning time' (ALT) and student learning. Once more, the research results identified a significant positive relationship between achievement and student engagement in the classroom. Similar results have been reported in studies by Soli and Devine (1976) and Lahaderne (1968).

The studies cited clearly showed that academic achievement is positively related to on-task student behavior. A previous review of related research (Keough, 1980) noted as well that these studies indicated that more on-task interaction with peers takes place among high achievers than with lower achieving students. This particular review further suggested that off-task peer interactions were given little consideration in any of the studies conducted to date, and that further research into peer-directed classroom behavior, as it relates to achievement, needs to be undertaken.

A conclusion could be made, based on this sampling of the literature, that a most important function of teaching would be to encourage the participation of the child in on-task behavior; specifically, the motivation of the child to engage in learning activities.

Much of the recent research into classroom interaction has been based on a humanistic view of education and has operated out of what may be termed a 'self-enhancement model' of motivation (Amidon & Hunter, 1967; Aspy, 1969; Wittmer & Myrick, 1974; Aspy, 1977). In the self-enhancement model, the teacher establishes a facilitative learning environment for students and then guides the learning by engaging the interest of the student in the instructional task. Following from the writing of Rogers (1951) and later Carkhuff (1967), the facilitative learning environment must be provided unconditionally so as to establish a genuine and effective relationship between the teacher and student. The general line of thought is that the provision of the facilitative climate will promote positive feelings of self within the student; this positive self-image then encourages the student to engage in on-task learning activities.

Viewing the self-enhancement model (see Figure 1) in terms of engaging the student in on-task behavior, motivation may be seen as the teacher behaviors aimed at engaging the student's involvement in the learning activity.

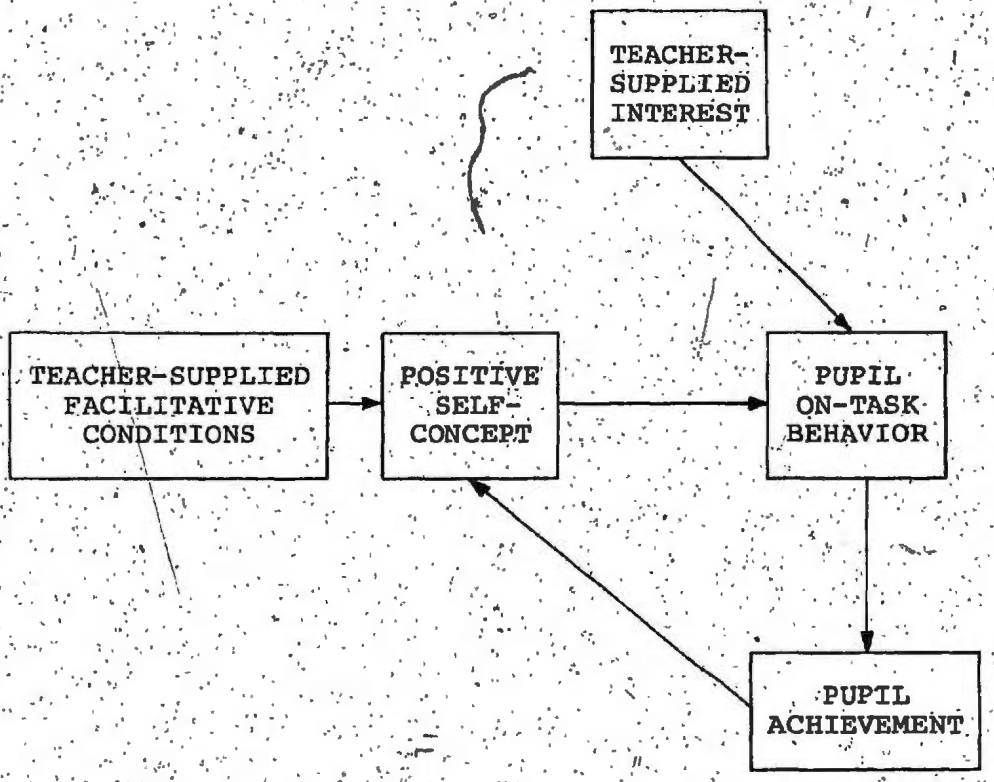

FIGURE 1. The self-enhancement model.

In this view, the provision of a supportive and facilitative learning situation is separate from the teacher process of motivating students to become involved in the learning activity. Teacher-supplied interest is the added ingredient which serves to direct the child to on-task behavior.

There are two areas of concern with this model of motivation. First, it does not explain why the student should actually engage in on-task behavior. Since the

facilitating conditions are supplied unconditionally, the student benefits whether he is engaged in learning or not. Pupil on-task behavior relies solely on the teacher's ability to captivate the interest of students. To suppose that a teacher is capable of 'entertaining' all students successfully, especially when a task is of itself dull and uninteresting, is simply not realistic. Secondly, there is research which suggests that this model does not work for all students. Though the level of achievement does in fact rise in classrooms which exhibit facilitative climates, the variability of the achievement increases as well (Thurstone, 1936; Johannesson, 1967). The facilitative environment, therefore, influences some children more than others.

A second motivational model may be found within the 'social learning theory' of Bandura and Walters (1964). Their approach to learning maintains that the learning of appropriate behaviors requires a model of the desired behavior and contingent reinforcers of that behavior. The type of reinforcement and the manner in which they are applied determine, to a great degree, the effectiveness of this reinforcement in teaching children. Assuming that students have fundamental needs as formulated by Maslow (1962), reinforcement could be seen as the satisfaction or deprivation of these needs.

Maslow's theory presents a set of human needs which are essentially of two types, growth and deficiency (Maslow, 1962). The deficiency needs are comprised of the physiological, safety, love and belonging, and esteem needs. These needs, according to Maslow, are satisfied mainly through social interaction. It follows, then, that children would behave in ways which lead to satisfaction of these needs and conversely would avoid behaviors which fail to satisfy or which increases the likelihood of deprivation of these needs. The second category of needs, the growth needs of self-actualization and aesthetics, are quite different in that satisfaction of these needs comes from within the self, through the experiences encountered by the child. Maslow (1962) concluded that deficiency needs must be satisfied in order to maintain the mental health of the individual. Growth needs, on the other hand, could be deferred or not be satisfied at all without any deleterious effect to the child's emotional well being.

The satisfaction of the child's deficiency needs requires only that the reinforcement be gained through social interaction, regardless of the source. Therefore, within the classroom context, the critical issue becomes the control of the sources of contingent reinforcement. There are two primary sources of reinforcement available within the classroom--the teacher and the child's peers. The teacher who has control of the satisfaction of a pupil's

deficiency needs could make this satisfaction contingent upon the pupil's participation in on-task behavior. Motivation, for the teacher, therefore, becomes synonomous with controlling the on-task behavior of her students through manipulation of contingent reinforcement. The schematic presented in Figure 2 demonstrates how this approach to motivation could explain classroom functioning.

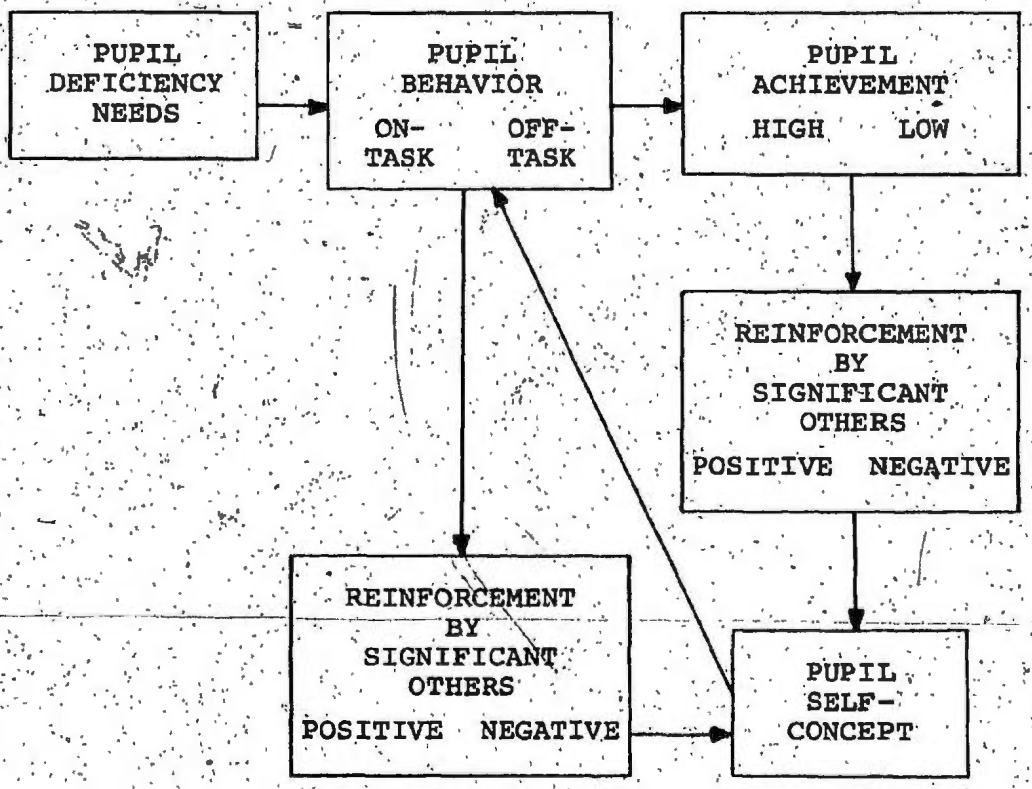

FIGURE 2. Social learning model.

The application of this theory to actual classroom practice might for a number of reasons be rather difficult. For example, the student might be achieving satisfaction of his needs through an entirely different source and would, therefore, not become as involved in on-task behavior, having no reason to do so. The student might view the teacher as being an inadequate source of reinforcement and would in this case as well be beyond the control of the teacher in terms of application to on-task activity. Parents are an obvious primary source of needs satisfaction for most children and would undoubtedly serve to confuse the 'clean' lines of the model. However, it is generally accepted within education that there must be some agreement between teacher and parent on the reinforcement a child receives if that child is to achieve in school at or near his potential. Teachers themselves might be self-defeating, while still operating within the confines of the model. The child might, for example, be reinforced for his behavior even if it is not on-task and actually learn a route to needs satisfaction which would in fact decrease academic achievement.

The satisfaction of growth needs does not fit within this model, since growth needs require no external reinforcement. The satisfaction to be gained comes mainly from the activity itself, that is, from the actual 'doing'. Growth motivation further implies that

deficiency needs are already being satisfied. As Figure 3 indicates, the teacher must in this situation seek to interest the child in taking part in the on-task activity. The teacher might accomplish this by either providing high interest activities or by some means inculcating interest within the child.

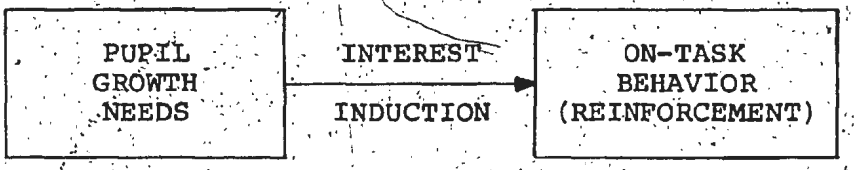

FIGURE 3. Satisfaction of growth needs

The 'self-enhancement model' arose in part from the great weight of research which had consistently found that a facilitative climate did, in fact, enhance pupil achievement. A comparison of the two models presented will demonstrate that the 'social learning model' does provide for the integrity of this research. The 'social learning model' has applicability here since a parallel can be drawn between the 'facilitative conditions' and the satisfaction of needs as proposed by Maslow (see Table 1). Acceptance corresponds to the love and belongingness need, praise or esteem enhancement corresponds to the esteem need, and interests correspond to the actualization and

TABLE 1

Aspects of the Social Learning Model

Needs	Sources of Satisfaction	Examples of Contingent Reinforcers
Deficiency:		
Physiological	Family	Token reinforce-
	Teacher	ment with candy
Safety	Family	Threat of physical
	Teacher	punishment
Love and belongingness	Family	Acceptance, respect
	Teacher	Threat of rejection
	Peers	Peer acceptance
Esteem	Family	Praise, esteem enhancement
	Teacher	Threatened loss of status
	Peers	Peer approval
Growth:		
Self-actualization	Activity	Interest
Aesthetic		Challenge

aesthetic needs. The 'social learning model' provides for the deficiencies of the 'self-enhancement model' because of the fundamental difference inherent in the provision of these facilitative conditions. While the humanistic approach would provide these conditions unconditionally, social learning theory demands that they be provided contingent upon the on-task behavior of the student.

Viewing the motivation of students from the social learning model has several distinct advantages in explaining some classroom phenomena related to pupil achievement (Turpin, 1981). A teacher might, for example, generalize his motivational technique to the entire class. Such being the case, a child whose needs were different from those being provided for might not engage in the activity desired by the teacher. If, on the other hand, a teacher were selective in applying reinforcement, consequently giving more to some than to others, the students receiving the greater degree of reinforcement would be expected to achieve more than those students receiving less reinforcement. The model also serves to explain a situation wherein a teacher simply does not provide sufficient reinforcement in terms of acceptance, esteem enhancement, and interest. In these conditions, children would not participate in appropriate learning activities and would, therefore, achieve below expected levels. In cases where the primary source of needs satisfaction is the child's peers or parents, the teacher's

lack of control of student learning behavior would also be demonstrated in terms of this model of motivation.

These relationships would require study and research. The validity of the model must be examined and its usefulness explored. Such research requires that careful observation of teacher-student interaction be undertaken and reliably recorded. A careful review of the available classroom observation scales revealed that no instrument was ideally suited to the type of observation which this model required. An instrument was needed which was consistent with the requirements of the model, particularly with respect to student on-task or off-task behavior and to the contingent reinforcers to be examined. This study was aimed at providing such an instrument.

Essentially, the 'social learning model' required a 'bifocal' observational scale, one which focused both on the student and teacher in interaction within the classroom. With respect to the student, the scale had to be sensitive to the observable interest demonstrated by the student toward learning activities, as well as to the types of interactions entered into both with the teacher and with fellow students. The teacher-focus aspect of the scale had to be directed primarily toward the motivating or non-motivating activity of the teacher. In terms of teacher motivation, the instrument would be required to discriminate between positive and negative forms of contingent reinforcers. Most importantly, the

instrument needed the capability of classifying these reinforcers as occurring on one of three levels, depending upon which student need was being reinforced by the teacher.

In addition, some discrimination needed to be made between reinforcers which directly rewarded or punished a child and those which served as cues or reminders that reinforcement was contingent upon a particular type of behavior. This 'direct-indirect' aspect of the scale would be of importance in distinguishing between those children who had learned the acceptable behaviors leading to need satisfaction, and those who had not.

The observation instrument presented in this report was designed to meet the requirements of the 'social learning model'. The instrument was intended for research into teacher-student interaction from within the framework of this model. Specifically, the instrument was designed for use in ongoing and intended research into the relatedness of pupil self-concept and degree of peer adjustment to the level and type of motivation employed by teachers to elicit on-task pupil behaviors.

Chapter II will consider some of the literature related to previous work in producing classroom observational instruments. Particular emphasis will be paid to some of the more historical and representative observational systems and to the problems associated with development and implementation of this form of research instrument.

CHAPTER II

REVIEW OF RELATED LITERATURE

From Humble Beginnings

Much has been written concerning teacher-student relationships and the outcomes devolving from those relationships. The earliest work in this area concentrated mainly on trying to identify the characteristics of 'good' teachers. An article by H.E. Kratz entitled "Characteristics of the best teachers as recognized by children," published in 1896 marked the beginning of published work in this regard. Kratz's study, as the title implied, used students as observers. These students were asked to describe the 'best' teachers and the ensuing descriptions then formed the basis of a list of 'good' teacher qualities. Early research in this field was indeed characterized by the use of student descriptions arising from their perceptions of 'good' teachers (Medley, 1972).

This form of research continued on into the mid-1900's. A major study by P.A. Witty (1947) used letters written by approximately 12,000 students as the medium from which was drawn a compilation of qualities of the 'best' teachers. These students, ranging in age from 9 to 17,

wrote on the topic "The teacher who has helped me most." A content analysis of these letters then revealed that the two major traits found in these 'good' teachers were; (i), a cooperative and democratic attitude; and (ii), kindliness and consideration for the individual.

This particular approach to the study of teaching behaviors was found not to be effective, the apparent reason being that students proved to be no more aware of the qualities of effective teachers than was anyone else. This led to a movement toward the use of rating scales in which experts attempted to identify the characteristics of good teachers. This method as well proved to be quite unreliable and by the 1930's there was a growing argument for the use of objective measures in studies of teacher behavior and characteristics (Medley, 1972).

As in most cases, this change too was slow in coming. A review of the literature by Medley and Mitzel (1963) uncovered only some 20 studies which employed objective means in analyzing teachers' classroom behavior.

With the swing from subjective to objective measures of behavior came also a movement from involvement with presage type variables toward a greater interest in process variables as they related to student outcomes. There was a moving away from the notion of trying to identify the attributes common to 'good' teachers and a placing of more emphasis on the behaviors brought to the classroom by

teachers who appeared to produce success in terms of pupil growth.

This point was brought out most clearly in two reviews written in 1954 concerning the relationship between teacher characteristics and pupil achievement (Morsh & Wilder, 1954; Ackerman, 1954). In all but one of the 25 studies reported by Wilder or Ackerman, the teacher characteristics included presage variables such as age, intelligence, experience and behavior assessed by rating scales marked by pupils or superiors. The overall conclusion by these researchers was that the results were inconclusive and at times contradictory. Both reviews recommended the use of systematic observation techniques in future studies.

The educational research community did in time react. Barak Rosenshine, in a review written in 1969, reported on over 20 investigations which had employed 'category' systems to determine relationships between teacher input and pupil outcomes. Since the 1960's there has been a virtual blossoming of instruments and investigations which endeavoured to probe the teacher-student interaction for its effect on both participants.

In summary, the past half century has witnessed a movement toward systematic observation of classrooms as a means of investigating the teaching-learning relationship. Though evolution was slow, systematic observation has

occupied much of the research time devoted to education in recent years.

A Brief Word on Teachers and Students

The past two decades have witnessed a tremendous degree of interest in the results of differing teacher behaviors on their charges. The general finding of most of the research into teacher behaviors has been that positive teacher behaviors are more frequently associated with student growth than are negative teacher behaviors (Reed, 1961; Turner & Denny, 1969; Klein, 1971; Aspy, 1972; Redd et al., 1975; Woolfolk, 1978). Positive teacher behaviors may be loosely identified as accepting and encouraging behaviors while the negative variety would be exemplified by rejecting and critical teacher remarks.

The characteristics described by Rogers (1951) as the facilitative conditions might serve to describe what has come to be accepted in the literature as the 'positive' teacher. These 'facilitative conditions' of acceptance, understanding, genuineness and respect do have an apparent influence on what students learn (Aspy & Roebuck, 1977; Fox & Peck, 1978). Aspy and Roebuck (1977) in a review of the literature concluded that teacher levels of the facilitative conditions were positively related to their students' achievement, IQ gains and school attendance. A previous review by Soar (1972) also found a positive relationship

between facilitative teacher behavior and student achievement growth, and as well concluded that teacher criticism tended to be negatively related to achievement gain. However, an independent study by Soar (1971) of kindergarten and first grade children, although in agreement with the findings linking positive affect with student growth, found that the presence or absence of negative teacher behavior had no effect on pupil cognitive growth. In general though, the teacher most liked by children and the teacher from whom they will most apparently learn "is the one who is quiet and friendly, the one who can talk easily with them and share the occasional joke" (Nash, 1976, p. 91).

A very comprehensive review of research relating pupil achievement and teacher behavior by Rosenshine (1971) reinforced the findings related by Soar (1972) and Aspy (1977). In addition, Rosenshine reported that, while extremes of criticism such as continued disapproval appeared to be negatively related to achievement, milder forms of criticism such as telling a child he was incorrect or calling for attention were not related to achievement in a negative sense. In fact, some studies showed a positive correlation between student achievement and mild criticism. Another interesting finding related by Rosenshine was that frequency of praise such as a teacher's saying "alright" were just as effective as strong praise and encouragement.

A large number of research studies have examined the differential treatment given students on the basis of their ability, social status, approval, and various other characteristics (Silberman, 1969; Jackson et al., 1969; Good & Brophy, 1973). One study found that teachers interacted more directly with pupils rated higher on a continuum from a typical 'worst' to 'best' student (Dalton, 1969). A number of studies have consistently shown that positively behaving students elicit positive behavior on the part of the teacher (Rosenfeld, 1967; Sarbin & Allen, 1968; Thomas & Becker, 1968; Klein, 1971).

Results of research have also been fairly consistent in finding that highly achieving students receive more teacher praise and support than do underachievers (de Groat & Thompson, 1949; Hoehn, 1954; Simon, 1966; Brophy & Good, 1970; Good, 1970). Brophy and Good (1970) concluded that teachers, in fact, did communicate differential performance expectations to different children through their in-class behavior. Teachers are "more likely to accept poor performance from students for whom they hold low expectations and are less likely to praise good performance from these students when it occurs, even though it occurs less frequently" (Brophy & Good, 1970, p. 367). Good's (1970) study showed as well, that teachers extended to high achievers significantly more opportunities to respond in class discussions. In fact, teachers actually deprived low

achievers of opportunities to respond in competitive situations.

To summarize, results of the majority of studies in this area of teacher-student interaction demonstrated that teachers do in fact, by means of their behavior, affect student growth. There is, however, a good deal of confusion still as to what is the appropriate balance between positive and negative teacher behaviors. As well, there appears to be good evidence that different pupils require essentially a different degree of this balance in their teachers before they will reach a maximum level of growth in an educational environment (Thurstone, 1936; Carkhuff & Truax, 1967; Johannesson, 1967; Soar, 1969; Marliave, 1976; Aspy, 1977).

Some Selected Observational Systems

Simon and Boyer (1974), in their anthology of observational instruments, included information on 99 systems developed to observe human communication. Of these, 78 belonged to the field of education. Each of these systems had a specific and usually unique focus on some aspect of communication and human interaction. Six observational systems were of particular relevance to the present study. Four systems were among those reported by Simon and Boyer: (i) the H.H. Anderson System; (ii) the Flanders System of

Interaction Analysis (FSIA); (iii) the Observation Schedule and Record: Form 4 (OScAR 4V); and (iv) the Teacher-Child Dyadic Interaction System (DOS). The remaining systems were the Florida Climate and Control System (FLACCS), and the Classroom Observational Instrument (COI). For a more complete discussion of these and other systems reference should be made to the anthology of Simon and Boyer (1974) and to excellent reviews by Medley and Mitzel (1963), Boyd and De Vault (1966), Rosenshine (1970), Flanders (1970) and Borich and Madden (1977).

There are basically two different forms of classroom observational systems, those being 'category' and 'sign' systems (Medley & Mitzel, 1963; Biddle, 1967; Dunkin & Biddle, 1974). The interpretation of these labels varies somewhat depending on the source and much confusion could exist concerning the proper classification of a particular system. A 'sign' system is usually considered to be one in which an observer is given a list of behaviors to watch for and is then requested to record that behavior each time it occurs during a given period of time. In contrast, a 'category' system is characterized by having observers make judgements on behaviors and then decide where a particular behavior should be assigned from among a list of categories. A 'sign' system then endeavours to give a frequency count for listed behaviors within a classroom, while a category system seeks to provide a more sequential and flowing

account of all behaviors which occur.

Category systems are generally considered to provide more information and more flexible information than do 'sign' systems (Dunkin & Biddle, 1974). Category systems, because of their sequential nature, lend themselves more to the study of actual patterns of interaction. Two of the systems presented, the OScAR technique developed by Medley & Mitzel (1958) and the FLACCS (Soar et al., 1971) are considered to be 'sign' systems. The remaining four are essentially 'category' systems.

The Anderson System

Harold Anderson is considered as one of the pioneers in the work of recording the process of human interaction (Simon & Boyer, 1974). Anderson and his associates studied both the contacts of teachers with children and pupil behaviors themselves (Medley & Mitzel, 1963). A method for simultaneously observing pupil and teacher was developed which attempted to distinguish between socially dominative versus integrative teacher behavior, while at the same time noting pupil behavior change in varying circumstances. There was no attempt to categorize students' interaction with classmates.

Each child was observed for five minutes at a time, and interaction between that child and the teacher only were recorded. The reliability of the instrument in terms

of observer agreement was generally quite high (Anderson, 1967). The number and type of categories used varied with the purpose of the study being done and only a small sampling of these categories are given in Tables 2 and 3. Essentially, the work of Anderson and his colleagues led the way for further research into classroom interaction, but it was over a century later that the next significant development occurred with the work of Ned Flanders.

The Flanders System of Interaction Analysis (FSIA)

The Flanders system has been the observation instrument most frequently used in analysis of the influence of teachers' classroom behavior. The instrument is 'bifocal,' but interest is keyed primarily to teacher behaviors which restrict or increase the student's freedom within the classroom (Borich et al., 1977). The instrument contains only 10 behavior categories (see Table 4) with recording done on a 10 x 10 matrix. The low number of categories combined with the ease of scoring makes the FSIA a relatively simple system to use in classroom observation. This simplicity has naturally enabled the users of the system to fairly consistently achieve coefficients of observer agreement in excess of 0.85. The matrix system itself actually provides a visual diagram of the interaction pattern within a classroom and to a large degree this preservation of information regarding the sequence of behavior made the

TABLE 2

Sampling of Categories Used by Anderson to Record Teacher Contacts*

Domination with conflict	
DC - 1	Determines a detail of activity in conflict
DC - 3	Relocates a child
DC - 4	Direct refusal
DC - 6	Warnings, threats, reminders
DC - 11	Punishment
Domination with no conflict	
DN - 1	Determines a detail of activity, mostly of a routine sort with no conflict
DN - 6	Warnings, threats, reminders, conditional promises with no evidence of conflict
DN - 9	Lecture method, questions
Integration without working together	
IN - 19q	Question regarding possible, though not expressed interest or activity of the child
IN - 19s	Statement regarding possible, though not expressed interest or activity of the child
Integration with evidence of working together	
IT - 14	Helps child to define or advance problem
IT - 16	Approval, thanks, acceptance of spontaneous behavior of child
IT - 17	Questions regarding the child's expressed interests

*Adapted from Anderson et al. (1946, 22-27).

TABLE 3

Sampling of Categories Used by Anderson to Record Pupil Behaviors*

1. Nervous habits
2. Leaves seat
3. Child domination of other children
4. Nonconforming to teacher's commands
5. Holds up hand
6. Answers when called upon
7. Fails to answer when called upon
8. Problem-solving
9. Tells experience
10. Brings something to school
    a. voluntary
    b. in response to invitation
11. Suggestion
    a. voluntary
    b. in response to others
12. Offers services
    a. voluntary
    b. in response to others

*Adapted from Anderson et al. (1946, 27-30).

TABLE 4

Categories for the Flanders System of Interaction Analysis*

TEACHER TALK	INDIRECT INFLUENCE	1.	Accepts feelings: accepts and clarifies the feeling tone of students in non-threatening moments
		2.	Praises or encourages: praises student action or behavior
		3.	Accepts or uses ideas of student
		4.	Asks questions: with the intent that the student answer
	DIRECT INFLUENCE	5.	Lecturing: giving facts or opinions
		6.	Giving directions: including commands
		7.	Criticizing or justifying authority
STUDENT TALK		8.	Student talk--response: a student makes a predictable response to the teacher
		9.	Student talk--initiation: talk by student which he initiates
		10.	Silence or confusion: pauses, or periods where communication cannot be understood by observer

*Adapted from Simon and Boyer (1974, 235).

Flanders approach unique (Medley & Mitzel, 1963). The FSIA has been most successful in its application to teacher training activities as a means of providing teachers with feedback on their own teaching behavior.

The greatest weakness of the Flanders system is its over-emphasis on teacher influence. Only two of the 10 categories focus on the student. This tends to lessen the instrument's effectiveness in studying the 'whole' of teacher-student interaction (Borich et al., 1977; Mitchell, 1969).

The Observation Schedule and Record (OScAR 4V)

This system developed by Donald Medley and Harold Mitzel (1958) and later modified was intended to analyze teacher and pupil verbal behaviors. As with the Flanders system, the primary focus is once more on the teacher (Simon & Boyer, 1974). This system attempts to divide all classroom interactions into two distinct sets of behavior; interchanges and monologues. The interchanges focus on the teacher's behavior alone, noting how the teacher initiates an interchange and how he responds to a pupil answer (see Table 5). The observer uses a flow chart to record behavior in this system. The intention here is to build a picture of the flow of interaction; however, the system breaks down since it is essentially a 'sign' system wherein the observer records his overall impressions of a

TABLE 5

A Sampling of Categories for Behavior in OScAR 4V*

I. Statements

A. Teacher statements--utterances which neither respond to nor solicit a response from a pupil are classified as follows:

1. AFFECTIVE. A statement revealing sensitivity to pupil feelings is classified as CONSIDERING. A statement criticizing pupil conduct is classified as REBUKING.

2. SUBSTANTIVE. A statement containing no affect but referring directly to content such as INFORMING if it contains a fact or PROBLEM STRUCTURING if it sets up a question to be solved.

3. PROCEDURAL. A statement which contains neither affect nor substance is classified as DIRECTIVE if it contains a command. A statement which does not fall clearly into one of the above categories is classified as DESCRIBING.

B. Pupil statements--utterances by pupils addressed to other pupils are classified as PUPIL STATEMENTS.

II. Interchanges

An interchange is an episode in which a pupil says something to the teacher and the teacher reacts.

A. Substantive interchanges--those in which the pupil's utterance refers to content to be learned.

*Adapted from Simon and Boyer (1974, 403).

given interval of classroom behavior (Dunkin & Biddle, 1974).

OScAR has been widely used, and as with the FSIA has been most helpful in applications to teacher training. The overemphasis on teacher behaviors has hampered its usefulness in researching the actual interaction in classrooms and doubtless contributed to its failure to get at any major aspects of classroom behavior related to pupil achievement (Medley & Mitzel, 1963).

The Teacher-Child Dyadic Interaction System (DOS)

The dyadic system developed by Good and Brophy (1969) is different from most observational devices in that it records separately the teacher's interactions with individual children. The authors of the device maintained that it is quite often inappropriate to treat the class as a whole and argued that the child should form the unit of analysis rather than a class of children (Borich & Madden, 1977). The system was designed to investigate the relationships between teacher expectancies and pupil achievement. The observation of individual students is intended to provide a record of teacher behaviors toward different types of learners in the classroom (Simon & Boyer, 1974).

The basic observation unit in the DOS in a sequence of behavior beginning with a teacher question, followed by a pupil response and the teacher's reaction to that response

(see Table 6). This approach makes possible the freezing of data relative to individual children and teacher behaviors directed toward those children (Emmer, 1972).

The system represented a new emphasis in research orientation away from a preoccupation with teacher behaviors toward an analysis of the actual interaction between teacher and learner. If there is a weakness with the Brophy-Good system, it is that it goes a little too far and actually discounts teacher interaction with the class as a whole. Class-directed comments and statements might, in fact, greatly affect individual students. The Dyadic system as well ignores student-student interaction and the impact these might have on both pupil and teacher behavior. With respect to the demands of the 'social learning model', the Dyadic approach as developed by Brophy and Good does not provide for the range of teacher motivating behaviors. The DOS focuses only on praise and criticism and makes no reference to acceptance and interest as pupil reinforcement.

The Florida Climate and Control System (FLACCS)

The fifth of the systems under discussion here was designed to record behavioral dimensions not covered by Interaction Analysis. These include classroom grouping, individual versus group work, and nonverbal affective expression in the classroom (Borich, 1977). The system was developed to aid in the study of the educational needs

TABLE 6

Sampling of Behavior Categories Used in Recording Teacher-Child Dyadic Interaction*

I. Response Opportunities

A. Directs questions
B. Open questions
C. Call-outs
D. Chorus questions
E. Discipline questions
F. Reading turns
G. Recitation opportunities

II. Level of Question

A. Process questions
B. Product questions
C. Choice questions
D. Self-reference questions

III. Quality of Child's Response

A. Correct response
B. Partially correct response
C. Incorrect response
D. No response

IV. Teacher's Feedback Reactions

A. Praise
B. Criticism
C. Product feedback
D. Process feedback
E. Repetition of question
F. Rephrasing of question
G. Asking a new question
H. Failure to provide feedback

(cont'd.)

Table 6 (cont'd.)

V. Work-related Contacts

A. Teacher-afforded
B. Child-created

VI. Behavior Evaluation

A. Praise
B. Warning
C. Criticism

VII. Procedural Contacts

A. Teacher-afforded
B. Child-created

*Adapted from Good and Brophy (1972).

of disadvantaged children, and was by design intended to be a 'low inference' instrument. The items used to describe behavior are, for the most part, specific, demanding little observer judgement (Soar et al., 1971). The difficulty with being specific, of course, is that a very large number of categories is required to provide adequate coverage. This is one of the major difficulties with the FLACCS. The system requires a two-page coding scheme containing some 180 items. A small sampling of these is presented in Table 7.

TABLE 7

A Sampling of Categories Used in the Florida Climate and Control System*

Teacher		Pupil
10	Teacher central	10 Pupil central
11	Leads singing, games	11 Pupil--no choice
12	Moves freely among pupils	12 Pupil--limited choice
13	Withdraws from class	13 Pupil--free choice
14	Uses blackboard, A-V equipment	14 Seat work without teacher
15	Ignores, refuses to attend pupil	15 Seat work with teacher
16	Attends pupil briefly	16 Works, plays with much supervision
17	Attends pupil closely	17 Works, plays with little supervision
18	Attends pupil in succession	
19	Attends simultaneous activity	
	Verbal Control	
		18 Resists, disobeys directions
		19 Obeys directions
20	Praises	20 Asks permission
21	Asks for status	21 Follows routine without reminder
22	Suggests, guides	22 Reports rule to another
23	Feedback, cites reason	23 Tattles
24	Questions for reflection, thought	24 Gives information
25	Correct without criticism	25 Gives direction
26	Questions for control	26 Gives reason
27	Questions, states behavioral rule	27 Speaks aloud without permission
28	Directs with reason	28 Engages in out of bounds behavior
29	Directs without reason	29 Collaborates with teacher
30	Uses time pressure	30 Task-related movement
31	Call child by name	31 Shows pride
32	Warns	

*Adapted from Soar et al. (1971).

Being a sign system, the data generated by the instrument does not retain the sequencing of behavior needed to adequately study classroom interaction. A search of the literature did not reveal information regarding applications of the system developed by Soar (1971). However, the system does serve to indicate the problems encountered in foregoing simplicity of construction in favor of highly specific, low-inference, behavioral categories.

The Classroom Observation Instrument (COI)

The most recent of the instruments reviewed was developed by the Stanford Research Institute for use in the Follow Through Project established by the U.S. Congress in 1967 (Stallings, 1975). Project Follow Through was originally designed to examine the differential effectiveness of some twenty-two programs based on various educational and developmental theories.

In order to test the effectiveness of these differing programs, the project first had to determine whether classrooms were actually implementing the various programs. This required systematic classroom observations. Since existing observational instruments were considered too limited in scope in that they usually focused on a single theory, the COI was developed.

Because it had to be sensitive to the concepts contained in several theories, the COI evolved as a massive

instrument. The final version of the observation scale contained 602 items of classification, divided into two major sections; the Observation Summary Form (OSF), and the Classroom Observation Procedure (COP). The OSF was used to record information such as the physical arrangement of classrooms, while the COP focused on more specific information about classroom structure and process. The COP was also divided into two sections. The first of these, the Classroom Check List (CCL), was used to code the conditions of instruction. The second section, the Five Minute Observation (FMO) was designed to summarize information about a selected individual called a 'focus person' who was observed for five consecutive minutes. The FMO consisted of seventy-six frames, with each frame having four sections. In practice, the observer coded the sections in each frame in sequence to form a 'sentence' describing the classroom action. The collection of the data in sequence enabled 'strings' of data sentences to be examined (Stallings, 1973).

The FMO section of the Classroom Observation Instrument contained 262 categories of behavior. A sampling of these behaviors is given in Table 8. The categories used, though focusing on positive and negative feedback, did not contain categories referring to levels of positive or negative reinforcement.

TABLE 8

A Sampling of Categories used in the Classroom Observation Instrument*

413a	Child not responding to adults
414a	Adult not responding to child
415a	Child waiting
416a	Children attentive to adults, nonacademic
417a	Children attentive to adults, academic
418a	Adults attentive to children, nonacademic
419a	Adults attentive to children, academic
420a	Adults attentive to a small group
421a	Adults attentive to individual children
422a	Positive behavior among children
423a	Positive behavior adults to children
424a	Positive behavior children to adults
425a	Child expressions of unhappiness
426a	Adult expressions of unhappiness
427a	Negative behavior among children
428a	Negative behavior adults to children
429a	Negative behavior children to adults
430a	Total adult affect
431a	Total child affect
432a	Adult punishment of children
433a	Child statements of self-worth
434a	Dramatic play, pretending
435a	Total academic verbal interactions
436a	Total interactions behavior control
437a	Children engaged in mutual activity

*Adapted from Stallings (1975, p. 115).

The scope of the COI is broad. A problem with the instrument might be that it attempts too much. Stallings (1973) admitted that some of the theories involved in the Follow Through Project were better reflected in the COI data than were others. The sheer size and scope of this instrument also combines to make observer training a monumental and expensive proposition.

In summary, the six observational systems presented here are believed to be representative of and indicative of the work performed in the field of systematic classroom observation. As the material presented indicated, this work has been quite varied. Most systems were developed with particular studies in mind and are, therefore, specific in purpose. A number of instruments such as the Dyadic system (Brophy & Good, 1969) and the COI (Stallings et al., 1973) had limited adaptability. The Brophy-Good system, though being the most relevant to the needs demanded by the 'social learning model' of motivation, does not lend itself to the total requirements of the research intended. A system meeting those requirements needed to be developed if the model and the theory were to be investigated.

The next section of this chapter will deal with some of the problems encountered in the development and implementation of classroom, observation instruments.

Considerations in Developing an Observation System

Observation systems are essentially tools for describing human communication. Simon and Boyer (1974) referred to observation instruments as "the meta-languages of communication". Observation systems are related to communication as parts of speech are to grammar. Observation instruments are used to analyze and dissect the elements of communication. These systems generally are comprised of a set of rules for classifying observed activities using a standardized procedure. The classification is usually accomplished by assigning the activity to a particular category from among a set of pre-defined categories. Since this categorization of behavior tends to 'abstract' communication, great care must be taken in ensuring that the system is efficient and valid. The major elements to be considered in the development of an observation system are dealt with in the following pages.

Type of System

There are two chief types of observation systems used within the educational context. Whether a 'sign' or 'category' system is to be developed is usually determined by the type of information which the researcher is interested in collecting. There is, therefore, no best system, but rather the system which is better suited to a particular

situation. Medley and Mitzel (1963) noted that 'category' systems were more likely to develop from studies which were based on a theoretical approach wherein the researcher was interested in looking at specific behavior to be examined in light of theoretical hypotheses. 'Sign' systems, on the other hand, tended to originate from studies which were looking for cause and effect relationships and were not guided by theory toward a particular set of behaviors.

The methodologies required in developing each type of system differ to some degree. The following discussion will focus on the construction of 'category' systems. Much of the discussion, though, has applicability to 'sign' systems.

Categorization

The first major decision an instrument developer must make is that of focusing the observation system. What behaviors are to be observed and how are they to be placed into categories? Since 'category' systems usually arise from some theoretical basis, the behaviors to be observed are in most instances dictated by the theory or model under consideration. However, how these behaviors are to be grouped together and then defined in concrete terms is most important for later application. Martin (1977) noted that the process of categorization is basically 'one of successive approximation'. The researcher groups and re-

groups behaviors until finally he arrives at a relatively small number of generic behavioral sets. There is general agreement that the number of categories should be kept as small as possible and the number 'ten' has been suggested as perhaps an optimal one (Martin, 1977; Medley & Mitzel, 1963).

Having grouped behavioral activities into categories, the researcher must ensure that the categories form a 'facet'; that is, the categories should form a mutually exclusive, all inclusive set (Simon & Boyer, 1974). This ideal situation is usually not possible and some behaviors will not fit into the categories provided. This makes necessary a catch-all miscellaneous category for such troublesome behaviors. Category 10, "silence or Confusion", in the Flanders system is an example of this type of provision (Karafin, 1973). A second problem may arise with behaviors which could be classified in two or more categories. This situation can be overcome only through extensive observer training.

Categories must be well defined. They must, in fact, be most specific. Categories, as such, cannot be theoretical constructs; they must be readily observable intervening variables. The dimension which can cause the greatest concern in the definition of categories is the level of judgement or inference required by the observer.

A 'category' system requires a small number of categories if it is to be effective. These categories must be as mutually exclusive as possible while still covering the spectrum of behaviors under study. This can only be accomplished by defining the categories in a relatively 'high inference' manner. The ultimate classification of behaviors will rely a great deal upon observer judgement. With proper training, though, observers can become quite adept at classifying 'high inference' behaviors such as teacher praise or enthusiasm (Brophy et al., 1976).

Conceptual Posture

The notion of 'conceptual posture' was addressed by Dunkin and Biddle (1974). Conceptual posture refers to the actual intent and orientation of the observation process, that is, 'what should be observed?' Is the study more interested in; (i), the intent of the observed behavior; (ii), the observed characteristics; or (iii) the effects of that behavior? Dunkin and Biddle pointed out that each of these might be recorded equally well by observers but might not all be useful depending on the requirements of a particular study.

Unit Sampling

The debate concerning the units to be used in sampling behaviors has continued from the first use of

observational systems. There are two basic approaches to behavior sampling. The unit of behavior might be a phenomenal one such as a statement or question, or change of speaker; or the researcher could use a time-unit method of sampling. The time unit is usually brief and the observer is normally asked to record one observation during each time unit. The major problem with time units (Dunkin & Biddle, 1974) is that they disrupt the normal rhythm of the classroom as it is being recorded. Since actions usually occur quite rapidly in classrooms, often the observer has to make judgements concerning which behavior to record, thus introducing subjectivity and error.

Brophy et al. (1976) have argued that time sampling is not psychologically valid and suggested that the unit of observation should be a 'molar' one as is used in the Dyadic Interaction System (Brophy & Good, 1969). An example of a molar unit would be: 'teacher asks question, student answers, teacher reacts with feedback'. An obvious fault with the molar approach is that it will ignore much behavior which does not fit the particular pattern under observation. Martin (1977) emphasized that time units have an advantage in that they lend themselves more to use by a number of observers across a number of categories. Medley and Mitzel (1963) cautioned that, though time units are useful, they should be kept relatively short to eliminate observer judgements as much as possible.

Recording

The greater proportion of observation systems have been designed to gather information live within the natural setting of the classroom. A small number of instruments have been designed specifically for use with recordings of classroom behavior. A large number of present systems could naturally be adapted for either 'in vivo' or 'in vitro' use. There has been a mild debate in the literature in recent years as to which approach is most efficient and valid. Generally, speaking, the cost of obtaining an audio or videotape record of classroom behavior and then training people to transcribe and analyze the recording is considerably higher than the cost of training and using live observers in classrooms (Nuthall & Church, 1972). The distinct advantage, of course, in having a permanent recording is that it can be extensively analyzed by replaying all or certain episodes any number of times.

In discussing research efforts using category observation systems, Biddle and Adams (1967) and Medley and Mitzel (1963) maintained that the simultaneous recording and codifying of classroom behavior by live observers has served to contaminate the data collected, contributing to the unreliability of the observational system. Biddle and Adams (1967) went on to argue that audio or video recordings ought to be made of classroom interactions and these recordings be subsequently analyzed, thereby separating

the functions of recording and coding.

The chief disadvantages of mechanical recording in classrooms appear to be related to the quality of the records received and to the effects of placing recording equipment in classrooms. In the first instance, the problem of quality is mostly related to the difficulty of getting both a good sound and video record from most classrooms. These problems arise mainly because of technical difficulties resulting from placement of cameras and sound microphones. There is, as well, a very rapid loss of quality when videotape is 'dubbed' or re-recorded on a second tape. Secondly, researchers have noticed that the introduction of recording equipment does have a pronounced effect on the behavior of both teachers and students. A study by Stukat and Engstrom (1967) found that in the presence of television cameras, teachers tended to speak more often and pupils less often. These effects apparently wear off after a few days. This is not to say that live observers do not have an impact on the behaviors of classrooms in which they are observing. On the contrary, most writers warn that observers should be kept to an absolute minimum and should be fully aware of their particular purpose in being within the classroom (Medley & Mitzel, 1963).

An interesting study by Long (1974) gathered information from classrooms using both live observers and

videotape recordings. An analysis of the data obtained by each method revealed very little difference in using either one. Long concluded that other factors such as cost or ease of data collection should be given prime consideration in deciding whether to use live or 'in vitro' data acquisition.

Reliability

The meaning and determination of reliability of observation instruments has been for several years and remains a very contentious issue. Generically, reliability refers to the accuracy and consistency with which the measures obtained by an instrument describe what it purports to describe (Herbert & Attridge, 1975). Frick and Semmel (1978) argued that the confusion existing with regard to reliability of observational measures arose from the failure to separate two statistical labels which are quite different conceptually: observer agreement coefficients and reliability coefficients. Of the studies which have included a discussion of reliability, most have in fact equated inter-rater reliability with the reliability of the instrument (Dunkin & Biddle, 1974).

The generally accepted view on the reliability of observational systems is that inter-rater reliability, though being important, is not sufficient to determine the reliability of an observation system (Tinsley & Weiss, 1975; Rowley, 1976; Rajaratnam, 1972; Frick & Semmel, 1978). The

classical definition of reliability does not apply to the reliability of observation systems since human raters, who replace tests in this context, are very rarely equivalent or identical in observation skills. In light of this, generalizability coefficients, based on generalizability theory (Cronbach, 1963; Cronbach et al., 1972) have been proposed as a means of determining reliability (Medley & Mitzel, 1963; Rowley, 1976; Tinsley & Weiss, 1975; Linhart, 1979; Cardinet et al., 1976; Erlich & Borich, 1979).

Medley and Mitzel (1963) defined the reliability of a 'category' observation system as "the extent that the average difference between two measurements independently obtained in the same classroom is smaller than the average difference between two measurements obtained in different classrooms" (p. 250). General agreement has been reached on the formula for reliability in terms of population parameters (Frick & Semmel, 1979). The coefficient so obtained is referred to as the intra-class correlation coefficient. The argument for application of generalizability theory to observational data is that agreement between observers is not at issue but rather the ability of the instrument to discriminate across classrooms or between teachers.

Several recent studies, however, have maintained that inter-rater agreement is still an important aspect in the development of observation instruments (Hurwitz,

1973; Emmer, 1973; Frick & Semmel, 1978). Frick and Semmel (1978) noted that if observer agreement is ignored, then a lack of reliability in data collection might in fact be due to a failure of observers to agree on the classification of items within the observation system. Therefore, while inter-rater agreement might not be sufficient to determine reliability it is nevertheless necessary.

There have as well been several notable advancements in methodologies to statistically evaluate inter-rater coefficients of agreement. While earlier system developers merely used percentages of agreement between observers, successive variations on a technique developed by Scott (1955) have led to more exacting measurement criteria. A specific coefficient developed by Cohen (1960) and later extended by Light (1971) is applicable in particular to nominal category systems. This coefficient is made considerably more powerful, if agreement is considered between an observer and a criterion measure, rather than agreement between two observers (Frick & Semmel, 1978). The reasoning here is that agreement between two observers simply indicates how one person coded a particular behavior relative to another. Criterion-related agreement, on the other hand, reflects the degree to which the observer is in agreement with the system itself and with its original category definitions. Information of the latter type would be of much greater value to the developer of an observational system.

This section has discussed some of the concerns related to the development and use of classroom observation scales. The material presented has demonstrated that debate is ongoing concerning several of these issues. The establishment of reliability of observational instruments is one area in particular which a number of instrument developers have neglected.

Summary

This literature review has briefly traced the historical development of observation systems and of research into the relatedness of pupil-teacher behaviors. This development has evidenced a gradual movement toward objective means of observation. As well, there has been a gradual increase in interest in the actual interaction between teacher and child.

The chapter has noted that, in the main, observation systems have been developed to meet specific circumstances and that these systems are not highly adaptable to new applications. No existing systems appeared to be appropriate to this particular study.

Finally, several issues involved in the development of observation systems have been discussed. The next chapter will demonstrate how these and other concerns were addressed in the development of the Classroom Motivation Observation Scale (Glasgow & Spain, 1978).

CHAPTER III

PROCEDURES

The procedures followed in the development and testing of the Classroom Motivation Observation Scale are outlined in this chapter. The material presented here is divided into three main areas of discussion. First, the categories of behavior used in the scale are presented along with some discussion of the sampling and coding procedure. This is followed by a discussion of the training package developed for use with the instrument. Finally, there is a presentation of the measures undertaken to determine the reliability of the observation scale. These divisions are not to suggest that each task was undertaken separately. On the contrary, the development of this instrument must be viewed as an integrated whole.

The Categories of Behavior

The 'social learning model' required a 'bifocal' observation scale which focused on both the student and teacher in interaction within the classroom. Consequently there are two distinct sets of categories; those dealing with pupil behaviors and those dealing with teacher behaviors observed at the same time.

Pupil Behavior Categories

The categories used in the coding of pupil behaviors focused on the attention to on-task behavior of the student. Subsequently, four major types of pupil behavior were identified. Three of these have been divided into more specific pupil behaviors (see Figure 4). In terms of 'conceptual posture', this coding scale viewed the categorization of pupil behavior from the teacher's perspective or from the effect of that behavior on the teacher. A child's behavior was coded 'on-task', for example, if a teacher would accept that behavior as 'on-task'. Similarly, a behavior would be coded as 'disrupting' if it did disrupt the class by causing the teacher to go off-task.

Definition of Major and Sub-Categories of Pupil Behavior

Major categories

I. Attending--Any on-task behavior which cannot be coded as peer-directed or teacher-directed action. Eye or body orientation is directed toward the task or teacher; or the student is otherwise involved in the ongoing classroom activity. Working with pencil and paper, listening to the teacher, laughing at some amusing class incident are examples.

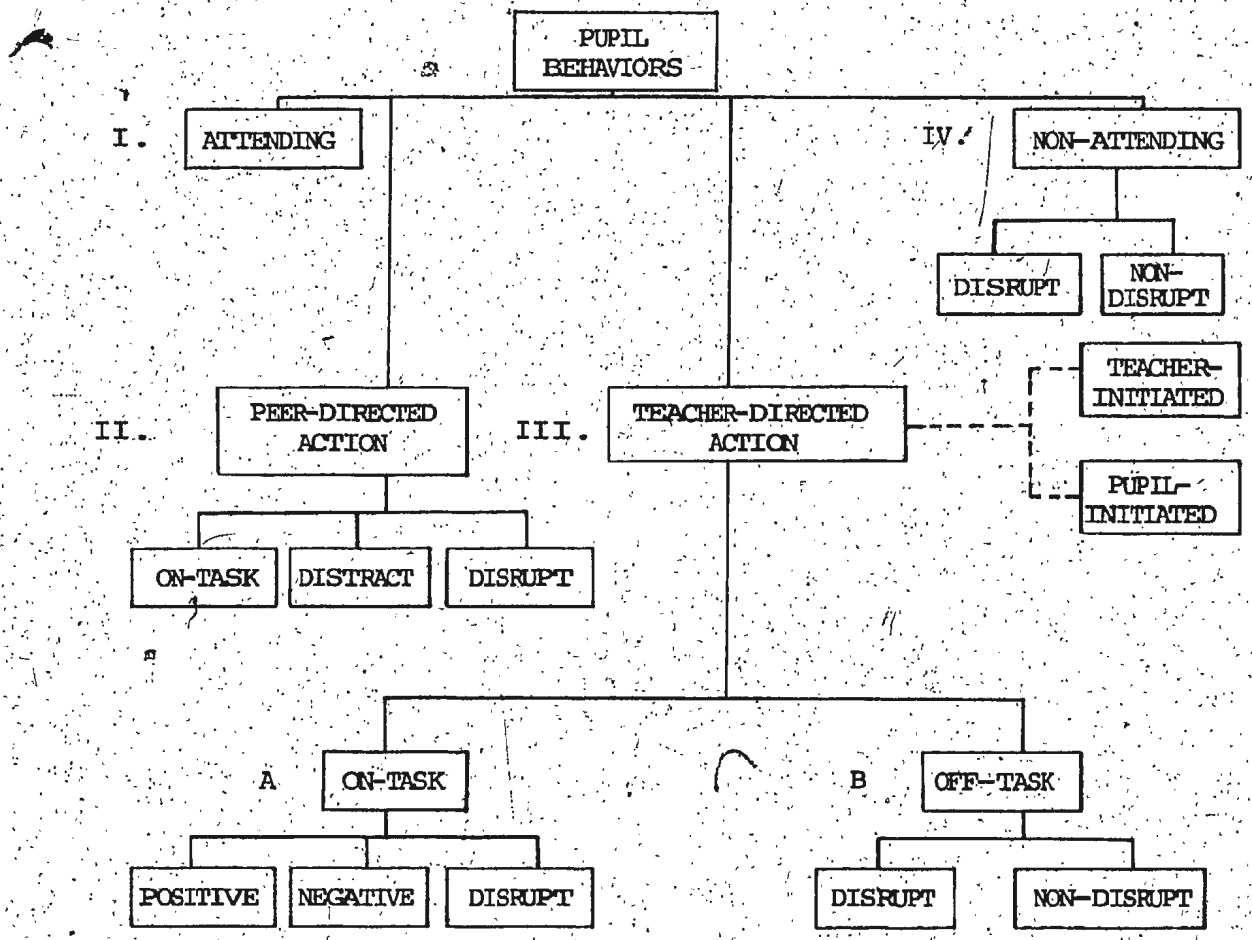

FIGURE 4. Major and Sub-Categories of Pupil Behavior

II. Peer-directed action--Any verbal or non-verbal action directed toward a fellow student or group of students. This category includes all physical acts and verbal interactions or attempts to communicate with peers.

III. Teacher-directed action--Any verbal or non-verbal action directed toward the teacher, including all interaction or attempts at interaction with the teacher.

IV. Non-attending--The student's attention is directed away from the teacher or task and the student does not appear to be involved in on-going class activity. This does not include non-attending behavior which may be coded as teacher- or peer-directed actions. Being turned away from the teacher, and playing with objects on desk, are examples.

Sub-Categories

1. On-task--Any action which pertains to the task or activity intended by the teacher to be of immediate concern to the child in the classroom.

2. Off-task--Any action which is not related to the task or activity of immediate concern in the classroom.

3. Disrupt--Any pupil behaviors which elicit from the teacher an off-task response. Such pupil behaviors may be any one of peer-directed action, teacher-directed action, or non-attention to task.

4. Distract--Any peer-directed action which distracts

a fellow student or group of students from on-task behavior, but which does not elicit an off-task response from the teacher.

5. Non-disrupt--Any off-task, teacher-directed action or non-attending behavior which does not elicit an off-task, teacher response.

6. Positive action--Any teacher-directed action which, from the teacher's point of view, is considered to be a desirable behavior on the part of the student. Examples include raising hand to be recognized, giving correct answer, asking a pertinent question.

7. Negative action--Any teacher-directed action which, from the teacher's point of view is considered to be an undesirable behavior, but which does not elicit an off-task teacher response. Giving an incorrect response, failure to respond, or giving an incomplete answer are examples.

8. Pupil-initiated--A teacher-directed action by the student which occurs when that student is not specifically called upon or designated by the teacher.

9. Teacher-initiated--A teacher-directed action which is the result of a question or command directed by the teacher specifically to the 'target student' (that is, the student under observation).

Teacher Behavior Categories

The categories used in the coding of teacher behavior focused on the motivational aspect of the teacher's behavior and upon the primary targets of that behavior. A particular emphasis of the coding scheme was the determination of levels of motivation employed by a given teacher. Teacher behaviors have been divided into four major categories with two of these categories being further subdivided (see Figure 5).

Definition of Major Categories of Teacher Behavior

I. Non-motivating--Teacher behaviors which are not intended to obtain or reward student participation. Teacher lecturing and any administrative chores are examples.

II. Positive motivation--Any motivating teacher behaviors which directly or indirectly provide for the satisfaction or recognition of student needs. Such behavior may occur on one of three levels. These are: Accepting, Esteem-enhancing, and Interest-providing.

III. Negative motivation--Any motivating teacher behaviors which directly or indirectly deprive the student of needs satisfaction. Such behavior may occur on one of three levels. These are: Non-accepting, Degrading, and Interest-reducing.

IV. Indeterminate motivation--Any motivating teacher

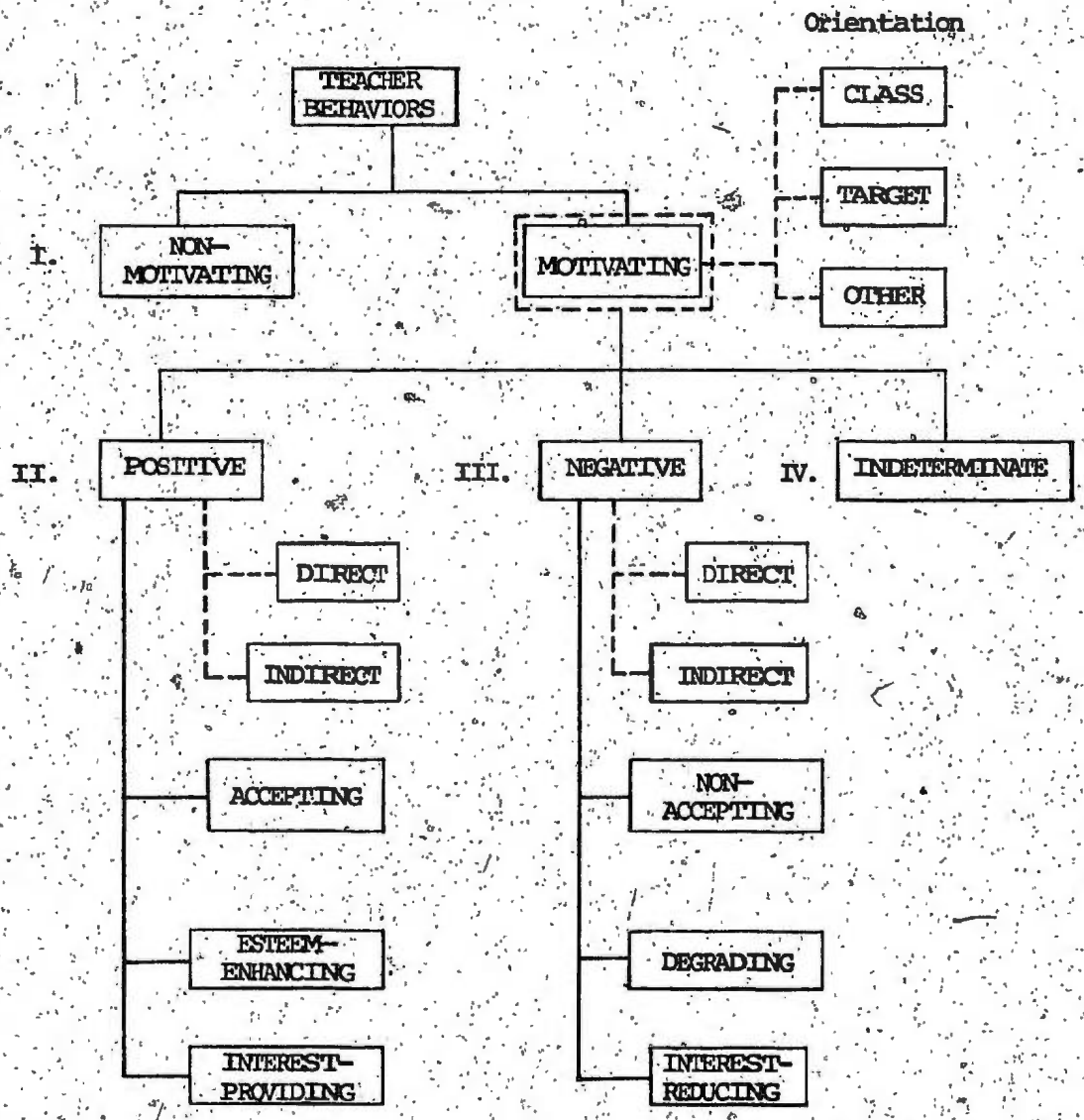

FIGURE 5. Major and Sub-Categories of Teacher Behavior.

behavior which cannot be classified as being either positive or negative, or which cannot be identified as occurring on any of the three levels of motivation under consideration.

The Sub-Categories of Teacher Behavior

Motivating behaviors on the part of the teacher can be viewed in terms of the teacher's manipulation of contingent reinforcers of student behavior. These reinforcers can in turn be viewed as behaviors which provide for or deprive the student of satisfaction of certain developmental needs. The student needs, under consideration here, consisted of three of the levels contained in the hierarchy of human needs identified by Maslow (1962). These needs were: (i) love and belongingness needs, (ii) esteem needs, and (iii) growth needs.

Correspondingly, motivational behavior which attempted to satisfy these needs also occurred on three distinct levels (see Table 9). Love and belongingness needs were satisfied by Accepting behavior on the part of the teacher, while esteem and growth needs were met by Esteem-enhancing and Interest-providing teacher behaviors, respectively. Conversely, Negative reinforcements would include any teacher behaviors which deprived the student of such need satisfaction, or which failed to recognize the existence of these student needs. Teacher behaviors which

TABLE 9

Teacher Motivating Behaviors

Student Needs	Positive Reinforcement	Negative Reinforcement
Love and belongingness	Acceptance	Non-acceptance
Esteem	Esteem-enhancement	Degradation
Growth	Interest-providing	Interest-reducing

frustrated the student's love and belongingness needs are labelled as Non-accepting. The thwarting of satisfaction of esteem and growth needs were, in turn, categorized as Degrading and Interest-reducing teacher behaviors. Provision was made in the scale, as well, for the differentiation of Positive and Negative reinforcement into Direct and Indirect teacher behaviors. This dimension might serve to distinguish those children who had previously learned acceptable behaviors leading to satisfaction of needs and who responded to the teacher as a source of that satisfaction. The sometimes subtle distinction regarding Direct and Indirect reinforcers is demonstrated in Tables 10 and 11.

TABLE 10

Positive Reinforcement

(Examples of Teacher Motivation at Varying Levels)

Level	Direct	Indirect
Acceptance	Helping behaviors, and demonstration of affection, such as smiling, etc.	Recognition of child, acceptance of student comment
Esteem-enhancing	Praise, encouragement	Promise of praise, or esteem building rewards in return for task accomplishment
Interest-providing	Rewarding student with high-interest activity following task completion	Promise of reward, or teacher use of activity to achieve student interest

Definition of Sub-Categories of Teacher Behavior

1. Accepting--Teacher behavior of a generally facilitative nature, involving warmth, positive regard and understanding. The teacher recognizes the student as a person of worth and communicates this recognition to the student.

2. Esteem-enhancing--Teacher behavior of an evaluative nature, aimed at enhancing the student's personal sense of worth or sense of pride in task involvement and accomplishment.

TABLE 11

Negative Reinforcement

(Examples of Teacher Motivation at Varying Levels)

Level	Direct	Indirect
Non-acceptance	Criticism, physical punishments, harsh commands	Impatient looks warning or threatening
Degradation	Evaluative comments, criticism or ability or performance	Threat of future humiliation or degrading action
Interest-reducing	Removal of interesting activities, refusing to allow student to progress when interest is there	Threatening to remove activities of interest to the student

3. Interest-providing--Teacher behaviors aimed at providing interesting and fulfilling activities to students.

4. Non-accepting--Teacher behavior which lacks warmth and understanding and fails to recognize the individual worth of the student.

5. Degrading--Teacher behaviors of an evaluative nature which diminish the student's personal sense of worth and/or sense of task accomplishment.

6. Interest-reducing--Teacher behaviors which reject opportunities to provide interesting activities for students,

or which tend to destroy already existing interest.

7. Direct--Any motivating teacher behaviors which of themselves provide immediate reinforcement to the child for engaging in present or past behavior.

8. Indirect--Any motivating teacher behaviors which serve as cues that direct reinforcement is contingent upon some future student behavior. Teacher statements of a promising or threatening nature are general examples of behavior within this category.

9. Orientation--The intended direction of any motivating teacher behavior. Any particular behavior may be directed toward the class as a whole, toward the target student, or to a student other than the one under observation.

The Sampling and Coding Procedure

The Classroom Motivation Observation Scale employed a time-unit behavioral sampling procedure. A thirty-second coding interval is to be employed with the observer's attention to be focused initially on the student and then on the teacher for a period of from fifteen to twenty seconds in total. In essence, a judgement concerning the student's behavior is to be made and then attention is focused on the teacher. The remaining portion of the half-minute interval is to be used in coding the behaviors in appropriate categories. Coding is done on a specially prepared form which permits coding for a five-minute

interval. Ideally, the observer would code classroom behavior for a five-minute period, take a short break, and once more repeat the cycle. The coding scale has been designed for use in observing several target students within a classroom. In practice, the observer makes an observation of one target student, then the next, and so on, on a rotating basis.

Both sets of behavior categories, pupil and teacher, are considered to be 'facets' and therefore all observed classroom behaviors should lend themselves to categorization within the scale. However, because of the length of the coding interval, it is quite likely that more than one codable response might occur in a given interval. In such circumstances, observers are to follow a particular set of priorities in coding behavior. These priorities in coding were determined to a great extent by the theoretical framework presented in Chapter I, and essentially have placed categories into a hierarchical scale.

Priorities in Coding Behavior

A number of priorities were established for coding classroom behavior. The intent was to capture those behaviors which were considered 'more valuable' either because of their theoretical importance or because of the infrequency with which they occur in classrooms. For example, behaviors to which the teacher makes a response

are given absolute priority since they represent a possible occasion of manipulation of contingent reinforcement within the classroom. The priorities followed in coding pupil behaviors are:

1. Behavior to which the teacher responds is to be coded before any other behaviors.

2. Where teacher response is not a factor, student action, whether Teacher- or Peer-directed is to be given coding priority over Attending and Non-attending behaviors.

3. Should a Teacher-directed and a Pupil-directed action occur during the same observation period, the Teacher-directed action should be coded.

4. If a target pupil engages in separate behaviors which might be coded as Attending and Non-attending, the observer should code the Non-attending behavior.

As well, the following rules were established for the proper coding of teacher behaviors. These are to be applied in the order given here.

1. All teacher behaviors directed toward the Target student are to be coded before any other behaviors.

2. Priority is next given to Negative teacher motivation. Thus, should the teacher respond both negatively and positively in the same interval, the Negative behavior should be coded.

3. The observer will then give priority to Direct as opposed to Indirect teacher motivation.

4. In terms of level of motivation, priority is consistent with the hierarchical position of the need satisfaction. Therefore, Interest-providing, Esteem-enhancing and Accepting will be coded in that order, should two or more occur in the same interval. A similar rule is followed for Negative reinforcement.

The reader is referred to the Coding and Training Manual for the Classroom Motivation Observation Scale (Spain & Glasgow, 1978) for a much more comprehensive treatment of the categories of behavior and of the general coding procedure to be followed.

Development of the Observer-Training Package

Following the recommendations of Thiagarajan (1973) and Brophy et al. (1976) a 'packaged' instructional manual was developed for use with this observation system. The package was meant to provide guidance and efficiency in the training of observers and was intended to be modified and re-designed during use. The package includes manuals for the instructor and observer, practice televised classroom coding sequences, and a video-taped criterion test.

The Video-Taped Criterion Test

Following the advice of Thiagarajan (1973), development began with the criterion test. He argued that working backward from the criterion would make objectives more

concrete and thus provide for a more vigorous instructional package. The criterion test has a twofold function in this situation, in that it is intended for use as a 'final exam', as a means of screening observers; and secondly, the criterion was designed to aid in assessing the reliability of the instrument itself. The criterion test is a videotape of 'edited' classroom interactions which include, as much as possible, the full spectrum of behaviors covered by the category system.

The raw material used in constructing the criterion test was obtained from videotapes of actual classroom interaction. Videotaping was done over a period of four days in a rural area of the Avalon Peninsula. Filming was carried out in two schools and videotapes of teachers and students were produced from five Grade One and Grade Three classrooms. These grade levels were used because it was felt that later application of the instrument would be within the primary grade level. Each classroom was visited on two occasions on separate days for a period of one hour on each occasion.

The actual taping was done using portable television cameras and videotape recorders. The particular cameras used were equipped with built-in microphones which were exceptionally sensitive and provided excellent sound reproduction. Two cameras were actually used in recording the classroom scenes. One of the cameras remained, for

the most part, trained on the teacher, while the other focused on various students in the classroom. The intention was to capture the interactive process including the non-verbal expression of both teacher and student.

The classroom videotapes were then screened by simultaneous playback of both camera angles. Synchronization was made easy by the identical sound tracks of the two tapes. By means of repeated viewings, selected samples of teacher-student behavior were identified and their tape position noted for later inclusion in the criterion test. The criterion tape was produced by 'editing' these small samples of behavior onto a second tape.

The episodes were then edited together. For each of the episodes on the criterion tape, the viewer first focuses on the child and then on the teacher for a total period of from fifteen to twenty seconds. A new episode begins at the end of each thirty-second interval.

A master code was then produced for each of the behaviors on the test tape. This master code was arrived at through a process of discussion and consensus on the part of the instrument developers.

A total of seventy-eight episodes make up the criterion test tape. However, a few of the categories are represented by only a small number of examples. Careful searching of the videotapes revealed no further examples:

One category, 'Teacher-directed off-task nondisrupt', is not represented on the test tape since good quality examples could not be found. The category is, however, dealt with on the training tapes. The categories were retained in the observation scale because they were considered essential. Rowley (1976) has agreed that researchers should not be deterred from measuring behaviors of interest merely because they occur infrequently in most classrooms. Once the completed test tape was in place, compilation of the training manual was begun.

The Training Program

The first section of the training manual presents a very detailed discussion of the coding system and of the categories used in the observation system. An attempt was made to present the material so that the trainees would go through a process of simple to complex skills acquisition. Numerous examples are cited and suggestions made regarding some of the more difficult coding situations.

When satisfied that the observer-trainees are familiar with the categories, the trainer moves on to the next portion of the manual which contains written accounts of classroom interaction. These descriptions were designed to promote discussion concerning the proper coding of behavior. Several of the examples given are borderline and require a thorough knowledge of the coding scheme in

order to be handled well. It was felt that these transcripts gave the coders very good practice in making both the more important and finer distinctions required by the observation system.

Following the work with these written behavioral incidents, the coders are asked to write their own definitions of the behavior categories which they have been using. At this point the instructor should be able to make some determination of the progress of individual trainees.

In the final portion of the training program, the observers are shown videotaped samples of classroom behavior and are once more asked to code these behaviors on a coding form. The videotape presentation was developed together with the criterion tape and the same process was used in compiling it. The training tape, however, is divided into three segments. First, the observers are asked to code only samples of student behaviors. This is followed by a presentation of teacher behaviors and finally, the trainees are asked to code actual interactions between teacher and student. These interactions are presented in the same 'bifocal' manner as described with respect to the criterion tape.

The advantage of using videotaped behavior in training is that the tape can be viewed any number of times so that the trainees may become very familiar with the

coding scheme and process. As in the case of the written episodes, discussion of the coding procedure was considered a vital element in the coding of these videotaped sequences.

The final step in the training of observers is the administration of the criterion test tape. In training observers to carry out actual classroom observation, coders who did not do well on the criterion measure could be retrained or dropped from the training program.

Master codes for both the training and test tapes are appendixed to the instructor's manual. Also accompanying the instructor's manual are written transcriptions of all episodes on these tapes. These were intended for clarification of items should this be needed.

A number of writers, including Medley and Norton (1971), have argued that observer competency should be determined by agreement in coding unambiguous examples of behavior shown on videotape. The development of this observation scale has rejected this as a viable approach and instead has included many borderline samples in both the training and criterion test tapes. This decision was certain to deflate inter-rater reliability to some degree, but it was believed that training observers in this manner was more likely to approximate conditions in real classrooms where behaviors are quite often most ambiguous in terms of categorization.

The Reliability Study

A great deal of confusion has existed regarding the application of tests of reliability to classroom observation measures. The present study has attempted a two-dimensional approach to the reliability question in an effort to measure the ability of the instrument to discriminate between types of classroom behaviors. Thus, the reliability study proposed to measure the observer agreement coefficient based on a comparison of trainee observers' performances on the criterion test with the master code developed for the test. This would determine the agreement of the observers with the categories as defined in the observation system. Secondly, the reliability study proposed to determine generalizability coefficients for the system based on an application of generalizability theory to observations obtained from 'live' classrooms.

Observer Agreement

For the observer agreement component of this study, eleven observers were trained on two different occasions. The trainees were representative of the type of person who might be considered for employment in any research study using the instrument. In educational terms, the trainees ranged from high school graduates to graduates from a Master's degree program in Educational Psychology.

A number of the observers received more training than others. Eight of the observers were trained for a period of ten hours, one observer received fifteen hours of training and the remaining two observers were given twenty-five hours of training over a period of one week. Following training, all observers were asked to independently code the behavior samples contained on the criterion test tape. The data obtained was then analyzed using statistical procedures designed to produce coefficients of criterion agreement.

Statistical Analysis of Agreement Data

A statistical procedure developed by Cohen (1960) and extended by Light (1971) for use with data from observational studies was employed to determine individual coefficients of agreement with the criterion code. Cohen's equation, when applied to a contingency table array of data proposed by Light (1973) has been found to be efficient in determining actual agreement of observation. The 'K' coefficient arising from this approach has been recommended as being a useful statistic, especially when employing a criterion standard of agreement (Frick & Semmel, 1978; Tinsley & Weiss, 1975). This study also employed a 'G' statistic developed by Light (1971) which provides for the collective comparison of a number of observers with a criterion measure, and Kpi, which is a coefficient

developed by Light (1971) to examine the level of agreement of a single category with a criterion.

Generalizability Coefficients

The two observers who underwent training for twenty-five hours over a week long period were employed in this portion of the reliability study. The schools employed in this study were situated in rural areas of the Avalon Peninsula outside St. John's, Newfoundland. A total of nine Grade Two classrooms were visited by the two observers. Eight students were chosen from each class as the target students to be observed. Four of the students selected in each class had been identified as low achievers while the remaining four were considered average achievers.

Observation Procedure

Each classroom was observed for a total of three morning and three afternoon sessions over a six-week period. That is, a total of fifteen hours of observation was done in each classroom, yielding a total of one hundred thirty-five hours of overall observation time. The two observers coded 11,900 frames of information during this time for an average of over 1,300 frames for each teacher or classroom. An observation frame refers to the thirty-second period during which a target student and the teacher were observed.

The observers were given the names of the students to be observed in each classroom as well as a copy of the seating plan which identified each child. Observation focused, in turn, on each target student and the teacher for thirty-second intervals until the cycle was completed. The observer would then begin with the initial student and repeat the cycle. A short break was taken following every five cycles. Observers were located so as to have an unobstructed view of both teacher and students. Since observations had been continuing for several months for another study, the presence of the observer had no noticeable effect on classroom behavior.

Statistical Analysis of Live Classroom Data

An analysis of variance based on generalizability theory (Cronbach et al., 1963) was performed on the data collected during the classroom observations. Generalizability coefficients (Cardinet et al., 1976) were then computed for each of the observation categories based on the design of students nested within classrooms crossed with achievement level. In order to standardize the data, the average frequency each category was observed for a target student was used as the unit of analysis.

Chapter IV presents an analysis of the data generated by this study.

CHAPTER IV

ANALYSIS OF THE DATA

In this chapter, the analysis of data undertaken with respect to the reliability study conducted on the Classroom Motivation Observation Scale is presented. The first section deals with the analysis pertaining to observer agreement. This section is divided into a discussion of the coefficients of agreement (K) for each observer in relation to the total scale and a discussion of the coefficients of agreement (Kp_i) for each category of behavior. The second section is an analysis of the information generated by the study implemented to determine the generalizability of the observation scale. Chapter V contains the summary, conclusions, and recommendations generated by the study.

Observer Agreement

Coefficients of Agreement (K) for Observers on all Categories

The initial step in the analysis of data related to observer agreement was the computation of coefficients of agreement for each observer in relation to the master code for the criterion test. These coefficients were

calculated using Cohen's K for inter-rater agreement (Cohen, 1960). The specific calculation followed Light's extension of K to a tabular array of data (Light, 1973). Consequently, contingency tables comparing each observer with the criterion code were drawn up for all eleven observers. The behavior categories (see Table 12) are displayed along the top and down the left hand side of the matrix, for the observer and criterion code, respectively. One of the contingency tables is given by way of example in Figure 6. This matrix shows the pattern of agreement for the pupil-focus categories based on the coding performed by one observer. Further arrays were produced for the teacher-focus categories.

In essence, K is a function of the number of agreements and disagreements with the criterion, where agreements fall along the main diagonal of the table and expected agreements are equivalent to the row totals for the criterion code. Light (1973) interpreted K as a measure of the distance of disagreement between two observers, or as in this case the distance of the observer from the criterion code. Since the agreement pattern for the observer displayed in Figure 6 is quite good, a fairly high value of K would be anticipated. The calculated value of the agreement coefficient is 0.90.

TABLE 12

Behavior Categories Employed in Construction of Contingency Tables

Pupil Behaviors	Teacher Behaviors
A. Attending	L. Non-motivating
B. Peer-directed action: --on task	M. Indeterminate motivation
C. Peer-directed action: --distract	N. Accepting
D. Peer-directed action: --disrupt	O. Esteem-enhancing
E. Teacher-directed action: --on-task, positive	P. Interest-providing
F. Teacher-directed action: --on-task, negative	Q. Non-accepting
G. Teacher-directed action: --on-task, disrupt	R. Degrading
H. Teacher-directed action: --off-task, disrupt	S. Interest-reducing
I. Teacher-directed action: --off-task, non-disrupt	
J. Non-attending: --disrupt	
K. Non-attending: --non-disrupt	

OBSERVER NO. 11

CRITERION	A	B	C	D	E	F	G	H	I	J	K	
A	18											18
B		5	1									6
C		1	8									9
D				5								5
E					10							10
F						6						6
G							3					3
H								2		1		3
I									0			0
J								1		5		6
K			3								9	12
	18	6	12	5	10	6	3	3	0	6	9	78

K = 0.90

FIGURE 6. Category-agreement matrix for all pupil focus categories: Observer 11 x criterion.

The K coefficient for each observer on all pupil-focus and teacher-focus categories are presented in Table 13. Also included in this table are the Z-values computed for each K as suggested by Light (1971). These values of Z ranged from 6.46 to 20.49 and all exceeded the critical value of Z at any reasonable level of significance. Thus, the null hypothesis of random agreement was rejected and it was concluded that all observed agreements exceeded chance.

The values of K for observer agreement range from .28 to .90 with those observers receiving the greater length of training time achieving the higher agreement levels. Also evident within the table was a slightly higher degree of agreement on pupil-focus as opposed to teacher-focus categories. This finding was quite consistent in that the teacher-behavior categories are generally of a higher inference nature than are the pupil categories. Consistent as well was the greater discrepancy in level of agreement between observers on the teacher-focus categories. Those observers with more training widened the gap between themselves and the others, possibly due to the greater degree of inference required on these categories. A fair degree of variability in agreement coefficients was also seen in Table 13, particularly among the eight observers who received the ten hours of training. The K scores for observers 3 and 5 on the pupil categories were quite low

TABLE 13

Observer Agreement Coefficients for all Behavior Categories

Observer No.	Hours Training	Pupil Focus K	Pupil Focus Z_K	Teacher Focus K	Teacher Focus Z_K
1	10	0.51	11.81	0.38	6.46
2	10	0.54	12.23	0.46	7.71
3	10	0.28	7.34	0.42	7.28
4	10	0.57	12.72	0.49	8.39
5	10	0.31	6.86	0.50	8.60
6	10	0.46	10.62	0.49	7.99
7	10	0.57	12.72	0.54	9.56
8	10	0.58	12.90	0.51	8.74
9	15	0.76	16.84	0.66	11.07
10	25	0.78	17.50	0.84	14.37
11	25	0.90	20.49	0.87	14.93

relative to the other observers in the group.

Interpretation of Cohen's K

The actual interpretation of K has not been very well developed. A review of the literature by Guttman et al. (1971) concluded that .65 was, by consensus, the lower limit of percentage agreement acceptable for research.

This would seem rather low, especially in terms of percentage agreement which was contaminated by chance. More conservative writers (Frick & Semmel, 1978; Tinsley & Weiss, 1975), though they did not recommend specific levels of acceptability for K, indicated that coefficients in the order of .75 to .80 would be sufficient for observer competence. When one recognizes that the K coefficient takes the factor of chance into consideration this would seem to be a most rigorous criterion.

There would appear as well to be some evidence that observer agreement coefficients obtained using videotape are generally lower than such coefficients obtained in live situations (Stallings, 1974; Sandoval, 1976). Such a reduction in reliability coefficients with video recordings would not be surprising in view of the loss of context and much of the non-verbal behavior, especially with edited recordings. Consequently, for K coefficients based on the criterion tape developed for this study, 0.70 or greater in all probability represented an acceptable level of agreement. Assuming that coefficients achieved on videotape represent a lower bound of actual observer agreement, observers who have achieved K coefficients $\geq$ 0.70 would be expected to perform quite well in actual classrooms.

Joint Agreement on all Categories of Behavior

The matrices shown in Figures 7 and 8 present the pattern of agreement found for all eleven observers in relation to the criterion on the pupil and teacher-focus categories, respectively. Some caution should be exercised in interpreting these matrices since the observers differ quite markedly in terms of training received. However, this form of data presentation was considered quite useful, particularly in examining errors in coding. Inspection of the pupil-focus matrix revealed that coders generally categorized pupil behaviors quite well. The agreement shown on categories A (Attending) and E (Teacher-directed action, on-task, positive), for example, was considered quite good. Specific errors in coding were identified on categories B (Peer-directed action, on-task) and G (Teacher-directed action, on-task, disrupt). In the first instance, observers were coding examples of 'Peer-directed, on-task' behaviors as 'Peer-directed, distractions' (category C). The most common error in the latter case was found to be a tendency for coders to place category G behaviors into E and F categories (Teacher-directed positive and negative behaviors, respectively).

Frequent mistakes made in the coding of teacher behavior (Figure 8) occurred on category M (Indeterminate motivation) and categories O (Esteem-enhancing) and R_q (Degrading). Indeterminate motivation was not coded in

ALL OBSERVERS

CRITERION	A	B	C	D	E	F	G	H	I	J	K	
A	160	0	0	0	26	1	0	0	2	4	5	198
B	4	28	26	3	0	0	0	0	1	0	4	66
C	5	6	52	2	4	7	0	3	2	3	15	99
D	1	1	7	23	1	1	5	6	1	9	0	55
E	5	0	0	0	88	6	6	4	0	1	0	110
F	5	0	0	0	27	34	0	0	0	0	0	66
G	0	1	0	0	9	9	13	0	0	1	0	33
H	3	1	0	1	2	0	4	15	3	4	0	33
I	0	0	0	0	0	0	0	0	0	0	0	0
J	1	0	0	1	0	0	5	16	0	38	5	66
K	17	0	9	0	2	0	0	1	12	5	86	132
	201	37	94	30	159	58	33	45	21	65	115	858

G = 40.41

FIGURE 7. Category agreement matrix for all pupil-focus categories: All observers x criterion

ALL OBSERVERS

CRITERION	L	M	N	O	P	Q	R	S	
L	45	4	5	0	0	1	0	0	55
M	17	25	23	3	6	3	0	0	77
N	16	13	127	4	5	10	1	0	176
O	2	2	32	38	3	8	3	0	88
P	1	0	4	1	27	0	0	0	33
Q	13	8	28	2	0	236	26	6	319
R	2	1	4	0	1	31	48	1	88
S	0	0	1	0	2	5	0	14	22
	96	53	224	48	44	294	78	21	858

G = 29.08

FIGURE 8. Category agreement matrix for all teacher-focus categories: All observers x criterion.

many instances where it occurred. Observers tended to classify these behaviors as either 'non-motivation' (category L) or 'accepting' (category N). 'Esteem-enhancement', though recognized as positive reinforcement, was quite often labelled as 'accepting'. A similar error was made in the categorization of 'degrading' behaviors which were misinterpreted as 'non-accepting' teacher behaviors.

Also given in Figures 7 and 8 are the values of G. This statistic was proposed by Light (1973) as a test for 'joint agreement' of N observers with a criterion. For large samples G is approximately normally distributed and is valuable in assessing the collective reliability of observations. The computed values of G were extremely high, resulting in the conclusion that the joint categorizations of all eleven observers agreed with the criterion more than would be expected by chance.

Coefficients of Agreement for Observers on Major Categories

In an effort to determine more clearly whether the observers were in greater agreement with the broader concepts of the observation scale, the eleven pupil-focus and eight teacher-focus categories were collapsed into the four major categories for each as defined in Chapter III. This analysis removed the requirements for finer distinctions among sub-categories and indicated the level of observer

competence for the more broadly defined categories.

Matrices were once more produced for each observer on both pupil and teacher-focus categories (see Figure 9). The resulting K and Z_k scores are presented in Table 14 for both pupil and teacher behaviors. The K coefficients obtained by the collapsing of categories ranged from 0.33 to 0.91 and once again the computed values of Z_k in all cases exceeded normally acceptable levels of significance. Examination of this table indicated that, as was the case for all categories, the K coefficient for the major categories increased with the longer training sessions. Still evident as well was the higher level of agreement on the pupil-focus as compared to the teacher-focus aspect of the scale. The discrepancy of K coefficients for observers 3 and 5 on the pupil-focus categories, relative to the remaining observers, became even more marked.

A comparison of Tables 13 and 14 indicated that the proportional gain in K with the categories collapsed was greater for observers 1 to 8. These were observers who received only ten hours of training.

Joint Agreement for Major Categories

The matrices in Figures 10 and 11 give the combined pattern of agreement for all eleven observers in terms of the major categories of pupil and teacher behavior. As with Figures 7 and 8, this presentation of data was most

OBSERVER NO. 9

CRITERION	Non-motivating	Indeterminate motivation	Positive motivation	Negative motivation	
Non-motivating	4	1	0	0	5
Indeterminate motivation	1	3	3	0	7
Positive motivation	0	3	23	1	27
Negative motivation	0	1	1	37	39
	5	8	27	38	78

K = 0.77

FIGURE 9. Category agreement matrix for major teacher-focus categories: Observer 9 x criterion.

TABLE 14

Observer Agreement Coefficients for Major Categories

Observer No.	Hours Training	Pupil-focus K	Pupil-focus Z_k	Teacher-focus K	Teacher-focus Z_k
1	10	0.66	9.88	0.52	6.53
2	10	0.66	10.01	0.58	7.00
3	10	0.33	4.89	0.57	6.21
4	10	0.71	10.73	0.70	8.30
5	10	0.45	6.74	0.57	6.82
6	10	0.66	9.89	0.69	7.78
7	10	0.69	10.51	0.60	6.73
8	10	0.74	11.33	0.59	6.77
9	15	0.91	14.01	0.77	8.79
10	25	0.90	13.62	0.86	9.78
11	25	0.91	13.88	0.88	10.09

useful in analyzing areas of poor categorization. The G-values indicating reliability of joint agreement remained quite high; therefore, joint agreement was beyond what would be expected by chance.

Examination of these matrices indicated that overall categorization was very good especially on the pupil-focus categories. The only areas of concern detected

ALL OBSERVERS

CRITERION	Attending	Peer-directed action	Teacher-directed action	Non-attending	
Attending	160	0	29	9	198
Peer-/directed action	10	148	31	31	220
Teacher-directed action	13	3	220	6	242
Non-attending	18	10	36	134	198
	201	161	316	180	858

G = 29.86

FIGURE 10. Category agreement matrix for all observers on major pupil-focus categories.

OBSERVER NO. 9

CRITERION	Non-motivating	Indeterminate motivation	Positive motivation	Negative motivation	
Non-motivating	45	4	5	1	55
Indererminate motivation	17	25	32	3	77
Positive motivation	19	15	241	22	297
Negative motivation	15	9	38	367	429
	96	53	316	393	858

G = 20.82

FIGURE 11. Category agreement matrix for all observers on major teacher-focus categories.

from Figure 10 were a tendency for observers to confuse some samples of 'Peer-directed action' and code these behaviors as either 'Teacher-directed' or 'Non-attending', and an apparent problem with some examples of 'Non-attending' behavior which were categorized as 'Teacher-directed' actions. The data in Figure 11 indicated that many coders were experiencing difficulty with the category of 'Indeterminate motivation' and were coding these behaviors as either 'Non-motivating' or 'Positive motivation'. Earlier examination of the data in Figure 8 had shown that this 'Positive motivation' was being coded mainly as 'Accepting behavior'. All of the other remaining major categories of teacher behavior were coded well.

Coefficients of Agreement (Kp_i) for each Category of Behavior

The final stage in the analysis of the data obtained from the criterion test focused on the categories of behavior. Coefficients of agreement for each observer on each category were computed for both major and sub-categories of behavior (see Tables 15, 16, 17, and 18). The coefficients computed were based on a statistic also suggested by Richard Light (1971). This statistic, Kp_i, allows comparison of agreement of each observer's score with a criterion score for any specific category.

Identification of an acceptable level of Kp_i should be attempted only with caution. Little is written concerning

TABLE 15

Agreement Coefficients (Kp_i)* for Major Categories of Pupil Behavior

Observer	Attending	Peer-directed action	Teacher-directed action	Non-Attending
1	0.71	0.47	0.92	0.58
2	0.72	0.41	0.86	0.69
3	0.37	0.53	0.62	0.00
4	0.85	0.63	0.86	0.51
5	0.69	0.28	0.59	0.33
6	0.45	0.63	0.85	0.71
7	0.85	0.51	0.86	0.57
8	0.72	0.57	0.87	0.84
9	1.00	0.81	0.88	1.00
10	0.93	0.86	1.00	0.79
11	1.00	1.00	0.94	0.73

$$*Kp_i = 1 - \left[\frac{1 - \left(\frac{\text{No. of agreements with criterion}}{\text{No. of times category appears on criterion}} \right)}{1 - \left(\frac{\text{No. of times observer coded category}}{\text{Total number of items on criterion test}} \right)} \right]$$

(Light, 1971)

TABLE 16

Agreement Coefficients (Kp_i) for Major Categories of Teacher Behavior

Observer	Non-motivating	Indeterminate motivation	Positive motivation	Negative motivation
1	0.75	0.23	0.55	0.54
2	0.77	0.22	0.60	0.64
3	0.58	-0.01	0.72	0.67
4	0.77	0.38	0.83	0.68
5	0.77	0.40	0.59	0.59
6	0.77	0.12	0.76	0.81
7	0.78	0.12	0.58	0.75
8	1.00	0.03	0.56	0.74
9	0.79	0.36	0.77	0.90
10	1.00	0.53	0.88	0.90
11	0.78	0.84	0.89	0.90

TABLE 17

Agreement Coefficients (Kp_i) for all Categories of Pupil Behavior

Observer	A	B	C	D	E	F	G	H	I	J	K
1	0.71	0.31	0.16	0.17	0.64	0.45	0.29	1.00	-0.01	0.65	0.61
2	0.72	0.16	0.17	0.78	0.62	0.47	0.65	0.32	1.00	0.81	0.60
3	0.37	0.32	0.38	0.38	0.73	-0.03	0.65	0.24	-0.22	0.00	0.00
4	0.85	0.32	0.61	0.19	0.64	0.63	0.31	0.30	-0.01	0.46	0.62
5	0.69	-0.01	0.15	0.00	0.46	-0.01	0.29	0.63	-0.01	-0.03	0.52
6	0.45	0.31	0.50	0.38	0.74	0.29	-0.03	0.32	-0.04	0.62	0.53
7	0.84	0.48	0.62	0.38	0.76	0.46	-0.03	0.29	1.00	0.62	0.44
8	0.72	0.64	0.52	-0.03	0.75	0.47	0.31	0.33	1.00	0.46	0.89
9	1.00	0.31	0.74	0.38	0.89	1.00	-0.33	-0.01	-0.01	0.81	1.00
10	0.93	0.82	0.51	0.79	1.00	0.64	0.32	0.65	1.00	0.82	0.80
11	1.00	0.82	0.87	1.00	1.00	1.00	1.00	0.65	1.00	0.82	0.72

TABLE 18

Agreement Coefficients (Kp_i) for all Categories of Teacher Behavior

Observer	L	M	N	O	P	Q	R	S
1	0.75	0.23	0.40	0.34	0.65	0.42	0.22	0.00
2	0.77	0.22	0.50	0.33	0.65	0.57	0.33	0.00
3	0.58	-0.01	0.70	0.23	0.64	0.39	0.56	0.48
4	0.77	0.38	0.74	0.22	0.65	0.47	0.34	0.48
5	0.77	0.40	0.57	0.47	1.00	0.38	0.33	1.00
6	0.77	0.12	0.49	-0.01	1.00	0.77	0.33	0.49
7	0.78	0.12	0.49	0.22	1.00	0.61	0.71	1.00
8	1.00	-0.03	0.53	0.45	0.66	0.57	0.57	0.49
9	0.79	0.36	0.76	0.47	1.00	0.73	0.46	1.00
10	1.00	0.53	0.76	0.86	1.00	0.90	0.86	1.00
11	0.78	0.84	0.92	0.86	0.65	0.90	0.86	1.00

the effectiveness of the statistic as a means of evaluation. The effect on the computation of the statistic, of the number of times the observer coded a particular category, can be most pronounced. The placement of this quantity in the denominator of the equation (see Table 15) can cause the value of Kp_i to change rather abruptly, especially when the category appears a small number of times on the criterion code. Perfect agreement (Kp_i) occurs whenever the observer codes all examples of a particular behavior correctly. The value of Kp_i will remain 1.0 even if the observer codes examples of other behaviors within this category. Conversely, should an observer make this type of error when all examples of the category in question have been coded incorrectly, the computation of Kp_i will yield a negative coefficient of agreement.

In spite of the volatile nature of Kp_i, it was of use here since it could identify categories which might be poorly defined or might require more extensive treatment in future training sessions. The coefficients displayed in Table 16, for example, would seem to indicate that the category of 'indeterminate motivation' caused some difficulty in categorization. However, the coefficients computed for observers 10 and 11 would indicate as well that much of the problem may be overcome through training. Further investigation of the Kp_i coefficients obtained from the coding of observers 10 and 11 (Tables 17 and 18)

tended to indicate that fairly reasonable levels of agreement could be reached for most categories of the scale, given adequate training. One area requiring further attention on the pupil-focus dimension appeared to be that of 'Teacher-directed on-task actions and teacher-directed off-task actions of a disruptive nature'. These categories (G and H) had the lowest levels of agreement with the criterion. On the teacher-focus behavior categories, the lone area of concern was once more seen to be category M, the category of indeterminate motivation.

Observer Agreement on some Remaining Sub-categories

The remaining coding discriminations were treated separately since they were not categories of behavior per se. These areas of discrimination were: (i) the initiator of a 'Teacher-directed action', (ii) the direct-indirect classification of teacher motivating behavior, and (iii) the orientation of teacher behaviors in terms of their intended direction. The percentage of agreement with the criterion code for each observer on all three dimensions is given in Table 19. Included, as well, are the averages for the two groups of observers based on duration of the training period.

The percentages of agreement were in all likelihood lower than would be expected in actual observations due to the difficulty in relating incidents of behavior

TABLE 19

Percentage of Agreement for Initiator, Orientation, and Direct-indirect Categories

Observer	Initiator	Orientation	Direct-indirect
1	82	62	56
2	68	69	52
3	23	74	47
4	82	78	70
5	46	59	50
6	77	75	47
7	82	82	49
8	82	74	56
9	73	85	50
10	91	97	84
11	82	88	70
Ave. 1 to 9	68	73	53
Ave. 10 & 11	87	93	77
Ave. for all	72	78	57

to the total context within the criterion test. Additionally, instances where the observers failed to code any of these areas were scored as incorrect. Therefore, lowered percentages of agreement may, in some cases, reflect lack of information rather than coder skill.

An examination of Table 19 showed that observers experienced the greatest difficulty in coding the Direct-indirect aspect of teacher motivation. The averages for the two groups, however, did tend to suggest that the longer training period enabled the observers to better discriminate on this category.

Generalizability of the Scale

The classroom observations conducted by observers 10 and 11 provided the data for testing the generalizability of the scale. The analysis was as described by Cardinet, Tourneur and Allal (1976) and yielded the coefficients presented in Tables 20 and 21 for the pupil and teacher categories. The data indicated that more variance was associated with the classroom of observation than with achievement level, suggesting that the observation scale, for most categories, discriminated quite well between classrooms and therefore between teachers. The generalizability coefficients were noticeably higher on the teacher-focus than on the pupil-focus categories. This was as

TABLE 20

Generalizability of Pupil-focus Observations for Classroom/Teacher Means

Pupil Observation Category	σ^2 (class)	σ^2 (error)	p^2 (class)
Attending	17.75	66.95	.734
Peer-directed on-task	4.26	6.19	.872
Peer-directed distract	11.33	15.28	.856
Peer-directed disrupt	0.006	0.12	.889
Teacher-directed on-task, positive	3.41	6.91	.800
Teacher-directed on-task, negative	3.03	0.42	.996
Teacher-directed on-task, disrupt	NOT COMPUTABLE - NO SAMPLES CODED		
Teacher-directed off-task, disrupt	0.05	0.34	.546
Teacher-directed off-task, non-disrupt	0.89	8.204	.464
Teacher-directed pupil-initiated	1.92	4.857	.652
Teacher-directed teacher-initiated	1.91	2.82	.885
Non-attend disrupt	0.03	0.08	0.737
Non-attend non-disrupt	6.78	38.21	0.650

TABLE 21

Generalizability of Teacher-focus Observations for Classroom/Teacher Means

Teacher Observation Category	σ^2 (class)	σ^2 (error)	ρ^2 (class)
Non-motivating	45.927	24.352	.938
Indeterminate	16.790	16.326	.892
Accepting	17.781	11.570	.925
Esteem-enhancing	2.114	3.229	.840
Interest-providing	2.076	0.625	.964
Non-accepting	18.783	8.015	.949
Degrading	0.088	0.303	.700
Interest-reducing	NOT COMPUTABLE - NO SAMPLES CODED		
Indirect	46.255	18.888	.951
Direct	5.768	6.792	.872
Class	27.122	7.337	.967
Other	55.374	18.437	.952
Target	0.00	7.326	0.00

expected since teachers, though having control over student behaviors, would nonetheless tolerate a certain variability of pupil behavior within their classrooms, thereby decreasing variance across classrooms.

The relatively low generalizability coefficients for 'Teacher-directed off-task' behaviors, coupled with equally low Kpi coefficients may indicate some problem of interpretation or definition of these particular categories. The coefficients contained in Table 21 for teacher-behavior categories were exceptionally high and only the coefficients recorded for the 'target' aspect of teacher orientation might be a cause for concern.

Presented in this chapter was the data analysis undertaken to determine the degree of reliability of the coding scale. Observer agreement has been analyzed with respect to: (i) observer agreement with the total scale, and (ii) observer agreement with specific categories. The data generated by the generalizability study has also been presented in the form of coefficients of generalizability for the individual categories.

The final chapter of this report will discuss the implications arising from this analysis of data. Chapter V will also present recommendations for further evaluation of the observation scale and for its use in classroom investigations.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary

The purpose of this study was the development and reliability testing of a classroom observation coding scale. The categories contained in the observation instrument have been presented, and the procedures followed in the development of a packaged training program discussed. The specific measures undertaken to field-test the reliability of observations using the scale were outlined, and the findings resulting from the reliability study were presented in Chapter IV.

The reliability study involved a two-dimensional approach to the determination of reliability of the observation scale. This approach involved the generation of coefficients of agreement to assess the degree to which trained observers could reach agreement with a criterion code based on a videotaped criterion test. The study also entailed the establishment of generalizability coefficients for each category of the scale, based on a series of live classroom observations.

Conclusions

Observers and the coding of behavior. The overall impression gained from the analysis of the categorizations of behavior made by observers in this study was that the behavior categories, as defined, permitted an acceptable level of agreement with the criterion. Generally, the pupil focus categories produced higher levels of agreement across observers. This finding was compatible with the differences in definition of pupil behavior categories as compared to teacher behaviors. The pupil category definitions were of a lower inference level than the teacher categories and one would expect greater ease of categorization with respect to the pupil behaviors. Though the observers who received only ten hours of training did not reach an acceptable standard, their performance on the major categories, taken together with the coefficients of agreement attained by other trainees would suggest that observers can indeed master the intricasies of the various category definitions.

There were nonetheless areas of concern on both the pupil-focus and teacher-focus aspects of the scale. The first problem with the pupil-focus set of categories surfaced during the analysis of the joint agreement of observers with the criterion. The joint agreement data on the major categorizations showed that behaviors within the 'Peer-directed' category were being incorrectly coded

as 'Teacher-directed' and 'Non-attending' behaviors. The analysis of joint agreement further demonstrated that there was confusion as well between the 'on-task' and 'distract' sub-categories of 'Peer-directed action', with 'on-task' items being coded as 'Distract'. However, later analysis of individual categories, based on coefficients of agreement for specific categories, demonstrated that observers who received more than ten hours of training achieved acceptable criterion agreement on the 'Peer-directed' categories. Therefore, this study has concluded that the problem with these categories was due mainly to a lack of adequate training time on the part of some observers.

A similar conclusion was reached in the analysis of the 'Teacher-directed, on-task disrupt' category. In this case, samples of this behavior were being coded as both 'Teacher-directed positive' and 'Teacher-directed negative' behaviors. This error too was shown to be a function of inadequate training since observers receiving additional periods of training did quite well on the category.

A more serious problem apparently existed with the two 'Teacher-directed, off-task' categories. These categories, differentiated as being disrupt and non-disrupt, were a source of confusion for the observers. The extended training received by two of the observers in this study lessened the problem considerably, but not satisfactorily. The relatively low generalizability coefficients (.546 and

.464) recorded for these categories indicated that the difficulties encountered in coding, pointed out by the observer agreement study, were apparently transferred to the observations done in live classrooms. An examination of the guidelines for coding these behaviors, presented in the training manual, indicated that the problem may have resulted from a lack of understanding of the original definitions. The conceptualization of these categories was complex and greater attention should perhaps have been paid to the proper interpretation of these definitions during observer training. A further problem with the 'Teacher-directed off-task nondisrupt' category was that very few examples were discovered for training purposes. No samples of this behavior were considered of good enough quality for inclusion on the test tape.

On the teacher-focus section of the observation scale, the category of 'Indeterminate motivation' caused the greatest difficulties in coding. The problem appeared to be mainly one of interpretation. There was a tendency on the part of observers to code samples of this behavior as either 'Non-motivating' or 'Accepting' behaviors. The category was included on the scale during the development of the instrument, because certain teacher behaviors defied categorization as either positive or negative reinforcement, even though they appeared to have motivational content. Observers apparently differed considerably in their ability

to categorize indeterminate behavior.

Examination of the individual agreement coefficients for the major categories of teacher behavior showed that observers differentiated quite well between positive and negative reinforcement and, in general, between behaviors which would be classified as motivating or non-motivating. There appeared, as well, to be no pattern linking ability to differentiate on these dimensions with ability to classify teacher motivating behavior as indeterminate. A possible answer might be found in the coding of observers who had the greatest amount of teaching experience. These observers, although recording fairly strong agreement on the other major categories of teacher behavior, recorded some of their lowest coefficients of agreement for indeterminate motivation. There may be a tendency for experienced teachers to recognize elements within these teaching behaviors which, in the overall teaching context, might be considered by them to be either a form of positive motivation or to be non-motivating. This would of course introduce a further inferential bias into the recording of observed behavior and might have implications for subsequent observer selection.

The data collected from live classrooms tended to support the original decision to include this category. This data did indicate that, though observers might differ in the proportion of teacher behaviors assigned to this

category, the category was generalizable across teachers. Examination of the agreement coefficients for the category of 'Indeterminate motivation', on the basis of training received, also indicated that the difficulty might be overcome through more extensive training.

Extended training also appeared to offer a solution to the difficulties experienced with regard to the coding of 'Esteem-enhancing' and 'Degrading' behaviors as well as in the categorizing of teacher motivation as 'Direct' or 'Indirect'. Each of these discriminations involves a high degree of inference and hence complexity of definition. Those observers who received the longer periods of training showed a marked ability to correctly code these behaviors relative to the remaining observers. In addition, the generalizability coefficients for these behavior categories were acceptable.

Observers and observer training. The data analysis consistently indicated that increased training time had a marked effect on observer skill in coding. This was especially so for the observers who received twenty-five hours of training and to a lesser degree for the observer who underwent a fifteen-hour training period. The collapsing of the behavior categories, in particular, demonstrated that the observers who received the shortest period of training were most likely deficient in making the finer distinctions required by the instrument. This appeared so

since the proportional gain in coding agreement was much greater for those observers who received only ten hours of training in relation to the others. A less obvious conclusion which might be drawn from this effect was that the lower proportional gain by the more highly trained observers indicated that those observers might have been approaching an upper limit of reliability. Should this be so, further training (beyond twenty-five hours) may not produce appreciable gain in coding ability.

There was also reason to believe that the coefficients of agreement reached on the major categories of behavior might in some cases be more indicative of agreements which would be achieved by observers in actual classroom observations. Many of the distinctions required in subcategory definition might be more easily made in live classrooms than from an edited videotape where the context of a given situation might be quite disjointed.

The background of observer-trainees could have had some effect on the effectiveness and patterns of coding behaviors, particularly teacher behaviors. Those observers who were experienced teachers might have inferred the intent of certain teacher behaviors as a result of their own teaching experience. Additionally, there was a most noticeable tendency during training for these same coders to question the correctness of codes given by the training manual. This was especially so with regard to some forms

of negative motivation and in the determination of levels of reinforcement, such as the discrimination between non-accepting and degrading behavior. Teachers tended to experience greater difficulty in categorizing behaviors as degrading than did non-teacher trainees. The data did indicate, however, that most observers overcame these problems with training. One teacher-observer did continue to experience difficulty with coding degrading behavior as non-accepting.

Conversely, the trainees with teacher training and/or experience were much more attuned to the terminology used in much of the observer training manual. These observers, therefore, presented much less difficulty in terms of interpretation than did those observers who had no professional educational training. Thus, experienced teachers, while presenting problems of inferential bias in coding, were more easier trained than other observers and did become relatively good coders.

The ten-hour training period given the initial eight observers also demonstrated that a short training period of this nature might be sufficient in determining those persons best suited for further training and subsequent employment as classroom observers. An examination of the results achieved by these trainees showed that two of the observers were quite deficient in coding skills as compared to the remaining observers. One might conclude,

based on this difference, that these particular trainees might experience great difficulty in achieving an acceptable level of coding skill. The economics of extended training would of course be at issue in making any decisions based on this tentative finding.

No conclusion can be reached, based on this analysis, of the effectiveness of this approach to training observers relative to other forms of training. However, this training program appeared to be effective, on the occasions of use cited, for the majority of observers participating. The training program may well have been more effective had more emphasis been placed on certain areas as previously noted. Several areas involving complex interpretation, owing to their conceptual definition, appeared to require additional training time.

Reliability of observation using the scale. The generalizability coefficients obtained from the analysis of data provided by the classroom observations, coupled with the excellent K coefficients registered by the observers involved, would indicate that observations recorded using this instrument are reliable for purposes of classroom research. The generalizability coefficients were most encouraging. The lowest recorded coefficients on either focus of the scale were those computed for the two 'Teacher-directed, off-task' categories. The much lower coefficients on these categories were obviously related to the difficulties

encountered in coding discovered by the observer agreement study.

The generalizability coefficients given for the teacher-focus categories of the scale contain an anomaly with respect to the coefficient computed for the 'Target' aspect of 'Orientation'. Interpretation of the coefficients of generalizability across classrooms and achievement groups, in this case, has yielded a coefficient of 0.0. Given the design of the instrument this finding was expected, as target students are also 'individual'. 'Class' and 'Other' components of this discrimination were generalizable across teachers, but the 'Target' aspect was not since it was not a condition of teaching per se. The 'Class' and 'Other' codes, on the other hand, appeared to be differentiating between teachers who interacted more with the whole class and those who interacted more with individual students. The 'Target' aspect was generalizable across achievement groups and this was consistent with the design of the study, where target students were found within achievement groups.

Overall, the generalizability coefficients for the various categories, the majority of which were greater than .85, indicated that the scale as a whole was indeed generalizable to teacher and student classroom behaviors and did, therefore, provide reliable information.

Recommendations

1. Subsequent use of this observation scale should ensure that adequate emphasis during training is placed on those areas of confusion in coding identified in this study.

2. Future training sessions should consider the implication of possible coder bias resulting from the background and training of observer-trainees. A direct attempt to investigate the effects of observer background might be most beneficial.

3. Further attempts to demonstrate the reliability of this instrument should consider using a longer training period to ascertain whether more training will produce appreciable gain in coding skill among observers.

4. Another consideration related to realiability of observer coding which should receive future attention is that of stability and the effects, if any, of coder 'drift' following training in the use of the scale.

5. The development of an expanded training package and a more comprehensive test tape would be a definite asset in the training and evaluation of observers. Such an expansion should endeavour to include more examples of each category of behavior and attempt to improve the overall quality of video production especially.

6. Since this study has demonstrated the reliability of observation provided by this observational instrument,

it is recommended that the instrument continue to be used in the examination of the theoretical implications posed in the opening chapter of this report.

7. The training package itself should be examined as to its effects on the classroom behavior of practising teachers. Several other observational instruments have been found to be useful in teacher training and professional development programs. A number of instances encountered during the training sessions have indicated that this instrument might also be useful as a training tool in the modification or reinforcement of teacher behavior.

REFERENCES

Ackerman, W.I. Teacher competence and pupil change. Harvard Educational Review, 1954, 24, 213-289.

Amidon, E., and Hunter, E. Improving Teaching. New York: Holt, Rinehart and Winston, 1967.

Anderson, H.H.; Brewer, J.E.; and Reed, Mary F. Studies of teachers' classroom personalities. III. Follow-up studies of the effects of dominative and integrative contacts on children's behavior. Applied Psychological Monographs, 1946, No. 11.

Anderson, H.H. The measurement of domination and of socially integrative behavior in teachers' contacts with children. In Edmund J. Amidon and John B. Hough (eds.), Interaction Analysis: Theory, Research and Application. Reading, Mass: Addison-Wesley, 1967, 4-23.

Aspy, D. The effects of teacher-offered conditions of empathy, positive regard and congruence upon student achievement. Florida Journal of Educational Research, 1969, 2, 39-48.

Aspy, D. Towards a Technology for Humanizing Education. Champaign, Illinois: Research Press Company, 1972.

Aspy, D., and Roebuck, F.N. Kids Don't Learn from People They Don't Like. Amherst, Massachusetts: Human Resources Development Press, 1977.

Bandura, Albert and Walters, Richard. Social Learning and Personality Development. New York: Holt, Rinehart and Winston, 1964.

Biddle, B.J., and Adams, R.S. An Analysis of Classroom Activities. Columbia, Missouri: Center for Research in Social Behavior, University of Missouri, 1967.

Biddle, B.J. Methods and concepts in classroom research. Review of Educational Research, 1967, 37, 337-357.

Borich, Gary and Madden, Susan. Evaluating Classroom Instruction: A Sourcebook of Instruments. Reading, Mass: Addison-Wesley, 1977.

Boyd, R.D. and De Vault, M.N. The observation and recording of behavior. Review of Educational Research, 1966, 36, 529-551.

Braun, Carl. Teacher expectation: Sociopsychological dynamics. Review of Educational Research, 1976, 46, 185-213.

Brophy, J.E., and Good, T.L. Teacher-child Dyadic Interaction: A Manual for Coding Classroom Behavior. Austin, Texas: Research and Development Center for Teacher Education, The University of Texas, 1969.

Brophy, J.E., and Good, T.L. Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. Journal of Educational Psychology, 1970, 61, 365-374.

Brophy, Jere E.; Everston, Carolyn M.; Anderson, Linda M.; Baum, Michael C.; and Crawford, John. Criterion-referenced observational measurement in the classroom. (Mini-training course presented at the annual meeting of the American Educational Research Association, 1976).

Cardinet, Jean; Tourneur, Yvan; and Allal, Linda. The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-135.

Carkhuff, R.R., and Truax, C.B. Towards Effective Counselling and Psychotherapy. Chicago: Aldine, 1967.

Cobb, J. Relationship of discrete classroom behavior to fourth grade academic achievement. Journal of Educational Psychology, 1972, 63, 74-80.

Cohen, J.A. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cronbach, L.J.; Rajaratnam, N.; and Gleser, G. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Cronbach, L.J.; Gleser, G.C.; Nanda, H.; and Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.

Dalton, W.B. The relationship between classroom interaction and teacher ratings of pupils: An exploration of one means by which a teacher may communicate her expectancies. (Paper presented at the annual meeting of the Southeastern Psychological Association, New Orleans, 1969).

de Groat, A.F., and Thompson, G.G. A study of the distribution of teacher approval and disapproval among sixth grade pupils. Journal of Experimental Education, 1949, 18, 57-75.

Dunkin, M.J., and Biddle, B.J. The Study of Teaching. New York: Holt, Rinehart and Winston, 1974.

Emmer, Edmund T. Direct observation of classroom behavior. International Review of Education, 1972, 18, 473-490.

Erlich, Oded and Borich, Gary. Occurrence and generalizability of scores on a classroom interaction instrument. Journal of Educational Measurement, 1979, 16, 11-18.

Fisher, C.W.; Filby, Nikola N.; Marliave, R.; Cohen, L.S.; Dishaw, M.M.; Moore, J.E.; and Berliner, D.C. Beginning Teacher Evaluation Study. Far West Laboratory for Educational Research and Development, San Francisco, California, 1978.

Flanders, N.A. Analyzing Teaching Behavior. New York: Addison-Wesley, 1970.

Fox, Ronald B., and Peck, Robert F. Personal characteristics of teachers that affect student learning. (Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978).

Frick, Ted and Semmel, Melvin T. Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 1978, 48, 157-184.

Glasgow, Conrad B., and Spain, William H. Coding and training manual for the Classroom Motivation Observation Scale. Institute for Educational Research and Development, Memorial University of Newfoundland, 1978.

Good, T.L. Which pupils do teachers call on? Elementary School Journal, 1970, 70, 190-198.

Good, T.L., and Brophy, J.E. Teacher-child Dyadic Interactions: A New Method of Classroom Observation. Austin, Texas: Research and Development Center for Teacher Education, The University of Texas, 1972.

Good, Thomas L., and Brophy, Jere E. Looking in Classrooms. New York: Harper and Row Publishers, 1973.

Guttman, H.A.; Spector, R.M.; Sigal, J.J.; Rakoff, V.; and Epstein, N.B. Reliability of coding affective communication in family therapy sessions: Problems of measurement and interpretation. Journal of Counselling and Clinical Psychology, 1971, 37, 397-402.

Herbert, J., and Attridge, C. A guide for developers and users of observation systems and manuals. American Educational Research Journal, 1975, 12, 1-20.

Hoehn, A.J. A study of social status differentiation in the classroom behavior of nineteen third grade teachers. Journal of Social Psychology, 1954, 39, 269-292.

Hurwitz, Richard F. The reliability and validity of descriptive-analytic systems for studying classroom behaviors. Classroom Interaction Newsletter, 1973, 8, 50-59.

Jackson, P.W.; Silberman, M.L.; and Wolfson, B.S. Signs of personal involvement in teachers' descriptions of their students. Journal of Educational Psychology, 1969, 60, 22-27.

Johannesson, Ingvar. Effects of Praise and Blame. Stockholm, Sweden: Almovist and Wiksell, 1967.

Karafin, Gail R. Discussion of considerations for selecting or developing an observational system. Classroom Interaction Newsletter, 1973, 8, 15-32.

Keough, Lorraine. The relationship of achievement and sociometric status to classroom behaviors of grade two students. Unpublished Master's thesis, Memorial University of Newfoundland, August 1980.

Klein, S. Student influence on teacher behavior. American Educational Research Journal, 1971, 8, 403-421.

Lahaderne, H. Attitudinal and intellectual correlates of attention: A look at four sixth grade classrooms. Journal of Educational Psychology, 1976, 68, 320-324.

Light, R.J. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 1971, 76, 365-377.

Light, R.J. Issues in the analysis of qualitative data. In R.M.W. Travers (ed.), Second Handbook of Research on Teaching: A Project of the American Educational Research Association. Chicago: Rand-McNally, 1973.

Linhart, Cynthia A. Application of Generalizability Theory to a Complex Rating Situation. (Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979).

Long, John V. Media effects upon classroom verbal interaction data. Classroom Interaction Newsletter, 1974, 10, 3-12.

Marliave, Richard. Observable classroom variables. (San Francisco, California: Technical Report 1-2, Beginning teacher evaluation study, Far West Laboratory for Educational Research and Development, 1976).

Martin, Jack. The development and use of classroom observation instruments. Canadian Journal of Education, 1977, 2, 43-54.

Maslow, A.H. Toward a Psychology of Being. Toronto: Van Nostrand Company, 1962.

McKinney, J.; Mason, J.; Perkerson, K. and Clifford, M. Relationship between classroom behavior and academic achievement. Journal of Educational Psychology, 1975, 67, 198-203.

Medley, D.M. Early history of research on teacher behavior. International Review of Education, 1972, 18, 430-439.

Medley, D.M., and Mitzel, H.E. A technique for measuring classroom behavior. Journal of Educational Psychology, 1958, 49, 86-92.

Medley, D.M., and Mitzel, H.E. Measuring classroom behavior by systematic observation. In Gage, N.L. (ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963, 247-328.

Medley, D.M., and Norton, D.P. The concept of reliability as it applies to behavior records. (Paper presented at the meeting of the American Psychological Association, Washington, D.C., 1971).

Mitchell, J.V. Education's challenge to psychology: The prediction of behavior from person-environment interactions. Review of Educational Research, 1969, 39, 699-710.

Morsh, J.E., and Wilder, E.W. Identifying the effective instructor: A review of the quantitative studies, 1900-1952. (USAF Personnel Training Research Centre, Research Bulletin No. AFPTRC - TR - 54 - 44, 1954).

Nuthall, Graham, and Church, John. Observation systems used with recording media. International Review of Education, 1972, 18, 491-507.

Perkins, H. Classroom behavior and underachievement. American Educational Research Journal, 1965, 2, 1-12.

Rajaratnam, Nageswari. Reliability formulas for independent decision data when reliability data are matched. Psychometrika, 25, 261-271.

Redd, W.H.; Morris, E.K.; and Martin, J.A. Effects of positive and negative adult-child interaction on children's social preference. Journal of Experimental Child Psychology, 1975, 19, 153-164.

Reed, H.B. Effect of teacher warmth. Journal of Teacher Education, 1961, 12, 330-334.

Rogers, C.R. Client Centered Therapy. Boston: Houghton-Mifflin, 1951.

Rosenfeld, H.M. Nonverbal reciprocation of approval: An experimental analysis. Journal of Experimental Social Psychology, 1967, 3, 102-111.

Rosenshine, Barak. Teacher behaviors related to pupil achievement. Classroom Interaction Newsletter, 1969, 5, 4-17.

Rosenshine, Barak. Evaluation of Instruction. Review of Educational Research, 1970, 40, 279-300.

Rosenshine, Barak. Teaching behaviors related to pupil achievement: A review of research. In Ian Westbury and Arno A. Bellack (eds.), Research into Classroom Processes. New York: Teacher's College Press, 1971, 51-117.

Rosenshine, Barak. Classroom Instruction. The National Society for the Study of Education. Seventy-Fifth Yearbook, Chicago, The University of Chicago Press, 1976.

Rowley, Glenn L. The reliability of observational measures. American Educational Research Journal, 1976, 13, 51-59.

Sandoval, Jonathan. The evaluation of teacher behavior through observation of videotape recordings. Beginning teacher evaluation study: Phase II, Final report: Volume III. California State Commission for teacher preparation and licensing. Sacramento, California, 1974.

Sarbin, T.R., and Allen, V.L. Increasing participation in a natural group setting: A preliminary report. The Psychological Record, 1968, 18, 1-7.

Scott, W.A. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-325.

Silberman, M.L. Behavioral expression of teachers' attitudes toward elementary school students. Journal of Educational Psychology, 1969, 60, 402-407.

Simon, Anita. The effects of training in analysis on the teaching patterns of student teachers in favored and

non-favored classes. Dissertation Abstracts International, 1967, 24, 2716A. (University Microfilm No. 67-3164)..

Simon, A., and Boyer, E. (Eds.), Mirrors for Behavior III: An Anthology of Classroom Observation Instruments. Philadelphia, Penn: Research for Better Schools, Inc., 1974.

Soar, Robert S. Optimum teacher-pupil interaction for pupil growth. Classroom Interaction Newsletter, 1969, 5, 38-45.

Soar, Robert S. A measure of teacher classroom management. (Paper presented at a symposium titled, "Observational methods for studying preschool environments," at the American Psychological Association Meeting in Washington, D.C., 1971).

Soar, Robert S. Teacher behavior related to pupil growth. International Review of Education, 1972, 18, 508-526.

Soar, Robert; Soar, Ruth; and Ragosta, Marjorie. Florida Climate and Control System. Gainesville, Florida: Institute for Development of Human Resources, University of Florida, 1971.

Soli, S., and Devine, V. Behavioral correlates of attention: A look at high and low achievers. Journal of Educational Psychology, 1976, 68, 103-116.

Stallings, Jane A. Follow-through Program: Classroom Observation Evaluation 1971-72. Stanford Research Institute, Menlo Park, California, 1973.

Stallings, Jane A. Implementation and child effects of teaching practices in follow-through classrooms. Monographs of the Society for Research in Child Development, 1975, 40, 1-133.

Stallings, Jane A., and Giesen, Phillip A. A study of confusability of codes in observational measurement. (Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974).

Stukat, C.G., and Engstrom, R. T.V. observations of teacher activities in the classroom. Pedagogisk Forskning, 1967, 11, 96-117.

Thomas, D.R., Becker, W.C.; and Armstrong, M. Production and elimination of disruptive classroom behavior by systematically varying teacher behavior. Journal of Applied Behavior Analysis, 1968, 1, 35-45.

Thurstone, L.L. The Measurement of Social Attitude. Chicago: University of Chicago Press, 1936.

Tinsley, Howard E.A., and Weiss, David J. Inter-rater reliability and agreement of subjective judgements. Journal of Counselling Psychology, 1975, 22, 358-376.

Turner, R.L., and Denny, D.A. Teacher characteristics, teacher behavior and changes in pupil creativity. Elementary School Journal, 1969, 69, 265-270.

Turpin, Edna. The relationship of teacher use of different reinforcement patterns to the self-concept development of second grade children with and without learning problems. Unpublished Doctoral Dissertation, The University of Maine at Orono, 1981.

Wittmer, J., and Myrick, R.D. Facilitative teaching: Theory and practice. California: Goodyear, 1974.

Witty, P.A. An analysis of the personality traits of the effective teacher. Journal of Educational Research, 1947, 40, 662-671.

Woolfolk, Anita E. Student learning and performance under varying conditions of teacher verbal and nonverbal evaluative communication. Journal of Educational Psychology, 1978, 70, 87-94.

APPENDIX A

COHEN'S K

Cohen's K* coefficient is given by the equation

$$K = \frac{P_o - P_e}{1 - P_e}$$

$$\text{where } P_o = \frac{1}{N} \sum_{i=1}^{c} n_{ii}$$

$$\text{and } P_e = \frac{1}{N^2} \sum_{i=1}^{c} (n_{i+})\ (n_{+i})$$

i) N = total number of items

ii) n_{ii} = total number of agreements for the ith category (main diagonal in tabular array)

iii) n_{+i} = marginal for the observer on the ith category

iv) n_{i+} = marginal for the criterion coder on the ith category

*Adapted from Frick and Semmel (1978, 170).