ED 021 809 56 SP 001 582

By- Schalock, H. Del; Beaird, James H.

INCREASING PREDICTION OF TEACHERS' CLASSROOM BEHAVIOR THROUGH USE OF MOTION PICTURE TESTS. FINAL REPORT.

Oregon State System of Higher Education, Monmouth. Teaching Research Div.
Spons Agency- Office of Education (DHEW), Washington, D.C. Bureau of Research.
Bureau No- BR-5-0836
Pub Date Jul 68
Grant- OEG-7-47-9015-284
Note- 101p.
EDRS Price MF-$0.50 HC-$4.12
Descriptors- ATTITUDE TESTS, CORRELATION, EDUCATIONAL EXPERIMENTS, ELEMENTARY SCHOOL TEACHERS, *FILMS, MEASUREMENT TECHNIQUES, OCCUPATIONAL TESTS, *PREDICTIVE MEASUREMENT, RESPONSE MODE, STATISTICAL ANALYSIS, STUDENT TEACHERS, *TEACHER BEHAVIOR, TEACHER RESPONSE, *TEST CONSTRUCTION
Identifiers- Minnesota Teacher Attitude Inventory, MTAI

The major portion of this report reviews a recent study that replicated and extended earlier research by Schalock, Beaird, and Simmons (1964, ED 003 620) on the use of situation reaction tests (using motion picture representations of classroom situations as test stimuli) to predict teaching behavior in the classroom. Chapters I and II summarize the 1964 study which pretested 40 student teachers using a paper and pencil attitude scale plus 3 situation reaction tests with different response modes; scores were correlated with observational measurements of management behavior in the classroom during student teaching. Chapter III reports the replication study which was extended to include situational data in the prediction scheme and which studied 39 experienced teachers in addition to 39 student teachers. In Chapters V and VI the results are discussed with reference to test theory, observational research methodology, and the measurement of teaching behavior in situation. Included are 20 statistical tables, a 16-item bibliography, and an example of the behavior profiles prepared for each teacher who took part in the study. A supplementary document to this report, "An Overview of the Teaching Research System for the Description of Teaching Behavior in Context" (developed for use in this study), is also in the ERIC collection. (JS)

FINAL REPORT
Project No. 5-0836
Grant No. OE-7-47-9015-284

INCREASING PREDICTION OF TEACHERS' CLASSROOM BEHAVIOR THROUGH USE OF MOTION PICTURE TESTS

H. Del Schalock
James H. Beaird

July 1968

teaching research

OREGON STATE SYSTEM OF HIGHER EDUCATION

R-68

FINAL REPORT

Project Number 5-0836
Grant Number OE-7-47-9015-284

INCREASING PREDICTION OF TEACHERS' CLASSROOM BEHAVIOR THROUGH USE OF MOTION PICTURE TESTS

H. Del Schalock
James H. Beaird

TEACHING RESEARCH
A Division of the Oregon State System of Higher Education
Monmouth, Oregon

July, 1968

The research reported herein was performed pursuant to a grant with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

Acknowledgments

Success of research projects carried out within the context of teacher education programs and the public schools depends upon the cooperation, effort and good will of a large number of people. Approximately 60 student teachers from two teacher education institutions within the State, their supervisors, department chairmen, deans of cooperating schools, and approximately 60 experienced teachers from twenty school districts, along with their principals and superintendents, shared these qualities. Their contribution to and support of the project is greatly appreciated. Special thanks also go to project staff: Mrs. Karen Davidson, Miss Margaret Faulkner, Mr. Ken Harrison, Mr. Robert Lange, Mr. Sid Micek, Mrs. Anne Rose, Mrs. Joyce Schalock, Mrs. Dorothy Sperling, and Mrs. Virginia Weigel; and to the data reduction staff. Without their continuous support and extra effort the management of the project would have been a much more difficult task than it was.

H. Del Schalock
James H. Beaird

TABLE OF CONTENTS

LIST OF TABLES

CONTINUED

LIST OF FIGURES

Chapter I

INTRODUCTION

Recently completed research by Schalock, Beaird and Simmons (1964) on the predictive power of tests which use motion pictures as test stimuli suggests that a methodology may now be at hand which will permit the prediction of teaching behavior in the classroom. Using student teachers as subjects, Schalock et al. were able to demonstrate multiple correlations of .69 to .87 between scores on a battery of situational-response tests (tests which use motion picture representations of classroom situations as test stimuli) administered prior to student teaching and observational measures of their behavior in the classroom during student teaching. This represents an unusual accomplishment, for typically studies in the behavioral sciences have not been able to account for more than 50 per cent of the variance in any criterion that has been predicted to, and when the criterion has been as complex as teaching behavior the level of prediction has nearly always been less. In the Schalock, Beaird and Simmons study at least 50 per cent of the variance was accounted for in each of the 15 separate criterion measures used (concrete behavior of teachers in the classroom) and as much as 75 per cent of the variance was accounted for in some.

Unfortunately, several factors tend to temper the confidence that can be placed in the findings that came from the study. First, a small N (40) coupled with a relatively large number of predictor variables (18) could have led to the multiple correlations being spuriously high.

Cronbach (1960) has warned that validity shrinkage is likely to be great from one study to another when many predictors are tried and when weights are determined from small samples. Dunn (1959) has gone so far as to say that multiple regression methodology, the strategy of analysis used in the Schalock, Beaird and Simmons study, is not a particularly reliable methodology. She attempted to predict to choice of field of study and success in it, using grades as criteria of success, and found that for her first sample (N=approximately 500) multiple R's ranged from .416 to .914, but in a cross-validation group, using the same predictors to the same measures, found correlations of -.433 to .160. These data, in combination with the large number of predictors and the relatively small number of subjects used in the Schalock _et al_. study, make the correlations coming from it suspect. In defense of the study, however, the number of predictors never exceeded N/2, a commonly applied rule of thumb in studies of this kind.

A second factor that leads to a tempering of confidence in the data stems from the somewhat unorthodox analyses applied to it. Seventy-five regression analyses were run, 15 (one for each of the criterion measures employed in the study) using _total test scores_ from each of the four instruments employed in the prediction battery, 45 using the _subscale scores_ found within three of these instruments (the four tests used in the study were made up of 1, 7, 11 and 12 subscales respectively), and 15 using a _combination of the best predictors_ from each of the four tests in the battery as these were identified in the subscale analyses.

While the full range of data were reported for the various analyses (see Chapter II), there is some question as to what to make of them. There is also some question as to the appropriateness of the procedure used in selecting the best of the subscale predictors for inclusion in the final set of regression analyses. Subscales were selected on the basis of per cent of criterion variance accounted for and it may have been more appropriate to select on the basis of the correlation of subscales with the criterion measures and other subscales. In any event, either or both of these factors could hav· caused spuriously high correlations to appear between predictor and criterion measures.

In contrast to the sources of error in the data that could have given rise to spuriously high correlations two sources of error could have acted to reduce the magnitude of the correlations. The first of these derives from the fact that the measures used in it were "prototypic" in nature. This was the case for both the predictor and criterion measures, for both were first generation in their development and representative of relatively unexplored approaches to measurement.[1] As such the conceptual framework which guided item development in the predictor and criterion measures was relatively primitive, the filmed episodes around which the predictive instruments were built were relatively weak, the item analyses used in their development were based on responses of experienced teachers whereas the instruments were subsequently

[1] The Schalock, Beaird and Simmons study was actually designed as a validation study of the three situational response measures that were used as predictors. Also, the observation system from which the criterion measures were derived was developed within the context of the study. Both sets of measures are described in the next section of the report.

used with inexperienced teachers, and the observation system used in obtaining the criterion measures suffered from relatively low reliability on the part of observers applying it. In combination these limitations led to a set of predictor and criterion measures which were more limited in range and quality than ultimately desired.

The second source of error in the study that could have acted to reduce the magnitude of the correlations found was the failure of the researchers to control for situational factors that interact with or are thought to influence teaching behavior in the classroom. Factors such as unplanned events, composition of the class, physical conditions within the classroom and the nature of the activity in which teacher and learners engage were not controlled, and since these are likely to be significant determinants of teaching behavior their omission or neglect should have reduced still further the magnitude of the correlations found in the study.

In light of these kinds of limitations in measurement it is remarkable that correlations of the magnitude demonstrated were obtained.

Given the data that derived from the study, and the many potential sources of error that accompanied them, a proposal was submitted immediately upon the completion of the study to the U.S. Office of Education for its replication and extension. Three factors led to the second proposal: (1) the essentially unprecedented results obtained in the parent study, (2) the numerous potential or real sources of error in it, and (3) the desire to avoid the pitfalls of uncritical test adoption, that is, the desire to forestall the users of tests from moving too quickly to adopt the instruments developed in the study for use in their own programs of research or evaluation. Since these instruments were new, and since the first predictive efforts with them were so promising, there was danger that the measures might be applied in areas where basis for their application did not exist. Cronbach (1960) states this danger well:

> When an investigator has once obtained a satisfactory validity coefficient he tends to install his program and stop research. Other workers, reading his report of the study, accept his test as valid and put it to work in their own situations. This practice is unsound. In the first place, any validation result is influenced by chance, and correlations will fluctuate from sample to sample. Consequently the test which proves best in one sample may not prove to be the best predictor in another similar sample. Even when the results are based on a large sample the particular score or the particular weights most effective in a multiple correlation are certain to change when a new group is tested. If the same formula is applied to other groups, correlation is sure to drop. Moreover, the supply of men and the conditions of training change according to time. It follows that the investigator must redetermine the validity of his prediction technique periodically.

Four major objectives guided the present study:

(1) to replicate the parent study;

(2) to extend the design of the parent study to experienced, primary grade teachers;

(3) to strengthen both replication studies by increasing the number of subjects used in each and including in them measures of situational variables that affect predictive accuracy; and

(4) to investigate the effects on prediction of deriving criterion measures from behavioral samples of varying lengths.

The rationale for objectives (1) and (3) has already been spelled out. The rationale for objective (2) was twofold: a) the desirability of testing the power of the predictive measures with a variety of teacher populations, and b) the theoretically based expectancy that the situational-response tests would predict the behavior of experienced teachers better than they would student teachers because of the wider background of experience they can draw upon in interpreting the situation before responding to it and because the items in the tests were validated initially against a population of experienced teachers. The rationale for objective (4) was simply that the systematic study of behavioral sampling and its relation to the stability of measures dependent upon it is long overdue. Observational methodology, especially as it applies to <u>prediction in situation</u>, is inescapably dependent upon behavioral sampling yet there has been no research to date to suggest clearly the nature of the sample needed to maximize prediction. While the present study did not permit an exhaustive investigation of the issue (length of behavioral samples were limited to one, two and three hours), it was hoped that it would provide a point of departure for subsequent work.

In passing it should be pointed out that the investigation of situational measures and their relationship to the predictability of behavior in situation was also exploratory in nature, with situational measures being limited to rather gross descriptions of classroom structure and composition, events and the orientation of school administrators toward classroom management. Much the same point of view underlaid this effort as underlaid the study of behavioral sampling: a great deal has been written about the significance of situational variables in research design, but as yet no one has gotten serious about their measurement. It was hoped that the present effort would represent a start in that direction.

A fifth objective evolved as the study progressed, namely, to strengthen the criterion measures used in it. This required extensive work on the observation system developed in the parent study, and led in part to a request for a 6-months extension of the study. A by-product of this extension is the accompanying monograph (see Attachment I) that provides an overview of the observational system that derived from the effort. The system is referred to generally as the Teaching Research System for the Description ofTeaching Behavior in Context, and provides the most exhaustive measure of teaching behavior presently available. As such, its development represents one of the major contributions of the project.

The one major source of error inherent in the parent study that could not be reduced in the replication study was that attributable to the quality of the predictor measures: they had to remain unchanged.

Because the research to be reported ties so closely to the Schalock, Beaird and Simmons study, the next chapter in the report is devoted to its review.

Chapter II

AN OVERVIEW OF THE SCHALOCK, BEAIRD AND SIMMONS STUDY

As indicated previously, the Schalock, Beaird and Simmons study was intended as a validation study of situational-response tests which used motion picture sequences of classroom behavior as test stimuli. The general hypothesis underlying the study was that in order to predict to complex human behavior the tests to be used as predictors had to reflect in their composition the complexity of the behavior to be predicted. Specifically, the hypothesis tested in the study was that as test stimuli increased in their representativeness of the behavior to be predicted, and as the opportunity for response to those stimuli approached "life-likeness" in their freedom, the predictive power of tests would increase accordingly. Motion picture sequences of classroom behavior were used in an effort to provide a stimulus situation comparable in complexity to that involved in real life teaching.

The Predictor Measures

Four predictor tests, varying on a continuum of stimulus and response complexity, were used in the study: 1) a traditional paper-and-pencil attitude scale, where the test stimulus was a statement describing an orientation to the teaching function and response was defined by agreement or disagreement to the statement (The Minnesota Teacher Attitude Inventory), 2) a situational-response test where the test stimuli were written descriptions of filmed classroom situations and response was defined by agreement or disagreement to statements made in relation to the situational descriptions (The Word Test), 3) a situational-response test where the test stimuli were motion picture sequences of classroom situations and response was defined as in (2) above (The Film Test), and 4) a situational-response test where the test stimuli were also motion picture sequences of classroom situations but the response was free, i.e., the subject responded to the filmed situation *as if she were the teacner in the situation* (The Simulation Test). It was hypothesized that the predictive power of the tests would vary in the order of their listing above, with the MTAI being the weakest predictor and the simulation test the most powerful. The relationship of these tests to one another on a continuum of stimulus and response complexity appears as Figure 1.

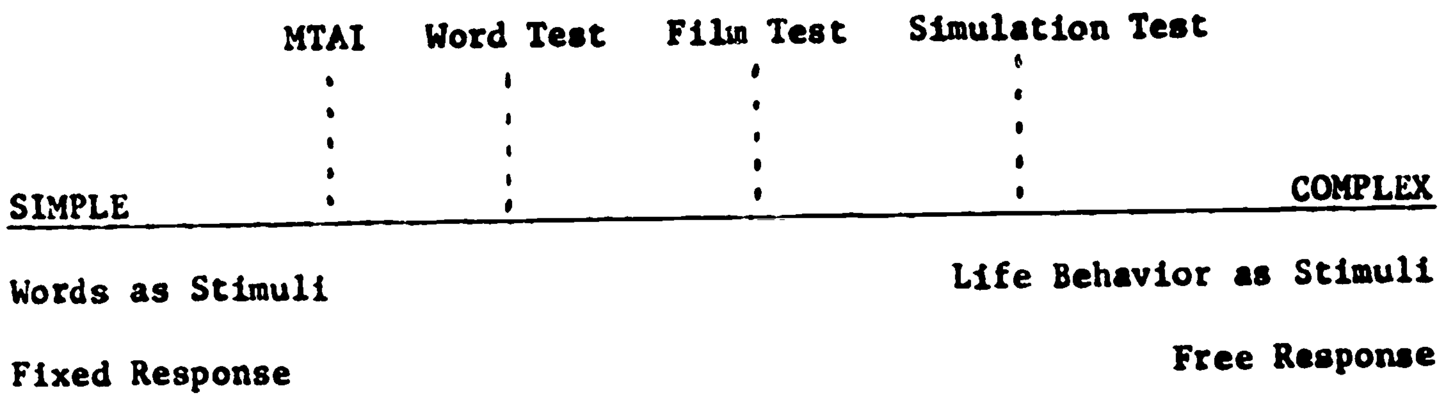


Figure 1. Continuum of test stimulus and response complexity.

The Word, Film, and Simulation Tests were constructed especially for the project. Generally speaking, they were designed to assess a teacher's orientation to classroom management and interpersonal relationships with children. No attempt was made to assess orientation to learner outcomes or the instructional strategies pertaining to them. Situations portrayed in the tests were identified as particularly challenging and representative of these dimensions of the teaching process by first, second, and third grade teachers.

Word Test. The word test consists of 13 written descriptions of actual classroom situations which occurred in the first, second, and third grades of the Campus Elementary School (CES) at Oregon College of Education. Each written description of a situation is followed by 12 to 22 statements about the situation to which respondents agree or disagree on a five point scale (Strongly Disagree to Strongly Agree). The test provides a total score and 12 subscale scores (see Table 1). Split-half reliability of the various scales range from .55 to .94.

Film Test. The film test consists of 13 motion picture sequences of actual classroom situations which occurred in the first, second, and third grades in CES. Each sequence is followed by 11 to 22 statements about the situation portrayed to which testees respond in the same manner as for the word test. The test provides a total score and 11 subscale scores. Split-half reliability for the various scales range from .51 to .92.

Simulation Test. The simulation test consists of 12 motion picture sequences of classroom events filmed in a single second grade at CES. The sequences are arranged chronologically to represent a single day. The test is accompanied by a cumulative folder detailing anecdotal and test information for each of the "main characters" portrayed in the film sequences. Sequences were filmed in such a manner that when viewing the films the children are looking directly at the respondent. Respondents record, verbatim, their reactions to the situations and then describe (1) why they responded in the way they did, (2) what they hoped to attain through their response, (3) their impression of the child (when a single child was involved) and (4) why they responded at the time they did. The test provides a total score and seven subscale scores. Scores are derived through a content analysis of written responses.

Subscales and reliability estimates for the Word and Film Tests are presented in Table 1. Subscales for the Simulation Test are presented in Table 2. Because the scores of the Simulation Test subscales are derived from judges' ratings of respondent behavior, reliability coefficients

Table 1. Subscales and Reliability Estimates for the Word and Film Tests

Word Test Subscales	r	Film Test Subscales	r
1. Management I	.782	1. Management I	.752
2. Interpersonal Awareness	.818	2. Interpersonal Awareness	.883
3. Technique Awareness	.702	3. Technique Awareness	.656
4. Management II	.834	4. Management II	.752
5. Orientation to Strategies	.546	5. Response to Deviation	.651
6. Orientation to Structure	.864	6. Philosophy of Structure	.510
7. Philosophy of Structure	.740	7. Approach to Structure	.718
8. Approach to Structure	.730	8. Teacher Characteristics	.631
9. Teacher Characteristics	.868	9. General Interpretation	.512
10. General Interpretation	.832	10. Specific Interpretation	.915
11. Specific Interpretation	.929	11. Specific Sanction	.785
12. Specific Sanction	.784	12. Total Test	.915
13. Total Test	.940		

of the usual nature were not determinable. Instead, inter-rater reliability was determined and was found to be consistently acceptable.

Table 2. Subscales of the Simulation Test

1. Management I
2. Interpersonal Awareness
3. Structure
4. Address to Individuals
5. Use of Questions
6. Trust
7. Academic Orientation

The Criterion Measures

In order to provide a rigorous test of the basic hypothesis, it was decided to use as criterion performance specific behavioral measures instead of more typically used global measures of teaching success. To this end systematic observational procedures were used as the primary data source in the study. Performance ratings, which have plagued the field of research on teacher effectiveness, were not used.

The system of observations used in the study, from which the criterion measures were derived, involved both preconceived category sets and rating scales. Category sets were developed for the description of specific interactive behaviors that occurred between teacher and child, and rating scales were used to assess some of the more global qualities reflected by the teacher in the situation. Both categories and rating scales were designed to assess the same parameters of the teaching process that the predictor measures were designed to assess, namely, orientation to classroom management and interpersonal relationships with children.

Interaction was conceptualized as following essentially a stimulus-response paradigm, where a stimulus (cue, demand) might or might not be responded to and a response might or might not serve as an invitation (stimulus) to a further response. The basic model for observation was a three-stage interaction sequence: (1) a stimulus (demand situation) operating upon the teacher within the classroom setting, (2) a response (or lack of response) of the teacher to the demand situation, and (3) the response of a child or group of children to the teacher's response. With this model behaviors of the teacher could be related explicitly to behaviors of children in her class. In turn some child behavior could be related to behaviors of the teacher. The model also permitted recording of interaction between teacher and child that continued over time, i.e., where there were more than three exchanges in the interaction sequence. The categories that made up the system appear in Tables 3 and 4.

Table 3. Classes of Child Behavior Descriptive of Stimulus and Response Conditions

Category Set 1		Category Set 3
Classes of Child Behavior as Stimuli to Teacher Behavior		Classes of Child Behavior as Responses to Teacher Behavior
1. Ignoring of group goal	a	Acceptance au Unqualified Acceptance aq Qualified Acceptance ar Acceptance with Reward
2. Intense social involvement		
3. Involvement in academic content		
4. Routine classroom functions	t	Tending
5. Rule breaking behavior	i	Exchange of Information
6. Conflict behavior	pp	Postpones
	ig	Ignores
	r	Rejection ru Unqualified Rejection rq Qualified Rejection r pers Rejection Through Attempts at Persuasion

Table 4. Classes of Teacher Behavior

Category Set 2

R	Judges behavior worthy of
Ru	Unqualified reward
Rq	Qualified reward
T	Tending
I	Exchange of information
D	Directing (non subject matter)
S	Structuring (subject matter)
Sq:	Direction giving (who, what, where, when)
Sh:	Explaining (how)
Si:	Information giving
Sc:	Correcting
Pp	Postpones
Ig	Ignores
C	Judges behavior worthy of change
Cu	Attempts change with unqualified power
Cq	Attempts change with qualified power
C pers	Attempts change through persuasion or suggestion

Affect and Intensity ratings also accompanied the recording of each category of teacher and child behavior. This represented an effort to obtain a measure of the feeling tone and/or intensity of the interaction. Four affect measures were used: (1) warmth, intensity, exuberance; (2) distance, aloofness, hostility; (3) upset, concern, anxiety; and (4) neutrality, or a lack of any of the above. Three levels of intensity were used: low, moderate and high. Intensity ratings were always made relative to the intensity of the situation in the classroom at the time.

In addition to the category sets nine rating scales were developed to measure some of the more general characteristics of teacher behavior. These were adapted from the scales developed by Schalock and O'Neil (1961) in relation to parent-child interaction, and included measures of (1) Tolerance for Changeworthy Behavior, (2) Warmth, (3) Respect for the Individuality of Children, (4) Comfortableness, (5) Intellectuality, (6) Consistency, (7) Tempo, (8) Organization, and (9) Harmony. All were rated on a five point scale.

Data became available from the measurement system in the form of category frequency counts and ratings. One of the unique features of the observation system was that each category of interaction was able to be identified as to who initiated what kind of behavior, and what the response to it was! Thus it was possible to determine not only the frequency with which a teacher responded to children with power, or ignored a child initiation, or gave help, but it was also possible to determine the kind or class of behavior that elicited such responses. It was also possible to identify such factors as the role the teacher played in the classroom, for example, whether she tended to be the center of things through lecturing, structuring, directing, etc., or whether she let the children assume the more active role; whether the children tended to initiate interaction with her or avoid her; and what kinds of behavior she tended to reward, punish or ignore.

The rating scale data tended to support and extend the basic data obtained through the direct behavioral measures.

The fifteen criterion measures used in the study were derived from these category and rating scale data. Three features characterized the criterion measures:

1) They were theoretically relevant, i.e., they related to dimensions of the model of teaching behavior used as a guide to instrument development throughout the study, and, as a consequence, exhibited a close tie to the predictive instruments that were developed;

2) They were complex in the sense that they represented a pooling of a number of conceptually related behaviors into a ratio or combination score. Theoretically this provided a more stable and comprehensive measure than would single classes of behavior; and

3) The measures took full advantage of the power of the observational system in the sense that they tied to (a) various classes of child behavior, (b) the teacher's response to classes of child behavior, and (c) the child's response to the teacher's behavior.

So far as we know, this is the first time that observational data have been used in this particular way, and on a priori grounds the measures derived by means of this procedure are most promising.

The measures derived from the category data appear in Table 5. The measures derived from the rating scale data appear in Table 6. The reliability of these measures will not be reviewed here, but on the basis of rather tenuous reliability data (see the original report) all measures were judged to be minimally adequate. Upon use it was found that the category based measures were essentially unrelated or independent measures (low intercorrelations) while the rating scale based measures were highly related.

Table 5. Criterion Measures Used in the Study That Were Based on Category Frequency Counts

Measure	Source
1. Permissive-Restrictive	1. $\dfrac{\text{Total Cu + Cq + Cpers*}}{\text{All Teacher Acts}}$
2. Power	2. $\dfrac{\text{Total Cu}}{\text{Total Cu + Cq + Cpers}}$
3. Consideration I	3. $\dfrac{\text{Total Ru + Rq + Cq}}{\text{Total Ru + Rq + Cq + Cu + Cpers + Pp + Ig}}$
4. Consideration II	4. $\dfrac{\text{Total T + I + Sh + si in response to child initiations in categories 2, 3 and 4}}{\text{All of the above plus all other teacher responses to child initiations in categories 2, 3 and 4}}$
5. Affective Orientation of Teacher	5. $\dfrac{\text{Total (+)}}{\text{Total (+) + (-) + (}\checkmark\text{)}}$
6. Teacher Success in Obtaining Cooperation or Compliance	6. $\dfrac{\text{Total child responses of au, +, i or i}\rightarrow\text{ to teacher Cu, Cq, Cpers, Sq, or D actions}}{\text{All of the above + all other child responses to these teacher actions.}}$
7. Teacher Approachableness	7. Total child category 2, 3, and 4 entries, including questions (2→, 3→, 4→) in Flow Pattern III.
8. Individual vs. Group Focus	8. $\dfrac{\text{Total teacher acts directed to group or part group}}{\text{Total Teacher Acts}}$
9. Teacher vs. Child Focus	9. $\dfrac{\text{Total Flow Pattern III entries}}{\text{Total Teacher Acts}}$
10. Directing vs. Facilitating	10. $\dfrac{\text{Total Sq}}{\text{Total Sq + Shi + Sh + Sc, including questions, in any of these categories}}$
11. Question vs. Statement	11. $\dfrac{\text{Total Si}\rightarrow\text{ + Sh}\rightarrow\text{ + Sq}\rightarrow}{\text{Total Si + Sh + Sq}}$

*Cu, Cq, Cpers, etc. are category labels used in recording (see Tables 3 and 4). For category definitions and examples, see Schalock, Beaird and Simmons, pp. 400-434.

Table 6. Criterion Measures Used in the Study that were Based on Rating Scales

Measure	Source
12. Permissive-Restrictive	Ratings on Tolerance for Changeworthy Behavior Scale
13. Consideration	Rating on Respect for Individuality Scale
14. Classroom Climate	Ratings on Warmth and Harmony Scales
15. Total Teacher Characteristics	Total of All Scale Ratings

Procedures

Subjects were senior women majoring in elementary education at Oregon College of Education and Oregon State University who were teaching in the primary grades during the 1963-64 school year. A total of 56 subjects participated in the study, although the sample attenuated for various reasons to a final N of 40.

Prior to the academic quarter during which the subjects were engaged in their student teaching, the four predictive measures (MTAI, Word Test, Film Test, and Simulation Test) were administered. The tests were administered in group settings and in a randomized order. Six half-hour records of interactive behavior for each subject in the classroom, collected on two separate days, constituted the behavioral sample. Each day two half-hour observations were made in the morning and one half-hour observation in the afternoon. A different observer observed each subject each day. All observations were made during the last two weeks of the subject's student teaching experience.

Results

Three levels of regression analyses were run: Level I related *total test scores* to each of the 15 criterion measures used in the study; Level II related the *subscale* scores from each of the three situational response tests to the various criterion measures; and Level III related *a combination of subscales from the various tests that proved to be effective predictors in the Level II analysis* to the criterion measures.

Results of Level I Analysis. Fifteen regression analyses were run, one for each criterion behavior, using in each case the total scores for the MTAI, Word Test, Film Test, and Simulation Test as the predictor variables. The percent of criterion variance (R^2) accounted for by total scores of the four predictors ranged from 8.0 to 32.6. The L test for ordered hypotheses (Page, 1963) was nonsignificant, failing to substantiate the basic hypothesis.

Results of Level II Analyses. Sixty regression analyses were run, one for each criterion behavior (15) for each of the four instruments used as predictors. For the MTAI zero order coefficients of correlation were computed since it does not have subscales. Per cent of criterion variance (R^2) accounted for by the MTAI ranged from zero to 5.3, with a mean of 1.7; per cent of criterion variance accounted for by subscales of the Word Test ranged from 14.9 to 50.1, with a mean of 30.9; per cent of criterion variance accounted for by subscales of the Film Test ranged from 18.7 to 49.6, with a mean percent of variance of 38.3; and per cent of variance accounted for by Simulation Test subscales ranged from 22.9 to 54.1 with a mean of 37.6. The L test for ordered hypotheses was significant at the .005 level, substantiating the hypothesis of difference in predictive effectiveness as tests moved toward the approximation of lifelikeness. It will be noted, however, that the hypothesis did not hold with respect to the predicted relationship between the Film Test and the Simulation Test.

Results of Level III Analyses. Fifteen Level III regression analyses were made, one for each criterion behavior. Each analysis utilized 18 predictor variables - the MTAI total score, the five subscales of the Word Test that proved to be the most effective predictors of a given criterion measure, the five subscales of the Film Test that were the most effective predictors of the same criterion, and the seven subscales of the Simulation Test. In most cases, the subscales used in any given regression analysis differed from those used in other regression analyses.

Per cent of criterion variance accounted for in Level III analyses ranged from 49.0 to 75.7 with a mean of 58.8. The L test for ordered hypotheses was significant for these data at the .05 level, with the Simulation Test subscales consistently outranking the other predictors in accounting for criterion variance. The MTAI consistently ranked last, with the select subscales of the Word and Film Tests accounting for essentially the same amount of variance in the criterion measures.

The Multiple R's that derived from the three levels of analysis are presented in Table 7. On the basis of these data it was concluded that in general the results supported the basic hypothesis tested in the study, namely, that as test stimuli become more representative of

Table 7. Per Cent of Criterion Variance (R^2) Accounted for in the Three Levels of Regression Analysis

	Level I	Level II			Level III
	(Total Test	(Subscale Scores)			(Selected Subscale
Criterion	Scores)	Word	Film	Simulation	Scores)
1	.176	.303	.281	.303	.563
2	.325	.384	.476	.360	.504
3	.185	.504	.476	.303	.624
4	.123	.314	.397	.436	.624
5	.176	.292	.384	.292	.504
6	.144	.221	.384	.533	.504
7	.090	.152	.270	.230	.476
8	.137	.348	.292	.384	.593
9	.102	.270	.449	.384	.608
10	.048	.221	.410	.397	.757
11	.073	.303	.360	.449	.723
12	.084	.325	.436	.384	.548
13	.194	.314	.436	.436	.689
14	.221	.410	.490	.384	.656
15	.0784	.176	.185	.360	.490

the behavior to be predicted and as the opportunity for response approaches the freedom characteristic of life situations the power of prediction increases.

Chapter III

AN OVERVIEW OF THE REPLICATION STUDY

As indicated previously, the present study represented an extension as well as a replication of the parent study. Two extensions were undertaken: 1) the addition of situational data, that is, descriptions of classroom structure, composition, unplanned events, etc., to the original prediction scheme, and 2) the repetition of the study, including the use of situational measures, with experienced elementary school teachers. In addition, the study was designed to provide information on the effect on prediction of using behavioral samples of differing lengths in obtaining the criterion measures. In providing an overview, each of these aspects of the study will be described.

The Replication Study

Every effort was made to exactly replicate the parent study. Subjects were drawn from the same population, the same predictive measures were used,[2] criterion measures were equivalent but strengthened (see below), and the same analyses were applied.

Subjects. Thirty-nine senior women, majoring in elementary education ✓ with specialization in the primary grades at either Oregon State University or Oregon College of Education, served as subjects for the replication

[2] Three of the tests used, the MTAI, the Word Test, and the Film Test, were completely equivalent since their administration and scoring required no coding or interpretation; the Simulation Test was as "equivalent as possible" considering its reliance upon coders for its scoring.

study.[3] Subjects were drawn from the pool of students who did their student teaching in Winter and Spring terms of the 1965-66 academic year and Fall and Winter terms of the 1966-67 academic year. Only students who volunteered to take part in the study and who did their student teaching within a 60 mile radius of Oregon State University were eligible for inclusion in the study. These were the same criteria used in the parent study, and approximately the same proportion of students met these criteria as they did in the parent study. No consistent differences appear in predictor or criterion measure scores for the students from the two institutions.

Predictor measures. The same predictor measures that were used in the parent study, that is, the MTAI and the Word, Film and Simulation Tests, were also used in the replication study. They were administered in group settings at the close of the term that preceded the term in which the subjects did their student teaching. In contrast to the parent study, however, a totally random order of test presentation was not followed. The four tests required approximately five hours to complete, and it turned out to be impossible to get all subjects to arrange for a block of time of that length. It was possible to get everyone to arrange for a half a day of testing, however, so the expedient of having them take three of the four tests during the scheduled time and one of

[3] The project proposal called for 40 to 45 subjects but an effort was made to increase this number to 60. With the aid of an extension to the project, fifty-eight student teachers were tested and/or observed to some degree of completeness, but due to illness, schedule conflicts, and other "end of the term" complications (student teachers had to be observed within a two week period at the end of their student teaching experience) complete data was obtained for only 39 of them.

them at home was followed. This was a workable solution since two of the tests, the MTAI and Word Test, were self-administered, paper and pencil measures. Operationally this meant that the Film Test and Simulation Test were always administered under supervised conditions, and either the Word Test or the MTAI were administered under non-supervised conditions, i.e., at home. Furthermore, since it was critical that the Film and Simulation Tests be administered under supervised conditions, it also meant that these two tests were always administered in a one-two order, and that the MTAI or Word Test was always third in the order of presentation. While the Film and Simulation Tests were always assigned their order of presentation randomly, and the Word Test and the MTAI were always assigned to the non-supervised condition randomly, the inability to follow a totally random assignment of test order represented a source of error in the data and a departure from the procedure followed in the parent study.

Another departure from the parent study derives from the scoring procedures used with the Simulation Test. While the MTAI, Word and Film Tests require only the tabulation of the responses students make to them, the Simulation Test requires the coding or classification of descriptively written responses (protocols) that are made to it. This requires that coders master a category-rating scale system (see pp. 114-118 in Schalock, Beaird and Simmons, 1964) and demonstrate their reliability in applying it.

Generally speaking, evidence as to the accuracy with which coders could apply the simulation coding system was disappointing. As in the parent study, project staff worked in two-man teams in making the category and rating scale assignments. Each member of a team independently read each response and independently assigned a score for each scale to that response, but, after comparing codings arrived at a joint decision as to the "correct" ratings or category placement when there was disagreement.

Evidence as to the reliability with which teams assigned their codings was obtained by having each team score eight protocols and then compare their codings. The results of this comparison appear in Table 8.[4] As in the parent study, it was decided arbitrarily to identify as inter-team disagreement any variance of three or more frequencies in either the numerator or the denominator of each score. Using this base, each measure was then checked to see the number of disagreements between teams that appeared across the eight protocols. The cells that are enclosed with heavy lines in Table 8 are the scores on which the coder teams were judged unreliable by applying this criterion.

It will be seen from these data that the coding teams were unable to agree upon categorization or scale placement much more than 60 per cent of the time. In most studies this level of agreement would be judged inadequate and further training of coders or refinement of the category-rating scale system used in the coding would be indicated.

[4] It will be noted that 15 category and scale scores appear in Table 8 while only 7 predictor measures come from the Test as a whole. This apparent discrepancy is accounted for by the combination of some of the 15 scales into single predictor measures. The factor analytic data upon which these combinations rest are reported in the parent study (see pp. 117-119).

Table 8. Reliability of Teams in Coding the Simulation Test Protocols.*

Scale Subject		Restrictive-Permissive	Power	Focus of Address: Individual	Focus of Address: Group	Facilitating-Directing	Question	Statement	Sensitivity	Focus of Concern: Group	Focus of Concern: Individual	Trust II	Trust III	Unique vs. Categorical Orientation	Academic Orientation	Personal Orientation
1	Team A	9/11	6/10	7/10	3/10	5/10	3/10	5/10	2/10	8/10	4/10	6/10	15/20	2/10	8/10	3/11
	Team B	8/11	9/10	8/10	3/10	2/7	6/10	8/11	5/10	7/10	5/10	3/10	18/21	5/10	9/11	2/8
2	Team A	5/11	6/11	9/11	2/11	2/11	1/11	8/11	5/11	7/11	4/11	2/11	15/21	3/11	10/11	3/11
	Team B	6/11	5/10	8/11	4/11	4/11	2/11	9/11	1/10	5/8	7/9	1/10	17/21	2/9	9/11	3/10
3	Team A	5/10	4/11	8/11	2/11	8/9	1/9	9/11	8/10	4/10	8/10	10/11	18/22	5/10	6/11	4/10
	Team B	8/11	7/10	6/7	3/7	5/9	2/9	9/11	5/10	7/10	5/11	6/11	18/22	8/11	9/11	6/10
4	Team A	5/11	7/9	9/10	1/10	7/8	4/8	6/10	9/11	9/11	3/11	3/11	19/22	3/11	7/11	5/11
	Team B	1/10	9/9	8/11	2/11	9/11	4/11	7/11	8/11	7/11	5/11	6/11	18/22	5/11	5/11	7/11
5	Team A	9/11	7/8	7/10	3/10	4/10	2/10	8/11	7/10	10/11	2/10	1/11	20/22	7/11	8/10	2/11
	Team B	9/11	8/8	7/10	4/10	5/10	4/10	7/10	8/10	9/10	4/10	4/11	17/21	8/11	8/11	4/11
6	Team A	8/10	5/9	8/11	2/11	8/10	4/9	6/9	1/11	3/11	9/11	10/11	18/19	2/11	6/11	6/11
	Team B	6/11	8/11	9/10	3/10	11/11	7/9	3/9	4/10	3/11	6/10	6/11	20/22	6/11	6/11	2/11
7	Team A	7/11	4/10	7/11	4/11	3/10	2/9	7/10	0/11	7/10	4/10	8/9	18/22	1/11	10/11	0/11
	Team B	8/10	5/11	6/11	7/11	6/11	3/9	8/11	1/11	10/11	2/10	7/10	16/22	3/11	10/11	2/11
8	Team A	10/11	6/9	6/10	5/10	11/11	1/10	9/10	2/10	6/11	5/11	10/11	22/22	7/10	8/11	5/11
	Team B	9/10	6/10	7/10	5/10	10/11	1/10	10/11	7/10	7/10	6/10	7/11	20/22	8/10	10/11	3/11

*Heavy lined squares represent cells where observer teams were in sufficient disagreement as to be judged unreliable.

This was not a feasible solution in the present study for two reasons: 1) additional training in the system did not appreciably affect reliability scores (the category definitions and scoring rules described for the measure in the final report of the parent project were sufficiently inadequate to make the demonstration of reliability impossible, no matter how intensive the training) and 2) the system could not be refined or altered as that would lead to a predictor measure that was different from that used in the parent study. As a consequence, the level of reliability demonstrated in Table 8 had to be deemed acceptable even though the data that thereby derived from the simulation measure were of a highly unreliable quality. It is interesting to note, however, that even with this degree of unreliability in the data the measures that derived from the data were relatively independent (see Table 9).

Table 9. Intercorrelation Matrix for Subscales of the Simulation Test

Scales	1	2	3	4	5	6	7
1 Management	1.00	-.40	-.63	.48	-.19	-.12	.22
2 Interpersonal Awareness		1.00	.43	-.52	.13	-.22	-.70
3 Structure			1.00	-.40	.13	-.006	-.45
4 Address to Individuals				1.00	-.18	.12	.33
5 Use of Questions					1.00	.01	-.05
6 Trust						1.00	.26
7 Academic Orientation							1.00

Criterion measures. At the time that the parent study was undertaken (1962 - 1964) none of the classroom interaction measures that then existed were particularly appropriate to the purposes of the study. The focus of the predictor measures was upon discipline or classroom management behavior, and with the exception of the work of Hughes (1959), and to some extent that of Medley and Mitzel (1958), existing measures did not take that dimension of teaching behavior into account. As a consequence an effort was made to develop a system for describing teacher-learner interaction that focused upon both classroom management and instructional behavior. The decision to undertake such an effort stemmed from a history of experience in the application of observational methods to the study of parent-child interaction (Moustakas, Sigel and Schalock, 1956) (Schalock and O'Neill, 1960) and a deep dissatisfaction with the superficiality of measures of teaching behavior being proposed at that time by Hughes (1959), Flanders (1960), Smith (1960), and Medley and Mitzel (1958).

As anyone who has attempted to develop an observational system knows, it is a time consuming and difficult task. As a consequence, while it was possible to develop a system of observation that provided the kind of data needed in the parent study, the system itself was little more than a first approximation to the system ultimately desired. This became clearer and clearer as the present study progressed, and as a result the decision was made to extend the system within the context of the present study to a more finished state. It was partially toward this end that a six month extension to the study was obtained.

The observational system that evolved as a by-product of the study represents an effort to develop a conceptually sound, relatively exhaustive measure of teaching behavior and the contextual variables which influence it. In developing the system, advantage has been taken of the work of others who have been interested in describing teaching behavior, for example, Hughes (1959), Flanders (1960), Smith (1964), Bellack (1963, 1965), Aschner and Gallagher (1963), and Taba (1964); the work of Bales (1950) in the study of small group interaction; and the work of Bishop (1951), Moustakas, Sigel and Schalock (1956), and Schalock and O'Neill (1960) in the study of parent-child interaction. An effort has been made in the present system, however, to move beyond earlier efforts and to overcome many of their limitations (Schalock, 1967). Specifically, an effort has been made to tie the system conceptually to that which is known about cognitive development and the teaching-learning process, to include in it a running account of both teacher and learner behavior, to make it inclusive of both the instructional and the management parameters of teaching, to use as a data base both the verbal and non-verbal aspects of teacher-learner interaction, and to conceptualize teaching behavior so as to make the system applicable across a wide range of ages and settings, e.g., the home or nursery school, the playground or classroom, the elementary or the secondary school. In addition, the TR System provides a detailed record of the subject matter, classroom organization and activity in which a class is involved. In short, the observation system represents an attempt to develop a means of looking at teaching behavior wherever and whenever it occurs and to describe it as occurring in relation to the full range of factors which influence it. An overview of the system is presented in the accompanying monograph (see Attachment 1). Detailed category definitions, examples, and operational procedures appear in a training manual that is now being completed (Schalock and Micek, 1968).

The use of an expanded observation system in the replication study represented a potential problem: how does one use a "different" measurement system and still obtain essentially the "same" measures? A further complication stemmed from the decision to eliminate from the study the four criterion measures used in the parent study that depended upon rating scale data (measures 12 through 15 in Table 6, p. 18). As indicated previously, the intercorrelations between these measures were sufficiently high as to make them unacceptable as independent measures. With these changes, the decision was finally reached to include eight measures that were representative of the category based measures used in the parent study and three new measures made possible by the expanded observation system. The total set of criterion measures used in the replication study are listed in Table 10. The reliability of observers in applying these measures, as this is reflected in the comparability of criterion measures obtained by individual observers observing simultaneously but independently, is presented in Tables 11 and 12. While these data were not as supportive of observer reliability as had been desired, the press of the project schedule demanded that they be accepted so that field observations could be undertaken. Fortunately, even though some of the measures appeared to be relatively unreliable prior to formal observation, they proved to be relatively independent as formal measures. The intercorrelation data for the criterion measures, as these were derived from the final data pool on both student and experienced teachers is presented in Table 13.

Table 10. Criterion Measures Used in the Replication Study

Measure*	Source**
1. Degree of Control	$\frac{\text{All teaching moves which reflect censorship}}{\text{All teaching moves}}$
2. Orientation to the Use of Power in Maintaining Control	$\frac{\text{All censorship moves which rely upon power as a basis for behavioral change}}{\text{All censorship moves}}$
*3. Teacher Response to Deviant Behavior	$\frac{\text{All non-censoring responses to deviant behavior}}{\text{All responses to deviant behavior}}$
*4. Orientation to the Use of Positive Reinforcement	$\frac{\text{All instances of positive evaluation}}{\text{All evaluative moves}}$
5. Consideration in Response to Academic Initiations	$\frac{\text{All non-censoring responses to academic initiations (Flow III)}}{\text{All responses to academic initiations}}$
6. Consideration in Response to Non-Academic Initiations	$\frac{\text{All non-censoring responses to non-academic initiations (Flow III)}}{\text{All responses to non-academic initiations}}$
*7. Consideration in Response to All Non-Academic Behavior	$\frac{\text{All non-censoring responses to non-academic behavior (FlowII)}}{\text{All responses to non-academic behavior}}$
8. Affective Orientation	$\frac{\text{All instances of positive affect (+)}}{\text{All instances of affect (+)+(-)+(}\checkmark\text{)}}$
9. Teacher Approachableness	$\frac{\text{All instances of student initiations (Flow III)}}{\text{All instances of student and teacher initiations}}$
10. Individual vs. Group Focus	$\frac{\text{All instances of interaction with a single child}}{\text{All instances of interaction}}$
11. Use of Inquiry in Instruction	$\frac{\text{All instances of Inquiry in relation to academic matters}}{\text{All teacher acts in relation to academic matters}}$

*The measures that are new to the replication study are starred.

**The specific categories of behavior making up these ratio measures can be found in the monograph that provides an overview of the Teaching Research System of Observation (see Attachment 1) and the Training Manual that accompanies the system (Schalock and Micek, 1968).

Table 11. Reliability of Observers in Applying the TR Observation System in February, 1966, as this is Reflected in the Comparability of Criterion Measures that Derive from Simultaneous but Independent Observations.

Criterion	Observers*			Observers			Observers			Observers		
Measure	A	B	C	D	A	F	A	B	F	C	E	F
1	.03	.06	.02	.07	.09	.05	.05	.04	.04	.08	.04	.06
2	.85	.65	.72	.55	.67	.73	1.00	.85	.92	.75	.56	.56
3	.00	.00	.00	.00	.00	.00	.00	.00	.33	.00	.00	.00
4	.34	.45	.53	.32	.34	.37	.36	.44	.33	.40	.63	.67
5	.40	.32	.47	.22	.43	.61	.93	.76	.72	.92	.93	.81
6	.00	.25	.00	.33	.33	1.00	.00	.00	.33	.00	.00	.00
7	.25	.25	.50	.00	.33	.33	.00	.00	.00	.50	.25	.75
8	.00	.50	.75	1.00	1.00	1.00	1.00	1.00	.50	.00	.00	.00
9	.12	.16	.14	.07	.08	.10	.11	.15	.17	.54	.43	.56
10	.22	.27	.23	.44	.45	.52	.45	.31	.36	.21	.25	.18
11	.45	.36	.54	.30	.29	.33	.25	.32	.12	.26	.22	.28

Criterion	Observers			Observers			Observers			Observers		
Measure	C	D	E	A	B	D	B	C	E	D	E	F
1	.09	.15	.11	.08	.11	.06	.04	.02	.03	.07	.11	.08
2	.69	.87	.52	1.00	.53	.63	.73	.81	.76	.82	.96	.55
3	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
4	.62	.27	.54	.15	.17	.23	.27	.33	.26	.33	.35	.35
5	.81	.91	.69	.75	.78	.76	.78	.76	.92	.43	.68	.72
6	1.00	.50	.00	.00	.00	.00	.00	.00	1.00	.00	.00	.00
7	.00	.00	.00	.50	.00	.25	.00	.00	.00	.25	.33	.25
8	.50	.25	.50	.00	.00	.00	1.00	.25	.25	1.00	1.00	1.00
9	.30	.35	.34	.36	.38	.45	.28	.19	.17	.41	.43	.40
10	.10	.08	.06	.42	.31	.36	.11	.08	.13	.38	.32	.37
11	.06	.08	.07	.31	.32	.46	.19	.24	.17	.25	.27	.32

*The study required six independent observers to be in the field during the time of observation. To demonstrate their reliability with the observation system they each observed four times with two other observers. Three separate teachers were used in the observations. Each reliability observation lasted 20 minutes. Categories on which observers were especially unreliable are underlined.

Table 12. Reliability of Observers in Applying the TR Observation System in November, 1966, as this is Reflected in the Comparability of Criterion Measures that Derive from Simultaneous but Independent Observations

Criterion Measure	Observers			Observers			Observers			Observers		
	A	B	C	D	A	F	A	B	F	C	E	F
1	.07	.07	.01	.09	.10	.08	.05	.04	.04	.04	.02	.05
2	.92	.73	1.00	.80	.56	.58	1.00	.67	.25	.80	1.00	.83
3	.00	.00	.00	.00	.33	.00	.00	.00	.00	.00	.00	.00
4	.37	.46	.63	.32	.34	.50	.32	.44	.55	.25	.67	.61
5	.00	.67	.86	.29	.56	.40	1.00	.78	.60	.92	.96	.76
6	.67	1.00	.00	.00	.64	.50	.00	.00	.00	.00	1.00	.50
7	.25	.00	.25	.33	.14	.00	.33	.00	.00	.50	.00	.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.00	1.00	.00
9	.07	.13	.12	.11	.15	.17	.06	.19	.08	.30	.34	.35
10	.26	.26	.22	.21	.26	.19	.36	.33	.45	.44	.46	.48
11	.30	.27	.30	.45	.31	.55	.25	.32	.11	.22	.34	.34

Criterion Measure	Observers			Observers			Observers			Observers		
	C	D	E	A	B	D	B	C	E	D	E	F
1	.03	.06	.05	.09	.18	.10	.08	.11	.05	.08	.03	.07
2	.67	.83	.50	1.00	.67	.75	.70	.83	.80	.71	1.00	.50
3	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
4	.80	.14	.20	.12	.10	.27	.29	.10	.29	.75	.34	.36
5	.85	.90	.60	.75	.72	.96	.75	.76	.88	.94	.93	.97
6	.00	.00	1.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00
7	.00	.00	.00	.50	.00	.00	.00	.00	.00	.00	.00	.00
8	1.00	1.00	.00	.00	.00	.00	1.00	1.00	1.00	1.00	.33	1.00
9	.28	.19	.22	.54	.43	.51	.30	.26	.34	.42	.48	.58
10	.38	.43	.46	.08	.06	.06	.45	.39	.40	.11	.08	.09
11	.31	.31	.24	.25	.33	.19	.06	.06	.06	.19	.16	.14

Table 13. Intercorrelations for the Criterion Measures

		Measure											
		1	1.00	.05	.0004	-.23	-.13	-.198	-.05	-.35	.09	.09	.08
		2		1.00	-.07	-.08	-.17	-.16	-.01	-.23	.11	.19	-.29
S	T	3			1.00	.26	-.16	.12	.597	.13	-.09	-.09	.14
t	e	4				1.00	.30	.26	.15	.60	-.52	.24	.19
u	a	5					1.00	-.02	.02	-.01	.09	-.20	.18
d	c	6						1.00	.18	.36	-.18	-.05	.21
e	h	7							1.00	.15	-.09	-.17	.10
n	e	8								1.00	-.41	.35	.29
t	r	9									1.00	-.35	-.16
	s	10										1.00	-.30
		11											1.00

		Measure											
E		1	1.00	-.106	.062	-.573	-.764	-.468	.136	-.417	.145	.122	.098
x		2		1.00	.108	.021	-.091	.071	.0399	-.121	-.223	.23	-.090
p	T	3			1.00	-.097	-.293	.078	-.017	-.069	-.034	.151	.422
e	e	4				1.00	.478	.32	-.34	.29	-.198	.063	-.109
r	a	5					1.00	.262	-.134	.377	.126	-.206	-.081
i	c	6						1.00	.019	.23	.109	-.058	-.207
e	h	7							1.00	-.025	.175	.0028	-.12
n	e	8								1.00	.126	.24	-.31
c	r	9									1.00	-.34	-.29
e	s	10										1.00	.197
d		11											1.00

Classroom observations. The same procedures were followed in making classroom observations as were followed in the parent study. Each subject was observed for nine 30-minute periods during the final two weeks of her student-teaching experience. Observations were made on days approximately one week apart by three different observers. Three 30-minute observations were made each day; two in the morning and one in the afternoon. During each observation period, subjects had primary teaching responsibilities in their rooms. Morning observation periods were characterized by relatively structured activities involving students in group settings. Afternoon periods, on the other hand, were generally characterized by unstructured individual activities. Such a schedule was devised to obtain a ratio of observations of teacher behavior in structured and unstructured activities roughly equivalent to that found in regular school activities.

Prior to actual observation, participating school personnel and college supervisors were oriented to the project and procedures to be employed. In addition, a practice observation was made in each subject's room one week prior to actual observations. After the practice observation, the supervising teacher, subject, and students were permitted to ask questions and express concerns regarding the observation procedure.

When observers arrived to record actual observations, they spent ten or fifteen minutes becoming familiar with the nature of interaction in the classroom, the setting, the traffic patterns, etc. This was, in a sense, an "acclimatization" period for observers. Once observation began, it continued for 30 minutes uninterrupted. While observing, observers were seated in unobtrusive positions that enabled them to see the subject and hear all that she said to students. There was no interaction between observer and students or between observer and subject. When the 30-minute period of observation was complete, the observer quietly left the room, returning when the next observation period was scheduled.

Extension #1: The Addition of Situational Data to the Original Predictio. Scheme

As indicated previously, the prediction scheme in the parent study included only test scores: situational factors affecting the behavior being predicted were not taken into account. Factors such as unplanned events, composition of the class, physical conditions within the classroom and the nature of the activity in which teacher and learners engaged were not controlled. Since these are likely to be significant determinants of teaching behavior, the present study attempted to include them in it. The aim of the present effort was to obtain prototypic measures of such factors and include them in the prediction scheme as control variables to see if their inclusion would significantly increase the amount of variance accounted for in the criterion measures.

Toward this end, seven dimensions of the classroom setting were identified: (1) the subject matter and the activity being pursued, (2) the organization of the classroom, for example, small study groups, individuals around a large work table, individuals at their desks, (3) the number of learners in the classroom, (4) the general characteristics of the learners in the classroom, for example, their personality characteristics, their capabilities, age, and sex, (5) the physical characteristics of the classroom, for example, the space available per learner, the presence of individual desks or tables, heat, ventilation, lighting, the proximity to activity on the playground or in the halls, (6) the philosophy of the school administration, particularly the building principal, in relation to classroom activity, and (7) unplanned events which are disruptive to planned learning experiences, for example, a fire drill, an unanticipated visitor, a child becoming ill, building repair or workmen's activity nearby. Measures for all of these factors were developed. Two of them, the subject matter and activity in which the class is involved, and the organization of the classroom, are described in connection with and at the same time that teacher and learner behavior are described; that is, they are part of the observation system (sre Figure 2).

SUBJECT____________________ OBSERVATION 1 2 3

OBSERVER____________________ PAGE____________

DATE____________________

Classroom Structure	Activities and Topics	Progressive Record of Teacher-Learner Interaction

Figure 2. The form on which the categories descriptive of teacher-learner interaction are recorded.

A diary record of the unusual or unplanned events that occur during the day on which the observations are made is kept by the teacher. All of the other setting measures, that is, the number of children in the class and their characteristics, the physical characteristics of the classroom, and the philosophy of the school administration in relation to the activities that take place in the classroom, are obtained through interview, either prior to or subsequent to the observation. In the paragraphs which follow, each of the situational measures are described briefly.

Subject matter, activity, and classroom organization. The subject matter in which a class is involved, the activity being pursued within that subject matter, and the classroom organization that accompanies it are recorded at the same time and on the same recording sheet as is the teacher-learner interaction (see Figure 2). Each observation begins with a notation as to subject matter, activity, and classroom organization, and these notations continue opposite the recording of the interaction that is occurring throughout the observation period. Time also is noted so that it becomes possible to identify the length of time spent within any given activity, classroom organization, etc. By including time, activity, classroom organization and subject matter in the observation record it is possible to analyze teacher-learner interaction against any or all of these factors.

Number and characteristics of children in a classroom, the physical characteristics of a classroom, and the philosophy of the school administration toward conduct in the classroom. As indicated above, information on these variables is obtained through an interview with the teacher. The specific items in the interview schedule are listed in Figure 3. The items included in the schedule were identified by elementary school teachers as

Figure 3. The interview schedule used in obtaining a description of the situational factors affecting the management behavior of teachers.

TEACHER______________________________

GRADE LEVEL__________________________

DATE________________________________

I CLASSROOM RELATED FACTORS

A. Physical Features of the Classroom

1. Size of room in relation to size of class.

a) square footage

b) teacher's feelings about adequacy of space

2. Seating arrangements in the room, i.e., tables and chairs vs. desks, etc. (describe)

3. Facilities for toilet and drinking (if present, describe)

4. Susceptibility of room to noise and student traffic. (Teacher's estimate; if susceptible, have teacher describe the nature and/or amount.)

5. Availability of educational materials, teaching aids, etc. in the room (teacher's estimate of adequacy).

B. Characteristics of the Class

1. Number of students in the class, plus the number absent on day of observation.

2. Boy-girl ratio.

3. Number of exceptional children in the class, e.g., intellectually, physically, and emotionally handicapped, intellectually superior, etc. (List number by class of exceptionality.)

4. The number of children who are habitually disruptive of the class plus number absent on days of observation (obtain from teacher's records).

Figure 3, Continued

5. Principal's estimation of the socio-economic status of the families of the students in the school (provide one of three estimates: predominantly lower SEC, predominantly middle and/or upper middle SEC; fairly even cross-cutting of the lower and middle SEC).

6. Principal's estimate of the mobility of the student's families (provide one of three estimates: a high proportion mobile, e.g., service or migrant worker families; a high proportion permanent residents; a fairly even distribution of mobile and permanent residents).

II SYSTEM RELATED FACTORS

A. Official Policy Toward Classroom Discipline and Control

1. Policy toward noise in the classroom (describe; obtain through principal).

2. Policy toward the handling of "discipline problems" by teachers (describe; obtain through principal).

B. Classroom organization, e.g., self contained, cooperative or nongraded, team teaching, etc. (describe; obtain through principal).

C. Curricular innovations, e.g., the "new math," experimental biology courses, etc. (describe; obtain through principal).

factors which frequently and significantly influence that which occurs within their classrooms. Since the titles of the factors are self-explanatory, no further comment will be made about them. The interview is usually administered after the observation has been completed so as to obtain information on the number of children absent during the observation, but it may be administered before the observation if so desired. Also, the interview schedule, in the form of a questionnaire, may be given to the teacher to complete by herself.

Unanticipated events. One of the setting factors identified by teachers which often influences teacher-learner interaction is that of unanticipated events. These can range from a sudden snow storm or an unanticipated assembly to a child becoming ill or a stray dog finding his way into the room. By definition, an unusual event is one which interferes with that which is planned in relation to instruction. In order to obtain information as to the nature and occurrence of these events each teacher that is observed is asked to record at the end of the observation period any unanticipated events which occurred either prior to or during the time of observation that in her opinion had a significant influence upon that which occurred during the course of the observation. The recording form that is provided the teacher for this purpose appears as Figure 4.

Predictor measures derived from the descriptions of setting variables. Four global measures designed to reflect the complicating effects of setting factors upon the task of classroom management were derived from the descriptions of setting variables provided by the interview schedules outlined in Figures 3 and 4.

TEACHER______________________________

GRADE LEVEL__________________________

OBSERVATION DAY (circle day) 1 2 3

DATE_________________________________

It is well known by teachers that factors such as the temperature or ventilation of a classroom, the physical well-being of children, the anticipation of a special event or holiday, the appearance of an invited or uninvited animal, the occurrence of a fire or a co.istruction project nearby, or the well-being of the teacher herself can have a marked effect upon behavior occurring within the classroom. Since our research requires as "natural" a picture as possible of classroom behavior, would you please describe below any circumstances that you feel may have caused the behavior observed in your classroom to be different from that which usually occurs.

If unusual events did occur, would you indicate also the approximate time that they occurred.

The examples of unusual events cited above are, of course, only suggestive of the wide range of events which can affect a classroom. When you are thinking about that which may have affected behavior in your own classroom please feel free to include anything and everything that may have made it an "unusual" situation.

The observer will pick this record up from you at the close of the last observation period on each observation day.

Figure 4. The form for recording unusual events which affected or could have affected behavior in the classroom during the time of observation.

1. A descriptor of the physical setting. Items IA, 2, 3, 4, 5 and II B and C from the interview schedule outlined in Figure 3 were combined into a global, 3-point scale designed to reflect the teacher's feelings or judgment about the adequacy of the physical features of the setting in which she taught. A score of zero on the scale indicated that, all factors considered, the physical characteristics of the classroom seemed to be somewhat handicapping from the point of view of classroom management; a score of 1 indicated that they were neither particularly handicapping nor particularly facilitating; and a score of 2 indicated that they were facilitory of the management task.

2. A descriptor of the administrative setting. Item II A 1 from the interview schedule outlined in Figure 3 provided the descriptive data from which this measure was derived. The measure was scored in the same way as measure 1, namely, a score of zero indicated that the administrative setting handicapped the task of classroom management, a score of 1 indicated that it was neither particularly handicapping or facilitating, and a score of 2 indicated that it was facilitating.

3. A descriptor of the characteristics of the class. Items I B 1, 2, 4, 5 and 6 from the interview schedule outlined in Figure 3 provided the descriptive data from which this measure was derived. In contrast to measures 1 and 2, measure 3 represented an algebraic summation of each of the five factors that fed into the measure. Before summation each of the five factors was scored from 0 to 2, following the same rationale as was used in scoring measures 1 and 2. This procedure

permitted measure 3 to have a range of 0 to 10. The criteria followed in arriving at the various subscores were:

a) fewer than 18 students in a class yielded a score of 2 and more than 32 yielded a score of zero;

b) a ratio of girls to boys in the class that favored the girls, i.e., greater than a 1:1 ratio, yielded a score of 2 and a ratio of boys to girls that exceeded 2:1 yielded a score of zero;

c) a class in which no children were habitually disruptive received a score of 2, whereas a class which had 3 or more children in it who were habitually disruptive received a score of zero;

d) a class which was made up of children from predominantly middle and upper class families received a score of 2 whereas a class which was made up of children predominantly from either lower-lower or upper class children received a score of zero; and

e) a class which was made up of children from families which were predominantly permanent in the community received a score of 2 whereas a class which was made up of children from families which were predominantly mobile received a score of zero.

4. A descriptor of unusual events. The interview schedule outlined in Figure 4 provided the descriptive information from which this measure was derived. Like measure 1, the information obtained from the interview was forced into a single three-point scale describing the extensiveness and/or criticalness of the unusual events that occurred during a day that classroom observations were made. If no unusual events occurred, a value of zero was assigned; if three or more unusual events occurred, or if a single event was extremely disruptive, a value of 2 was assigned.

A fifth measure descriptive of the classroom setting was also used in the study, namely, a mearure describing the instances of behavior that were disruptive to the class during the classroom observations. This measure was derived from observational data rather than interview data and consisted simply of the number of such instances that occurred. Four scores actually derived from this measure: 1) a score representing the total number of such instances, 2) a score representing the number of incidents that were non-academic in nature and directed toward the teacher, 3) a score representing the number of incidents of the same kind directed to other children, and 4) the number of instances that had an academic focus but which were sufficiently inappropriate in nature to cause them to be disruptive.

By combining the four measures derived from the observation data and the four derived from the interview data, a total of eight setting or situational measures were available for use in the study as predictors. The intercorrelations for these measures appear in Table 14.

Extension #2: The Repetition of the Study, Including the Use of Situational Measures, with Experienced Primary Grade Teachers

With one exception the same measures and procedures as outlined in the replication and extension of the parent study with student teachers were followed in the extension of the parent study to experienced teachers. The one exception occurred in relation to the time periods in which tests could be administered and observations could be made. In contrast to the rather rigid schedule of testing at the end of the term prior to student teaching and observation within the last two weeks of the student teaching experience, the experienced teachers could be tested and observed at any time.

Subjects in the study were thirty-nine experienced primary grade teachers drawn from the school districts in which the student teachers who participated in the study did their student teaching, and in the same proportion. Only those who volunteered for the project were

Table 14. Intercorrelations for the Situational Measures

	Measure	1	2	3	5a	b	c	d	4
Experienced Teachers	1	1.00	.169	-.355	-.107	-.093	-.027	-.12	.16
	2		1.00	-.10	-.11	-.15	.17	-.24	.034
	3			1.00	-.30	-.196	-.14	-.34	-.11
	5a				1.00	.82	.65	.76	-.23
	b					1.00	.32	.51	-.14
	c						1.00	.17	-.23
	d							1.00	-.16
	4								1.00

	Measure	1	2	3	5a	b	c	d	4
Student Teachers	1	1.00	-.03	-.27	-.04	-.12	.27	-.16	.02
	2		1.00	-.16	-.23	-.23	-.06	-.13	-.12
	3			1.00	-.12	-.02	-.10	-.08	-.40
	5a				1.00	.77	.62	.47	.39
	b					1.00	.44	-.03	.28
	c						1.00	-.15	.18
	d							1.00	.27
	4								1.00

included in it. No restrictions were placed upon length of teaching experience beyond having taught for at least one year prior to taking part in the study. The testing and observations required by the study were fitted to the convenience of the teachers within a given district and to the time schedule of project personnel.

An Investigation of the Effects on Prediction of Using Behavioral Samples of Differing Lengths in Obtaining the Criterion Measures

As indicated previously, the rationale underlying the inclusion of an investigation of this kind in the present study rests upon the fact that the study depends upon behavioral sampling for its criterion measures, but as yet there is no conclusi e evidence as to the length or number or distribution of behavior samples needed in order to obtain stable or representative criterion measures. The problem derives from the fact that teacher behavior is situation bound, that is, that on any given day or on different occasions within a day situational influences can be expected to bring about a great deal of variation in observed behavior. This problem is not unlike other sampling problems encountered within the behavioral sciences, and it is generally assumed that situational influences can be balanced out when the sample of observed behavior is lengthened. The question still to be answered, however, is "What are the fewest number of observations required to obtain a performance measure reflective of a balance of situational influences?"

The plan of the present investigation was relatively simple: compare the magnitude of the correlations derived from the prediction scheme with criterion measures based upon 1 hour, 2 hours and 3 hours of observation time, respectively. This was to be done for both student and experienced teachers. While it was recognized that such a design was far too simple to answer the question of behavioral sampling in any final sense, it was felt that it was sufficient to provide information that would be of use in the present study and in the design of future studies on the issue. Criterion measures in the parent study were based upon two hours of observation.

The Procedure Followed to Make Participation in the Study as Meaningful and as Valuable as Possible

Researchers in education are frequently accused of advantage taking or "usury" in their dealings with institutions in the course of their research, for often they fail to maximize the returns that come from their research for these institutions and the personnel within them. In many cases this has led to hesitancy or resistance on the part of school personnel to get involved in educational research, or even to the closing of entire school districts to researchers. Because of the heavy demands that the present project made upon participants, special attention was directed to making participation in it maximally beneficial. Toward this end a two-pronged procedure was worked out: a) make it possible to receive an hour of course credit for participation in the study, and b) give as much information as possible about the study to those who participated in it.[5] Three types of information were provided: 1) a history of the research that led to the study, its significance, and the contribution which the present study could conceivably make to educational practice (this was provided during "recruiting meetings" prior to participation), 2) a visual summary of each participant's behavior, in the form of profiles, for each of the nine half-hour periods they were observed in the study (this was provided in the form of a seminar during the term following participation), and 3) a written summary of results at the completion of the study. The addition of the seminar at which teachers could view and discuss detailed records of their own behavior proved to be highly satisfactory, though costly of time and energy, and is recommended as a worthwhile procedure to follow when information of this kind is available. A copy of a memorandum describing the procedure and a copy of a behavioral profile given to teachers for discussion in the seminar appear as Attachments 2 and 3 respectively.

[5] Special thanks are due to Dr. Jack Hall who helped work out the "Information Feedback" procedure that is described below and apply it within the context of the Elementary Teacher Education Program at Oregon State University.

Chapter IV

RESULTS AND DISCUSSION

The data have been ordered according to the four issues investigated in the study:

1) Can the results obtained by Schalock, Beaird and Simmons (1964) be replicated?

2) Can the per cent of variance accounted for in teaching behavior by the prediction scheme used in the Schalock, Beaird and Simmons study be increased by including in the prediction equation measures of situational factors that affect teaching behavior?

3) Do the results obtained in (1) and (2) above with student teachers vary when the methodology is applied to experienced teachers? and

4) Do the results obtained in (1), (2), and (3) above vary as the behavioral samples on which the criterion measures are based vary?

According to this ordering three separate analyses would have had to have been run on each of the first three questions in order to answer question 4; that is, each question would have had to have been analyzed using criterion measures based on 1, 2, and 3 days of observation respectively. Operationally, this would have required 528 regression runs to be made, a cumbersome and costly procedure. In an effort to short-cut this process, and still obtain the essential information desired on the relationship between length of behavioral sample and stability of criterion measure, a straightforward analysis of the differences obtained in criterion measures as a function of length of behavioral sample was undertaken. The rationale underlying the analysis was one of economy: if no differences were found in measures as a function of length of behavioral sample then not only could the tripling of regression runs be avoided but a smaller amount of data (1 or 2 days' data vs. that of 3 days') be handled in preparing the needed regression runs. The criterion used in the analysis against which to compare differences was the 3 day behavioral sample.

Since whatever regression runs to be made in the study depended upon the results of this analysis, it was undertaken first.

An Analysis of the Relationship Between Length of Behavioral Sample and Stability of Criterion Measures

It will be recalled that three different behavioral samples were obtained on subjects: a day 1 sample (2 one-half hour observations on a given teacher with a given class, two in the morning and one in the afternoon) a day 1 + day 2 sample (both on the same teachers with the same class), and a day 1 + day 2 + day 3 sample (all on the same teacher with the same class). To determine the length of behavioral sample required to insure stability of criterion data, each individual was assigned three scores for each criterion measure. The first score was determined by summarizing the observations made on the first day, the second score by summarizing the observations made on the first two days, and the third score by summarizing the observational data obtained during all three days. Using the latter score as a standard, the first two scores were compared against it to determine the feasibility of utilizing shorter behavioral samples. The rationale underlying this procedure was straightforward: if it were found that scores based on one or two days of observation varied significantly from the final score one would have to conclude that a one or two day observation was not sufficient to insure a stable measure of teacher behavior. Also, if this were the case, the question of the length of behavioral sample required to obtain stability would remain unanswered. On the other hand, if it were found that either a one or two day sample of behavior provided essentially the same measures as did the three day standard then one would be justified in using either the one, two or three day sample in deriving criterion measures.

The data that derived from the analysis appear in Table 15. It will be seen from these data that for both the student teacher and experienced teacher samples, scores based on a single day's observation varied significantly from scores based on three days of observation. This was not the case, however, for scores based upon two days of observation. For both samples observed in the study no statistically significant differences were noted between scores based on two days of observation and those based on three days of observation. Thus, for purposes of the present study, it was concluded that utilization of criterion scores based on a single day's observation was not warrented, but that the utilization of two days of observation, when each day's observation time is based upon three one-half hour observational settings, provides as adequate or stable a picture of teacher behavior as do three days of observation.

Table 15. Mean Scores for Criterion Measures Obtained from Observational Periods of Different Lengths.

Criterion	Student Teachers			Experienced Teachers		
	1 Day	2 Days	3 Days	1 Day	2 Days	3 Days
1	.067	.067	.070	.068*	.077	.075
2	.547	.555	.551	.468	.384	.423
3	.251	.253	.253	.191*	.210	.221
4	.218*	.223	.228	.252	.234	.216
5	.835	.835	.839	.796	.798	.820
6	.745*	.769	.799	.623*	.649	.674
7	.185*	.251	.262	.316	.325	.359
8	.606*	.753	.703	.624*	.737	.750
10	.334*	.357	.356	.252*	.277	.284
11	.253	.259	.259	.321	.329	.328
12	.404	.407	.408	.366	.372	.349

* Difference between Day 1 and Day 3 significant at .05 level.
No significant differences appeared between Day 2 and Day 3.

On the basis of these data, two decisions were made: 1) to calculate criterion measures on the basis of day 1 + day 2 data (the same data base as used in the parent study), and 2) run only one set of regression analyses, instead of three, in replicating and extending the study.

While these data provided a basis for firm decision making in the present study, and supported the use of the two day sample in the parent study, they are not sufficient in and of themselves to answer the full range of questions that need answering in relation to the issue of behavior sampling. They do indicate that a 2 or 3 day sample is different from a single day, but how would a 2 day sample compare to a five or ten day sample? More importantly, what difference would it make if the basis for behavioral sampling were activities or subject matter topics or stages in the development of topics? These and other questions ultimately must be answered if the study of behavior in situation is to be undertaken seriously. The results of the present study represent a start in this direction, but a great deal more needs to be done.

A Comparison of the Results Obtained in the Present Study with the Results Obtained in the Schalock, Beaird and Simmons Study

Using the day 1 + day 2 observation sample as a basis for the calculation of criterion measures, analyses were run which essentially replicated the Schalock, Beaird and Simmons study. These data, and the data from the parent study, are presented in Table 16. It will be seen from these data that essentially the same results were obtained in the two studies, though the Word Test in the second study tended to yield higher correlations than it did in the first and the Simulation Tests tended to yield lower correlations.

Table 16. Percentage of Variance Accounted for in Student Teacher Behavior in the Parent and Replication Studies*

Criterion Measure	Word Test		Film Test		Simulation Test		Composite of "Best" Predictors	
	1st Study	2nd Study	1st Study	2nd Study	1st Study	2nd Study	1st Study	2nd Study
1	.303	.570	.281	.784	.303	.133	.563	.556
2	.384	.379	.476	.212	.360	.332	.504	.378
*3		.481		.194		.359		.655
*4		.287		.256		.112		.309
5	.292	.466	.384	.438	.292	.188	.504	.419
6	.221	.344	.384	.412	.533	.301	.504	.675
*7		.206		.328		.375		.526
8	.348	.264	.292	.353	.384	.244	.593	.545
9	.270	.284	.449	.203	.384	.023	.608	.365
10	.221	.419	.410	.310	.397	.164	.757	.480
11	.303	.436	.360	.340	.449	.162	.723	.539

*Criterion measures that were new to the replication study.

It will be recalled that in the parent study the Simulation Test was consistently the most powerful predictor of the three; in the replication study it was consistently the least powerful. As would be expected, because of the decreased effectiveness of the Simulation Test as a predictor, the composite measure also decreased in its predictive effectiveness.

These data are at one and the same time encouraging and disappointing. On the encouraging side is the fact that the Word and Film Tests maintained themselves as fairly adequate predictors of behavior in situation. On the discouraging side is the fact that the Simulation Test failed to replicate in its effectiveness. This is discouraging not only from the point of view of losing a potentially powerful measuring device, but from the point of view of its implications for test theory generally. It will be recalled that the hypothesis tested in the parent study was that as tests became more lifelike in their stimulus and response properties effectiveness of prediction would increase. In general the hypothesis was supported by the study. The new data indicate that this may not be so, especially in light of the strong showing of the Word Test. Whatever the long range conclusion regarding the hypothesis will be, it is clear that at this point in time it does not have unequivocal support.

While recognizing this, it also needs to be recognized that several potential sources of error entered the Simulation data in the replication study (see pp. 24-27) and it could be that the results are simply reflective of that error. The results with the Word and Film Tests would seem to support such an interpretation, for they do essentially replicate. Assuming this to be a genuine possibility, and recognizing that the hypothesis tested in the parent study is not only attractive logically but has in fact once been supported, the better part of wisdom would seem to be to maintain the hypothesis and set up a series of studies to test it more fully.

An Analysis of the Effects of Adding to the Prediction Scheme Descriptors of the Setting in Which Criterion Measures Were Obtained

The data reflecting the consequences of adding to the prediction scheme measures descriptive of the setting within which teacher behavior occurred are presented in Table 17. In making these calculations all eight setting measures (see pp. 41-45) were used as predictors. In the prediction runs involving the Word, Film and Simulation Tests individually, all of the setting measures were included; in the prediction run involving the composite of "best" predictors only those setting measures that were in fact "best" predictors in previous runs were included.

Table 17. Percentage of Variance Accounted for in Student Teacher Behavior by Tests and Situational Descriptors

Criterion Measure	Word Test	Word T & Situation Meas's	Film Test	Film T & Situation Meas's	Simulation Test	Sim. T & Situation Meas's	Composite of Tests	Compos. & Situation Meas's
1	.570	.644	.784	.865	.133	.252	.556	.551
2	.379	.534	.212	.318	.332	.438	.378	.584
3	.481	.576	.194	.449	.359	.518	.655	.587
4	.287	.581	.256	.561	.112	.429	.309	.397
5	.466	.545	.438	.517	.188	.335	.419	.475
6	.344	.663	.412	.655	.301	.494	.675	.571
7	.206	.489	.328	.565	.375	.565	.526	.568
8	.264	.772	.353	.675	.244	.626	.545	.472
9	.284	.693	.203	.585	.023	.653	.365	.264
10	.419	.684	.310	.619	.164	.484	.480	.553
11	.436	.685	.340	.553	.162	.530	.539	.415

It will be seen from the data in Table 17 that a surprising amount of variance in student teacher behavior was accounted for by adding descriptors of the situation in which they were behaving to the prediction scheme. Without exception, at least when dealing with the Word, Film, or Simulation Tests independently, the amount of variance accounted for in criterion measures was increased when the situational descriptors were added to the prediction scheme. In some cases the amount of variance accounted for was equivalent to that accounted for by the formal predictor measures, and in some cases it actually exceeded that accounted for by the formal measures. When added to the Word Test the situational descriptors accounted for as much of the variance in three measures (measures 4, 6 and 7) and more of the variance in two (measures 9 and 10) than did the subscales of the Word Test itself. The same was found to be the case with the Film Test, though for somewhat different measures: as much variance was accounted for by the situational descriptors in measures 4, 8 and 10 and more in measures 3 and 9. because of the generally low predictive power of the Simulation Test the situational descriptors accounted for variance equal in amount to the Simulation Test in two measures (measures 1 and 5) and more in five (measures 4, 8, 9, 10, and 11). It is interesting to note that the three measures most susceptible to setting influence (measures 4, 9 and 10) were, respectively, Orientation to the Use of Positive Reinforcement, Teacher Approachableness, and Individual vs. Group Focus. Considering that the setting measures were few and only roughly conceived, these are remarkable findings, and suggest yet another line of research to be undertaken if prediction to behavior in situation is to be pursued seriously.

As expected, because of the procedure followed in selecting predictors, the same gains in predictive power did not appear when the situational descriptors were added to the prediction scheme.

An Analysis of the Results of the Replication Study Extended to Experienced Teachers

Using the same observational base for the calculation of criterion measures, the same predictor measures, etc., the design of the replication study with student teachers was extended to a sampling of experienced teachers. These teachers were drawn from the same schools in which the student teachers taught and from the same grade levels. Table 18 contains the data that derived from this extension. Table 19 contains a comparison of these data to those derived in the replication study with student teachers.

Table 18. Per Cent of Variance Accounted for in Experienced Teacher Behavior Without Regard for Situational Factors

Criterion	Word	Film	Simulation	Composite
1	.400	.319	.226	.582
2	.442	.333	.243	.651
3	.601	.366	.100	.685
4	.466	.169	.300	.556
5	.342	.251	.144	.600
6	.385	.250	.258	.499
7	.365	.262	.274	.611
8	.304	.306	.074	.434
10	.366	.080	.140	.350
11	.229	.272	.030	.409
12	.408	.295	.052	.617

Table 19. A Comparison of the Per Cent of Variance Accounted for in Student and Experienced Teacher Behavior Without Regard for Situational Factors

Criterion	Word		Film		Simulation		Composite	
	Stud.	Exper.	Stud.	Exper.	Stud.	Exper.	Stud.	Exper.
1	.570	.400	.784	.319	.133	.226	.556	.582
2	.379	.442	.212	.333	.332	.243	.378	.651
3	.481	.601	.194	.366	.359	.100	.655	.685
4	.287	.466	.256	.169	.112	.300	.309	.556
5	.466	.342	.438	.251	.188	.144	.419	.600
6	.344	.385	.412	.250	.301	.258	.675	.499
7	.206	.365	.328	.262	.375	.274	.526	.611
8	.264	.304	.353	.306	.244	.074	.545	.434
10	.284	.366	.203	.080	.023	.140	.365	.350
11	.419	.229	.310	.272	.164	.030	.480	.409
12	.436	.408	.340	.295	.162	.052	.539	.617

Two general observations can be made about these data: 1) they tend to follow the same general pattern observed in the student data, in that the Word and Film Tests accounted for a greater portion of variance than did the Simulation Test, and the composite measure accounted for slightly less variance than it did in the parent study, and 2) the Word and Film Tests were differentially effective with the student and experienced teacher samples. By and large the Word Test was a more effective predictor with the experienced teachers and the Film Test was a more effective predictor with the students. Both were unexpected outcomes. In entering the study it was anticipated that prediction would be consistently better for experienced teachers than for student teachers because their experience would permit them to respond to the situations depicted in the tests in ways which were similar to ways in which they had responded and/or tend to respond to comparable situations in the classroom. This expected relationship between background of experience and predictability of behavior obviously did not appear.

Even more surprising, at least at first blush, was the finding that the Word Test was a better predictor of experienced teacher behavior than the Film Test. As initially conceived, the theory of testing from which the predictive measures derived led to the expectation that the Film Test would be superior. In retrospect it appears that the theory is too simple. It may be, for example, that the theory holds *only for persons who have had a limited background of experience in classrooms*, and thereby have only a limited backlog of concrete referents to bring to a testing situation like that presented by the Word, Film, and Simulation Tests. For these people, the concrete referents provided by the Film and Simulation Tests may be an advantage; *for persons with a broad range of classroom experience the same referents may be a disadvantage*, for they may limit their perception to a single situation which may in fact not be representative of the situations with which they generally deal. If this should be true then one would expect that the Word Test, with its more general class of referents, to be more effective as a predictor for experienced teachers. Whatever the eventual explanation may be, the results of the comparative study between student and experienced teachers suggests that an approach to measurement that is maximally effective with one may not be maximally effective with the other.

An Analysis of the Effects of Adding to the Prediction Scheme with Experienced Teachers Descriptors of the Setting in which Criterion Measures Were Obtained

The data reflecting the consequences of adding to the prediction scheme measures descriptive of the setting within which experienced teacher behavior occurred are presented in Table 20. In making these calculations the setting measures were used as predictors in the same way they were used in the replication study with student teachers.

Table 20. Per Cent of Variance Accounted for in Experienced Teachers by Tests and Situational Descriptors

Criterion Measure	Word Test	Word T & Situation Meas's	Film Test	Film T & Situation Meas's	Simulation Test	Sim. T & Situation Meas's	Composite of Tests	Compos. & Situation Meas's
1	.400	.765	.319	.648	.226	.545	.582	.569
2	.442	.692	.333	.580	.243	.420	.651	.581
3	.601	.632	.366	.587	.100	.234	.685	.670
4	.466	.725	.169	.495	.300	.597	.556	.557
5	.342	.590	.251	.513	.144	.307	.600	.564
6	.385	.827	.250	.651	.258	.642	.499	.429
7	.365	.500	.262	.402	.274	.358	.611	.574
8	.304	.546	.306	.464	.074	.298	.434	.481
10	.366	.610	.080	.342	.140	.299	.350	.355
11	.229	.519	.272	.572	.030	.247	.409	.287
12	.408	.530	.295	.392	.052	.108	.617	.497

As with the student teacher data, a surprising amount of variance in experienced teacher behavior was accounted for by adding descriptors of the setting to the prediction scheme. As might be expected proportionately more variance was accounted for when these measures were combined with the Film and Simulation tests than when they were combined with the Word Test measures, but essentially the data replicate that obtained with student teachers. Taken together, the data on the effectiveness of situational descriptors as predictors leads to the obvious conclusion that if prediction to behavior in situation is to be undertaken seriously then a great deal of attention will need to be directed to the measurement and/or control of situational factors.

Chapter V

SUMMARY AND CONCLUSIONS

Recently completed research by Schalock, Beaird and Simmons (1964) on the predictive power of tests which use motion pictures as test stimuli suggested that a methodology may now be at hand which will permit the prediction of teaching behavior in the classroom. Using student teachers as subjects, Schalock et al. were able to demonstrate multiple correlations of .69 to .87 between scores on a battery of situational-response tests (tests which use motion picture representations of classroom situations as test stimuli) administered prior to student teaching and observational measures of their behavior in the classroom during student teaching. This represented an unusual accomplishment, for typically studies in the behavioral sciences have not been able to account for more than 50 per cent of the variance in any criterion that has been predicted to, and when the criterion has been as complex as teaching behavior, the level of prediction has nearly always been less. In the Schalock, Beaird and Simmons study at least 50 per cent of the variance was accounted for in each of the 15 separate criterion measures used (concrete behavior of teachers in the classroom) and as much as 75 per cent of the variance was accounted for in some.

Several factors, however, tended to limit the confidence that could be placed in the findings that came from the study. Two factors could have led to spuriously high correlations: 1) the final set of subscales used as predictors in the study were selected in a somewhat unorthodox manner; and 2) the number of subjects tested in the study (40) was small and the number of predictor variables used (18) was large. Over and against these sources of error was 1) the fact that the measures used in the study were prototypic in nature and therefore probably not as powerful as such measures could ultimately become. and 2) the failure to control for situational factors that interact with or are thought to influence teaching behavior in the classroom.

Given the data that derived from the study, and the many potential sources of error that accompanied them, a proposal was submitted immediately upon the completion of the study to the U.S. Office of Education for its replication and extension. Three factors led to the second proposal: (1) the essentially unprecedented results obtained in the parent study, (2) the numerous potential or real sources of error in it, and (3) the desire to avoid the pitfalls of uncritical test adoption, that is, the desire to forestall the users of tests from moving too quickly to adopt the instruments developed in the study for use in their own programs of research or evaluation. Since these instruments were new, and since the first predictive efforts with them were so promising, there was danger that the measures might be applied in areas where basis for their application did not exist.

Four major objectives guided the present study:

(1) to replicate the parent study;

(2) to extend the design of the parent study to experienced, primary grade teachers;

(3) to strengthen both replication studies by increasing the number of subjects used in each and including in them measures of situational variables that affect predictive accuracy; and

(4) to investigate the effects on prediction of deriving criterion measures from behavioral samples of varying lengths.

A fifth objective evolved as the study progressed, namely to strengthen the criterion measures used in it. This required extensive work on the observation system developed in the parent study, and led in part to a request for a 6-months extension of the study. A by-product of this extension is a monograph (see Attachment I) that provides an overview of the observational system that derived from the effort. The system is referred to generally as the Teaching Research System for the Description of Teaching Behavior in Context, and represents the most exhaustive measure of teaching behavior that currently is available.

A Summary of the Replication Study

Every effort was made to replicate the parent study in its exact detail. Subjects were drawn from the same population, the same predictive measures were used, criterion measures were equivalent though strengthened, and the same analyses were applied.

Thirty-nine senior women, majoring in elementary education with specialization in the primary grades at either Oregon State University or Oregon College of Education, served as subjects for the study.

Subjects were drawn from the pool of students who did their student teaching in the Winter and Spring terms of the 1965-66 academic year and the Fall and Winter terms of the 1966-67 academic year. Only students who volunteered to take part in the study and who did their student teaching within a 60-mile radius of Oregon State University were eligible for inclusion.

Four predictor tests, varying on a continuum of stimulus and response complexity, were used in the study: 1) a traditional paper-and-pencil attitude scale, where the test stimulus was a statement describing an orientation to the teaching function and response was defined by agreement or disagreement to the statement (The Minnesota Teacher Attitude Inventory), 2) a situational-response test where the test stimuli were written descriptions of filmed classroom situations and response was defined by agreement or disagreement to statements made in relation to the situational descriptions (The Word Test), 3) a situational-response test where the test stimuli were motion picture sequences of classroom situations and response was defined as in (2) above (The Film Test), and 4) a situational-response test where the test stimuli were also picture sequences of classroom situations but the response was free, i.e., the subject responded to the filmed situation _as if she were the teacher in the situation_ (The Simulation Test).

The predictor measures were administered in group settings at the close of the term that preceded the term in which the subjects did their student teaching. In contrast to the parent study, however, a totally random order of test presentation under supervised conditions was not followed: the Film Test and Simulation Test were always administered under supervised conditions, and either the Word Test or the

MTAI were administered under non-supervised conditions, i.e., at home. Furthermore, the Film and Simulation Tests were always administered in a one-two order, and the MTAI or Word Test always in a three-four order. While the Film and Simulation Tests were always assigned their order of presentation randomly, and the Word Test and the MTAI were always assigned to the non-supervised condition randomly, the inability to follow a totally random assignment of test order represented a source of error in the data and a departure from the procedure followed in the parent study.

Eleven measures descriptive of the interaction patterns of teachers and learners in the classroom served as criterion measures for the study. All were derived from the category descriptions of classroom interaction provided by the Teaching Research System for the Description of Teaching Behavior in Context (Schalock and Micek, 1968). Three features characterized the measures:

1) They were theoretically relevant, i.e., they related to dimensions of the model of teaching behavior used as a guide to instrument development throughout the study, and as a consequence exhibited a close tie to the predictive instruments that were developed;

2) They were complex in the sense that they represented a pooling of a number of conceptually related behaviors into a ratio or combination score. Theoretically this provided a more stable and comprehensive measure than would single classes of behavior; and

3) The measures took full advantage of the power of the observational system in the sense that they tied to (a) various classes of child behavior, (b) the teacher's response to classes of child behavior, and (c) the child's response to the teacher's behavior.

Eight of the eleven measures were comparable to those used in the parent study; three were new. These were added to replace the measures used in the earlier study that derived from rating scales.

The same procedures were followed in making classroom observations as were followed in the parent study. Each subject was observed for nine 30-minute periods during the final two weeks of her student teaching experience. Observations were made on days approximately one week apart by three different observers. Three 30-minute observations were made each day; two in the morning and one in the afternoon. During each observation period, subjects had primary teaching responsibilities in their rooms.

A Summary of Extension #1: The Addition of Situational Data to the Original Prediction Scheme

Situational factors affecting the behavior being predicted were not taken into account in the parent study. Factors such as unplanned events, composition of the class, physical conditions within the classroom and the nature of the activity in which teacher and learners engaged were not controlled. Since these are likely to be significant determinants of teaching behavior, the present study attempted to include them in it. The aim of the present effort was to obtain prototypic measures of such factors and include them in the prediction scheme as control variables to see if their inclusion would significantly increase the amount of variance accounted for in the criterion measures.

Two sets of measures were used in this respect: 1) those derived through interview with the teachers immediately after they had been observed, and 2) those derived from the records of classroom interaction made during the course of observation. Four global measures were derived from the interview data:

1) A descriptor of the physical setting, i.e., the space available per learner, the presence of individual desks or tables, heat, lighting, proximity to activity on the playground or in the halls, etc.;
2) A descriptor of the administrative setting, i.e., the philosophy of the building principal as to the nature of desirable or undesirable classroom activity;
3) A descriptor of the characteristics of the class, i.e., their socioeconimic status, the ratio of boys to girls, the number of habitually disruptive children in the class, etc.; and
4) A descriptor of unusual or unplanned events that were disruptive to planned learning experiences.

Four setting measures were also derived from the records of classroom interaction. These measures consisted simply of the occurence of behaviors that were disruptive to the class during the classroom observations. The four measures used in this respect were:

1) A score representing the total number of such instances,
2) A score representing the number of such incidents that were non-academic in nature and directed toward the teacher,

3) A score representing the number of incidents of the same kind directed to other children, and

4) The number of instances that had an academic focus but which were sufficiently inappropriate in nature to cause them to be disruptive.

By combining the four measures derived from the observation data and the four derived from the interview data, a total of eight setting or situational measures were available for use in the study as predictors. For purposes of analysis these were simply added to the set of formal predictors that derived from the MTAI, Word, Film and Simulation Tests.

A Summary of Extension #2: The Repetition of the Study, Including the Use of Situational Measures, with Experienced Primary Grade Teachers

With one exception the same measures and procedures as outlined in the replication and extension of the parent study with student teachers were followed in the extension of the parent study to experienced teachers. The one exception occurred in relation to the time periods in which tests could be administered and observations could be made. In contrast to the rather rigid schedule of testing at the end of the term prior to student teaching and observation within the last two weeks of the student teaching experience, the experienced teachers could be tested and observed at any time.

Subjects in the study were thirty-nine experienced primary grade teachers drawn from the school districts in which the student teachers who participated in the study did their student teaching, and in the same proportion. Only those who volunteered for the project were included in it. No restrictions were placed upon length of teaching experience beyond having taught for at least one year prior to taking part in the study. The testing and observations required by the study were fitted to the convenience of the teachers within a given district and to the time schedule of project personnel.

A Summary of the Investigation of the Effects on Prediction of Using Behavioral Samples of Differing Lengths in Obtaining the Criterion Measures

The rationale underlying the inclusion of an investigation of this kind in the present study rests upon the fact that the study depends upon behavioral sampling for its criterion measures but as yet there is no conclusive evidence as to the length or number or distribution of behavioral samples needed in order to obtain stable or representative criterion measures. The plan of the investigation was relatively simple: compare the magnitude of the correlations derived from the prediction scheme with criterion measures based upon 1 hour, 2 hours and 3 hours of observation time, respectively. This was done for both student and experienced teachers. Rather than pursue this plan, however, a straightforward comparative analysis of criterion measures was made to see if measures based upon 1, 2 or 3 day samples of behavior differed from one another. The 3 day sample was used as a standard in the analysis.

Conclusions

1. Were the results obtained by Schalock, Beaird and Simmons able to be replicated? Yes and no. Essentially the same results were obtained in the two studies with the Word and Film Tests, but different results were obtained with the Simulation Test. It will be recalled that in the parent study the Simulation Test was consistently the most powerful predictor of the three, whereas in the replication study it was consistently the least powerful. As would be expected, because of the decreased effectiveness of the Simulation Test as a predictor, the prediction scheme that combined the "best" subscale predictors from the three measures also decreased in its predictive effectiveness.

What do these results mean for the prediction of teaching behavior in the future? What do they mean for test theory? On both counts they are both encouraging and discouraging. On the encouraging side is the fact that the Word and Film Tests maintained themselves as fairly adequate predictors of behavior in situation, giving rise thereby to hope that teaching behavior may in time become a fairly predictable phenomenon. On the discouraging side is the fact that the Simulation Test failed to replicate in its effectiveness. This not only casts doubt on the trustworthiness of the measure that proved to be the most effective predictor in the parent study but also on the viability of the theoretical position underlying the study, that is, that as tests become more lifelike in their stimulus and response properties effectiveness of prediction should increase. Fortunately, these doubts may be more severe than they need to be, for there is reason to believe that the relative ineffectiveness of the Simulation

Test in the present effort was a function of the coding scheme applied to it rather than the test itself. Assuming this to be the case, it may well be that neither the test nor the hypothesis need to be discarded. The better part of wisdom would seem to be to maintain the hypothesis, devise new tests or new scoring procedures, and undertake a series of studies designed to test the methodology thoroughly.

2. Was the percent of variance accounted for in teaching behavior by the prediction scheme used in the Schalock, Beaird and Simmons study increased by including in the prediction equation measures of situational factors that affect teaching behavior? Unequivocally, yes! Without exception, at least when dealing with the Word, Film, or Simulation Tests independently, the amount of variance accounted for in criterion measures was increased when the situational descriptors were added to the prediction scheme. In some cases the amount of variance accounted for was equivalent to that accounted for by the formal predictor measures, and in some cases it actually exceeded that accounted for by those measures. For example, when added to the Word Test the situational descriptors accounted for as much of the variance in three measures and more of the variance in two than did the subscales of the Word Test itself. The same was found to be the case with the Film Test and the Simulation Test though in the latter case, because of the generally low predictive power of the Simulation Test, the situational descriptors accounted for equal variance in two measures and more in five. Considering that the setting measures were

were few and only roughly conceived these are remarkable findings, and suggest a critically needed line of research to be undertaken if prediction to behavior in situation is to be effective.

3. Do the results obtained in (1) and (2) above with student teachers vary when the methodology is applied to experienced teachers? Essentially no. By and large the same level of relationship was found between predictor measures and the classroom behavior of experienced teachers as was found between these measures and the classroom behavior of student teachers. Also, essentially the same results were obtained with experienced teachers when descriptors of the setting were added to the prediction scheme. The only major point of variance in the results obtained in the two studies was the finding that the Word and Film Tests were differentially effective with the student and experienced teacher samples. By and large the Word Test was a more effective predictor with the experienced teachers and the Film Test was a more effective predictor with the students.

These results were essentially unexpected. In entering the study it was anticipated that prediction would be consistently better for experienced teachers than for student teachers because their experience would permit them to respond to the situations depicted in the tests in ways which were similar to ways in which they had responded to comparable situations in the classroom. This expected relationship between background of experience and predictability of behavior obviously did not appear. Even more surprising, at least at first blush, was the finding that the Word Test was a better predictor of experienced teacher behavior than the Film Test. As initially conceived, the theory of testing from which the predictive measures derived led to the expectation that the Film Test would be superior. In retrospect it appears that the theory is too simple. It may be, for example, that the theory holds only for persons who have had a limited background of experience in classrooms, and thereby have only a limited backlog of concrete referents to bring to a testing situation like that presented by the Word, Film, and Simulation Tests. For these people, the concrete referents provided by the Film and Simulation Tests may be an advantage; for persons with a broad range of classroom experience the same referents may be a disadvantage for they may limit their perception to a single situation which may in fact not be representative of the situations with which they generally deal. If this should be true then one would expect the Word Test, with its more general class of referents, to be more effective as a predictor for experienced teachers. Whatever the eventual explanation may be, the results of the comparative study between student and experienced teachers suggests that an approach to measurement that is maximally effective with one may not be maximally effective with the other.

4. Do the results obtained in (1), (2), and (3) above vary as the behavioral samples on which the criterion measures are based vary? This question is unable to be answered directly, but on the basis of indirect evidence the answer would appear to be yes. In order to answer the question in the form it was asked three separate analyses would have had to have been run on each of the first three questions, that is, each question would have had to have been analyzed

using criterion measures based on 1, 2, and 3 days of observation respectively. Operationally, this would have required 528 regression runs to be made, a cumbersome and costly procedure. In an effort to short-cut this process, and still obtain the essential information desired on the relationship between length of behavioral sample and stability of criterion measure, a straightforward analysis of the differences obtained in criterion measures as a function of length of behavioral sample was undertaken. The rationale underlying the analysis was one of economy: if no differences were found in measures as a function of length of behavioral sample then not only could the tripling of regression runs be avoided but a smaller amount of data (1 or 2 days' data vs. that of 3 days') be handled in preparing the needed regression runs. The criterion used in the analysis against which to compare differences was the 3 day behavioral sample.

The data that derived from the analysis indicated that for both the student teacher and experienced teacher samples scores based on a single day's observation varied significantly from scores based on three days of observation. This was not the case, however, for scores based upon two days of observation. For both samples observed in the study no statistically significant differences were noted between scores based on two days of observation and those based on three days of observation. Thus, for purposes of the present study, it was concluded that the utilization of two days of observation, when each day's observation time was based upon three one-half hour observational settings, provides as adequate or stable a picture of teacher behavior as do three days of observation.

While these data provided a basis for firm decision making in the present study, and supported the use of the two day sample in the parent study, they are not sufficient in and of themselves to answer the full range of questions that need answering in relation to the issue of behavior sampling. They do indicate that a 2 or 3 day sample is different from a single day, but how would a 2 day sample compare to a five or ten day sample? More importantly, what difference would it make if the basis for behavioral sampling were activities or subject matter topics or stages in the development of topics? These and other questions ultimately must be answered if the study of behavior in situation is to be undertaken seriously. The results of the present study represent a start in this direction, but a great deal more needs to be done.

References

Aschner, Mary Jane McCue. The analysis of verbal interaction in the classroom. In A. A. Bellack (Ed.) Theory and Research in Teaching. New York: Teacher's College, Columbia University, 1963, pp. 53-78.

Bales, R. F. Interaction Process Analysis. Cambridge: Addison-Wesley, 1950.

Bellack, A. A., and Davitz, J. R., in collaboration with Kliebard, H. M., and Hyman, R. T. The language of the classroom: meanings communicated in high school teaching. Part I. U. S. Office of Education, Cooperative Research Project No. 2023. New York: Institute of Psychological Research, Teacher's College, Columbia University, 1963.

Bellack, A. A., in collaboration with Hyman, R. T., Smith, Frank L., Jr., and Kliebard, H. M. The language of the classroom: meanings communicated in high school teaching. Part II. U. S. Office of Education, Cooperative Research Project No. 2023. New York: Institute of Psychological Research, Teacher's College, Columbia University, 1965.

Bishop, Barbara M. A study of mother-child interaction. Psychol. Monogr., 1951 No. 11 (Whole No. 328).

Cronbach, L. J. Essentials of psychological testing. (2nd ed.) New York: Harpers, 1960.

Dunn, Frances E. Two methods for predicting the selection of a college major. J. counsel. Psychol., 1959, 6, 15-26.

Flanders, N. A. Teacher influence--pupil attitude and achievement. Final Report, Cooperative Research Branch Project #397, U. S. Office of Education, 1960.

Hughes, Marie, and Associates. A research report: Assessment of the quality of teaching in elementary schools. Salt Lake City: University of Utah Press, 1959.

Moustakas, C. E., Sigel, I., & Schalock, H. D. An objective method for the measurement and analysis of adult-child interaction. Child Develpm., 1956, 27, 109-134.

Schalock, H. Del. Issues in the conceptualization and measurement of teaching behavior. Paper read at the AERA meetings in New York, 1967, Mimeographed.

Schalock, H. D., Beaird, J. H., and Simmons, Helen. Motion pictures as test stimuli: An application of new media to the prediction of complex behavior. 1964, U. S. Office of Education, Title VII Project No. 971. Monmouth, Oregon: Teaching Research Division, Oregon State System of Higher Education.

Schalock, H. D. and Micek, S. The *TEACHING RESEARCH System for Describing Teaching Behavior in Context: A Training Manual.*

Schalock, H. D. and O'Neill, P. J. A reconceptualization of parent behavior. Research completed for the National Institute of Mental Health, U.S.P.H., Grant #M-3636(A). 1960.

Smith, B. O. and others. *A tentative report on the strategies of teaching.* U. S. Office of Education Cooperative Research Project No. 1640. Urbana: Bureau of Educational Research, College of Education, University of Illinois, 1964.

Taba, Hilda; Levine, S., and Elzey, F. F. *Thinking in elementary school children.* U. S. Office of Education, Cooperative Research Project No. 1574. San Francisco State College, 1964.

ATTACHMENT 2

March 3, 1966

MEMORANDUM

TO: Dean Zeran

FROM: Del Schalock

RE: A seminar in which participants in the prediction of classroom behavior project may enroll and receive 1 (one) hour of credit

As you know, Dr. Jim Beaird and I are replicating the research that we did several years ago with student teachers in the primary grades from OSU and OCE. You will recall that the research calls for the students to take a series of 4 tests prior to their student teaching experience and to be observed in the classroom for three half-hour periods on three separate days during the last few weeks of their student teaching experience. Observations are focused upon management behavior and involve the systematic description and recording of all management interchange between teacher and children.

Thus far, 18 student teachers from OSU have participated. Spring term students soon will be contacted about their interest in the study, and I anticipate that another 15 or so students will become involved. We also have approached the cooperating teachers about their participation in the project and approximately 2/3 of them wish to take part.

In order to make participation in the research as meaningful as possible, Dr. Jack Hall and I worked out a plan last fall whereby we prepare for each participant in the project "behavior profiles" for each half-hour that they are observed and discuss these with them at the end of the year within the framework of a 1 or 2 day seminar meeting. We anticipated that both student and cooperating teachers would attend the seminar, though it would be a voluntary matter, and that the discussion would center around individual profiles, the contrast between various student teacher and various cooperating teacher profiles, and the relationship between behavior patterns and situational factors. Also, individual teacher behavior will be related generally to a model of teaching behavior that has been developed in relation to the project.

It is also understood that I am to plan the seminar in cooperation with interested members of the Department of Elementary Education staff. This has awaited the completion of some writing on my part, but will take place during spring term.

In discussing the possibility of the seminar with Dr. Hall and his staff it was suggested that an hour of credit be attached to it with the thought that this would represent a formal record of the students' and cooperating teacher's participation in the project, as well as serving as an added inducement to participation. While the face-to-face contact within the seminar itself would not constitute a sufficient basis for an hour of credit, it was thought that the half day required in examination, the three days required in observation, and the informal contacts throughout a term with project staff would combine with the day of face-to-face contact in discussion to provide a legitimate basis for 1 hour of credit. All subjects have been approached from this standpoint and are expecting to register for the hour of credit spring term. Saturday, May 28th, has been set tentatively as the date for meeting with the student teachers while Saturday, June 4, has been set tentatively as the date to meet with the cooperating teachers.

In order that the participants may register for the seminar, a course number and class cards will have to be set up and available at spring term registration. I trust that this is possible, and I hope that it will not cause you inconvenience to arrange it at this time. Participation in the project has seemed to be a meaningful experience to the students and I think that they are looking forward to the discussion within the seminar.

If I may be of further help in arranging the seminar, please call upon me.

cc: Dr. Jack Hall

ATTACHMENT 3

PATTERN AND FOCUS OF CLASSROOM INTERACTION

TEACHER 65 B

ACTIVITY Health, Arithmetic

LENGTH OF TIME OBSERVED 40 Minutes

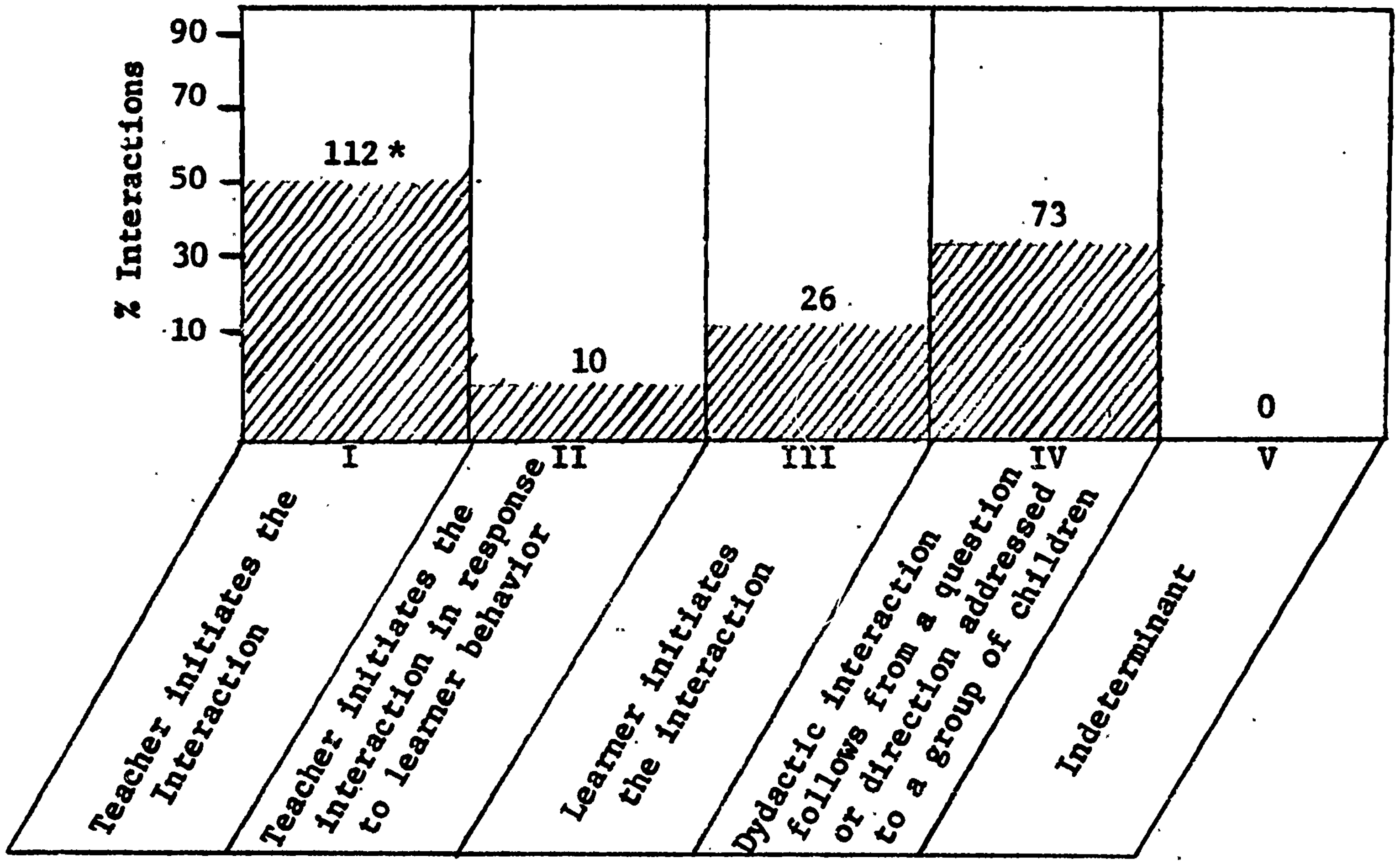


Figure 1. A graphic representation of the total number of interactive exchanges which occurred between teacher and learner(s), and a proportional analysis of the way these interactive exchanges were initiated.

* The numbers above the shaded areas represent the absolute frequency with which the category appeared.

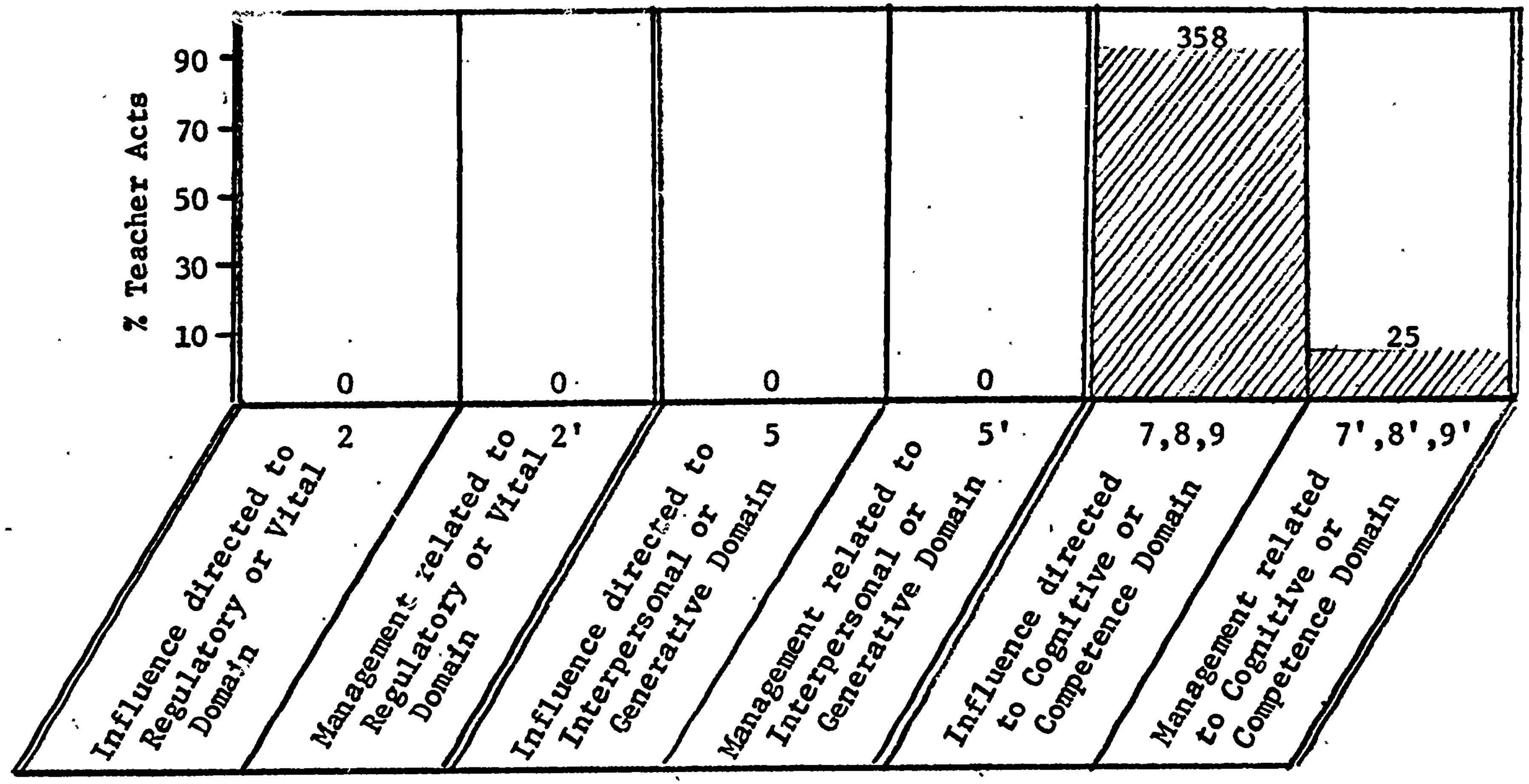


Figure 2. A graphic representation of the total number of teacher acts (messages) which were sent to the learner(s) and a proportional analysis of the distribution of these acts by the three major domains of human development on which a teacher focuses. A further classification scheme is used to identify: 1) those teacher acts that facilitate the development and/or maintenance of the domains of human development, i.e., those teacher acts that serve a management function, and 2) those teacher acts that directly influence the development and/or maintenance of the domains of human development, i.e., those teacher acts that serve a developmental function.*

* The prime symbol, e.g. 2', 5', 7', 8', 9', identifies all teacher acts which served a facilitory function.

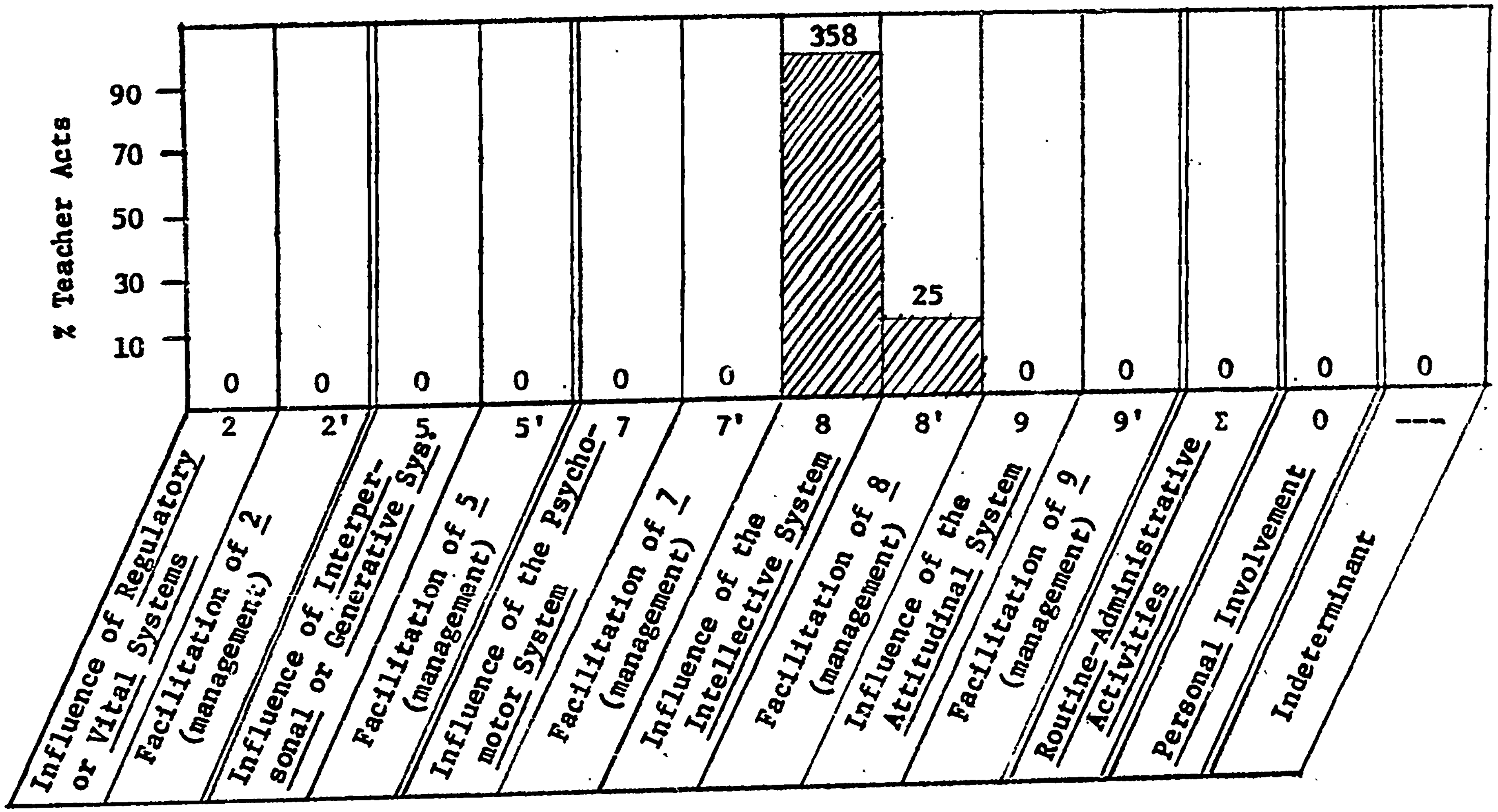


Figure 3. A graphic representation of the proportion of teacher acts as they relate to the teacher's FOCUS, (the areas of human development to which the teacher directs her attention). The areas within which a teacher focuses are: 1) the Regulatory or Vital Domain, 2) the Interpersonal or Generative Domain, 3) the Cognitive or Competence Domain. Within the Cognitive Domain three adaptive systems are identified: a) Psychomotor, b) Intellectual, and c) Attitudinal. In addition, the TR system records behavior that focuses upon Routine-Administrative activities, and the Personal Involvement of the Teacher

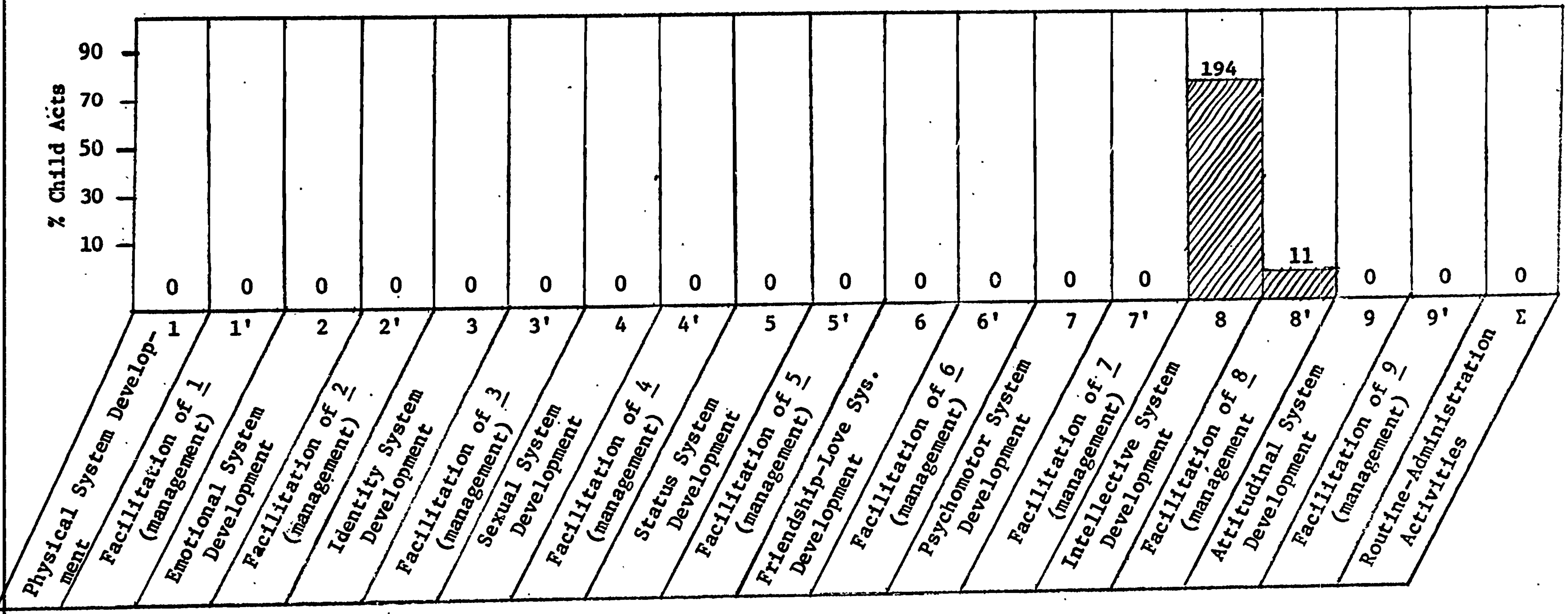


Figure 4. A graphic representation of the proportion of learner acts which were directed to the teacher as they relate to the learner's FOCUS (the area of human development in which the learner is involved). The areas within which a learner focuses correspond to the areas in which the teacher focuses, though the adaptive systems within the Vital and Generative Domains are made explicit when recording learner behavior whereas only the Domains are used when recording teacher behavior.

TEACHING OPERATIONS USED IN INSTRUCTION

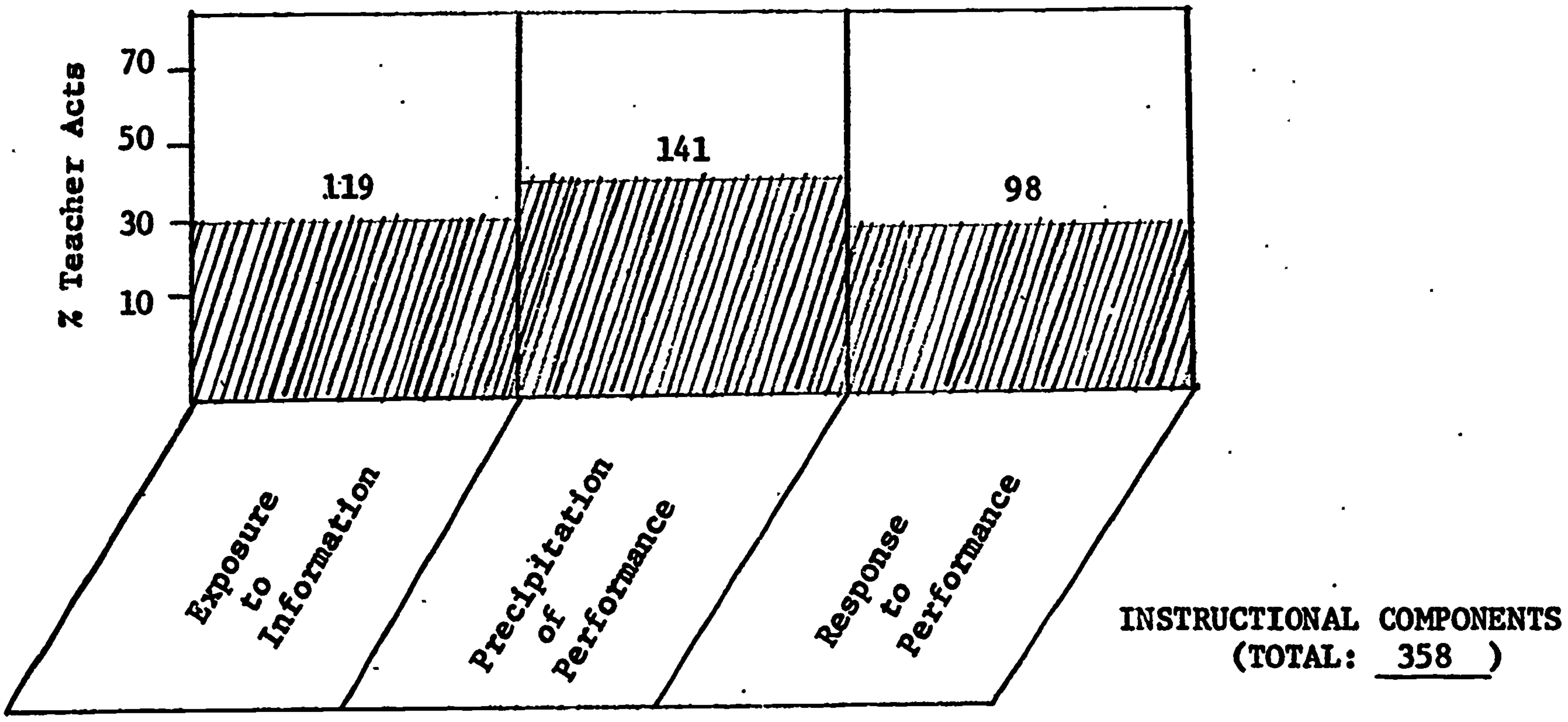


Figure 5. A graphic representation of the proportion of instructional acts used by a teacher which fall into one of the three major components of instruction. The component analysis represents the first level of analysis used in classifying TEACHING OPERATIONS.

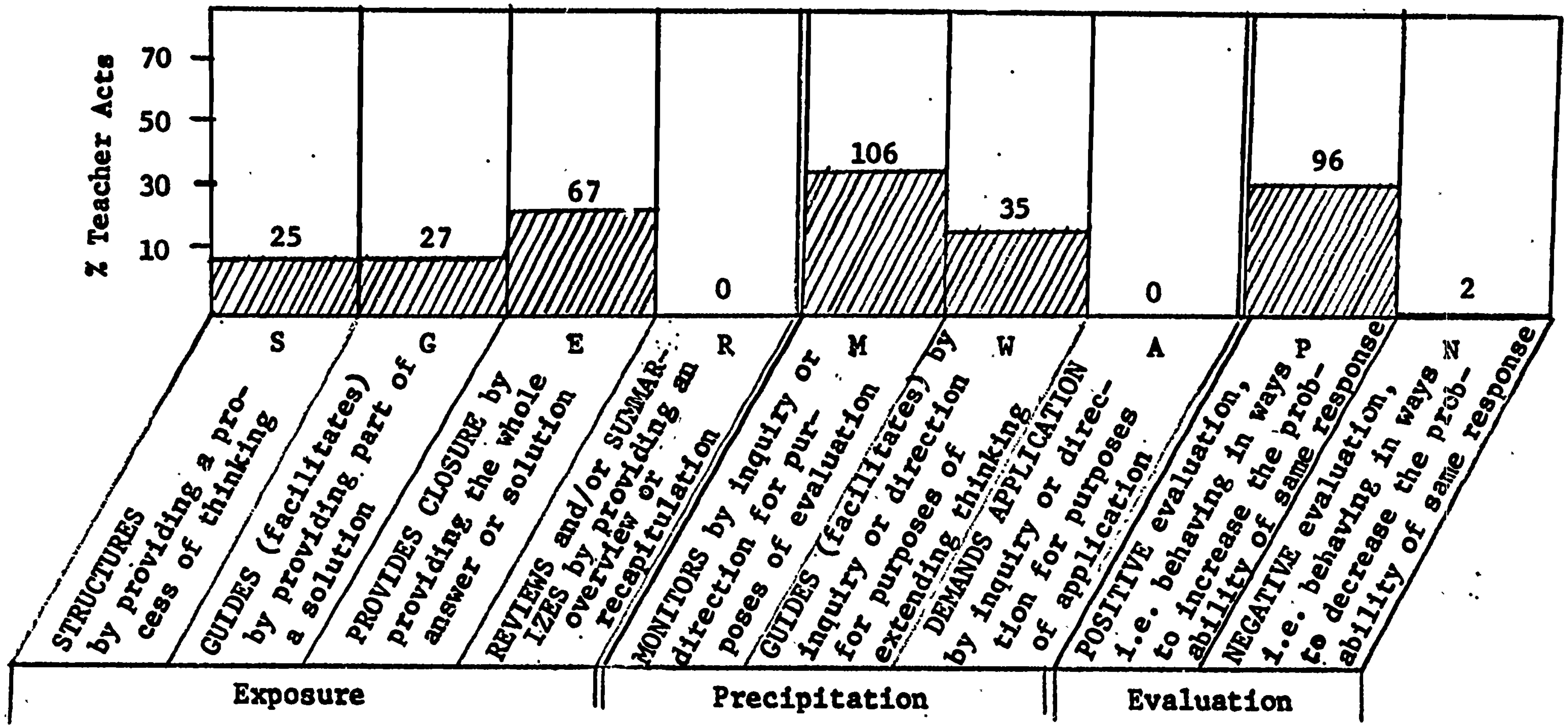


Figure 6. A graphic representation of the proportion of instructional acts used by a teacher which fall into one of the major functions served by teaching operations. The function analysis represents the second level of analysis used in classifying TEACHING OPERATIONS.

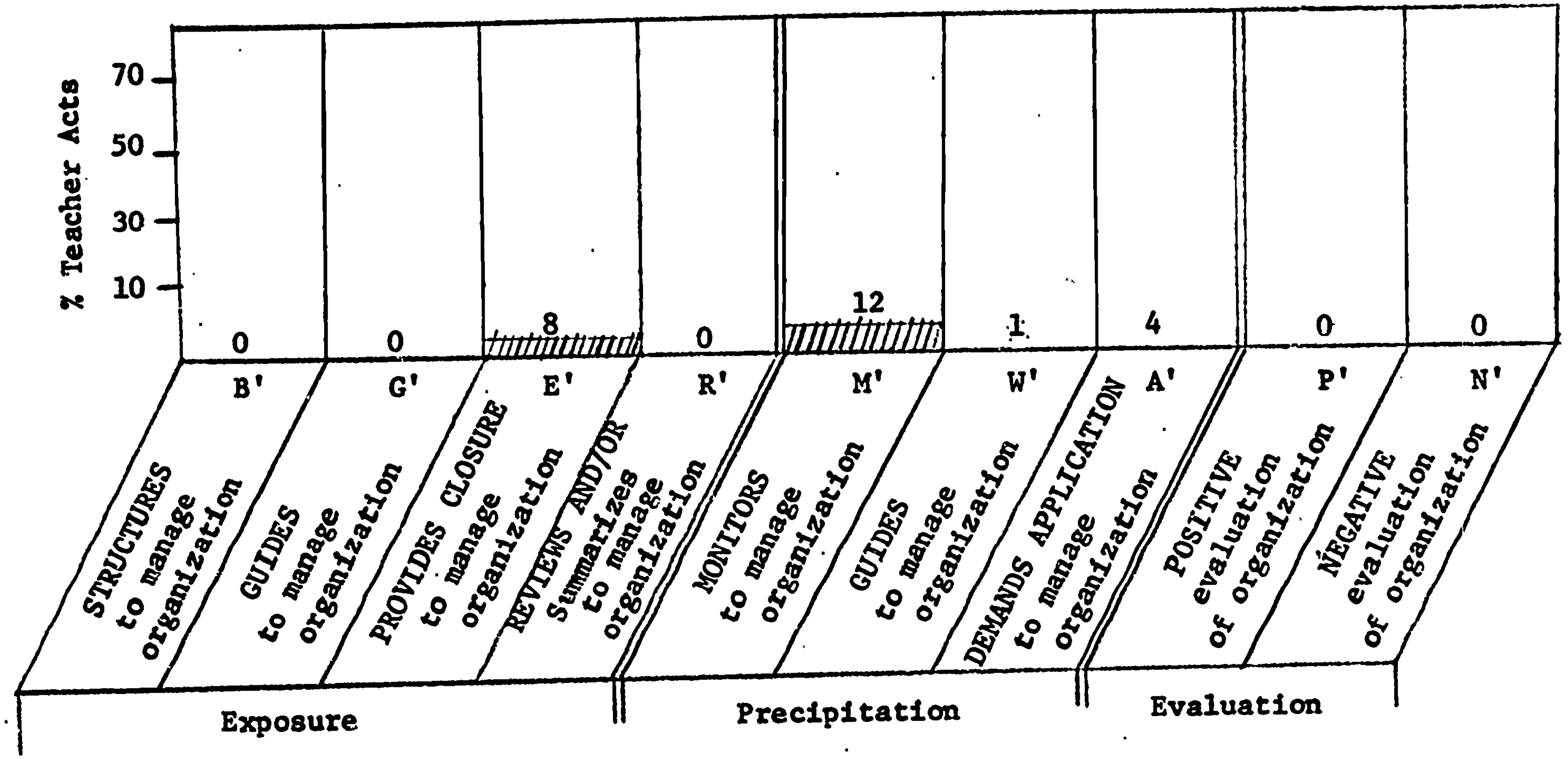


Figure 7. A graphic representation of the proportion of facilitory acts by the teacher which fall into one of the functions served by teaching operations.

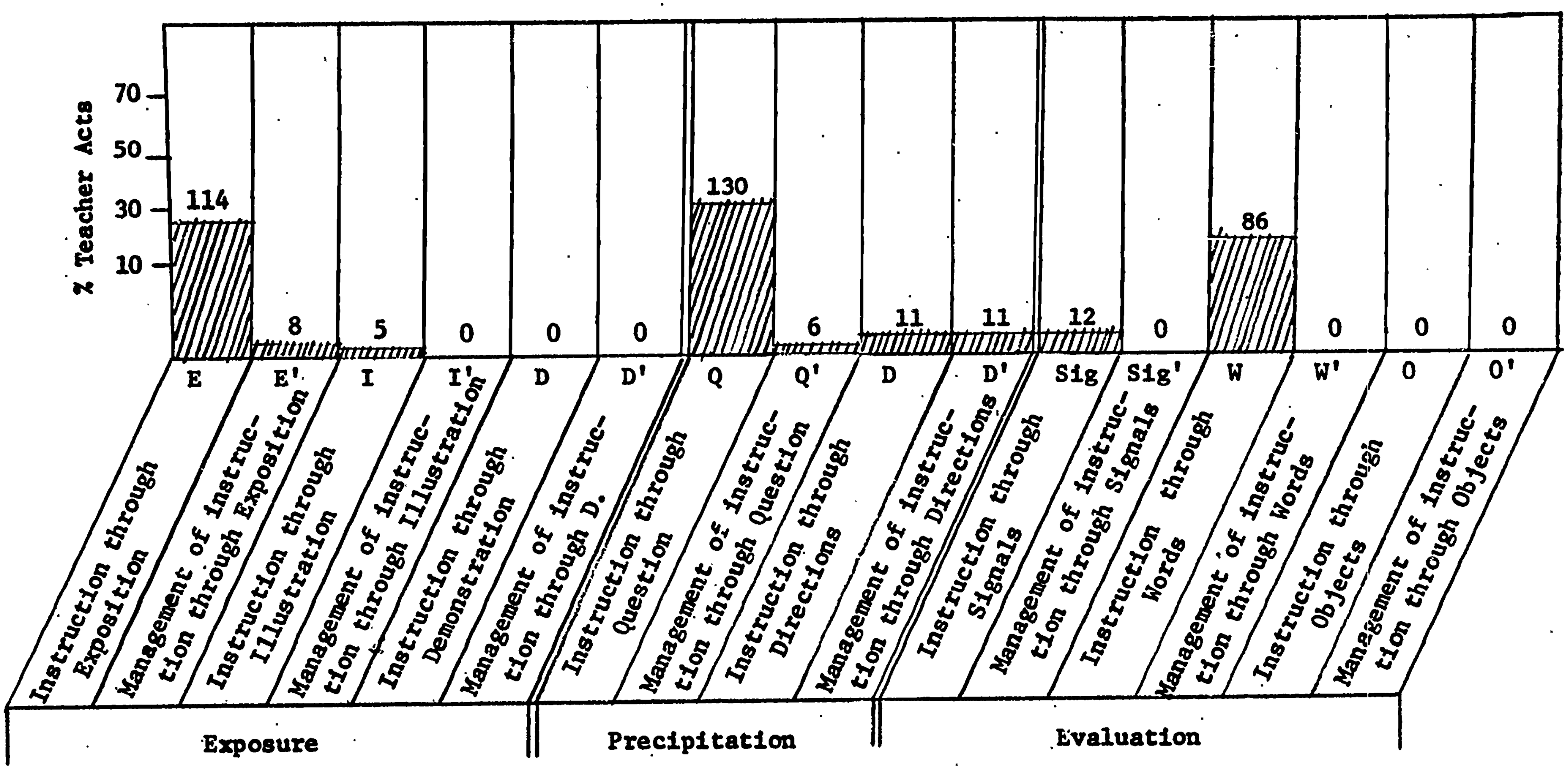


Figure 8. A graphic representation of the proportion of both developmental and facilitory acts used by a teacher classified according to the instructional tactics it represents. The tactic analysis is the thir. level of analysis in classifying TEACHING OPERATIONS. The tactic analysis identifies "how" a teacher sends a message to a learner.

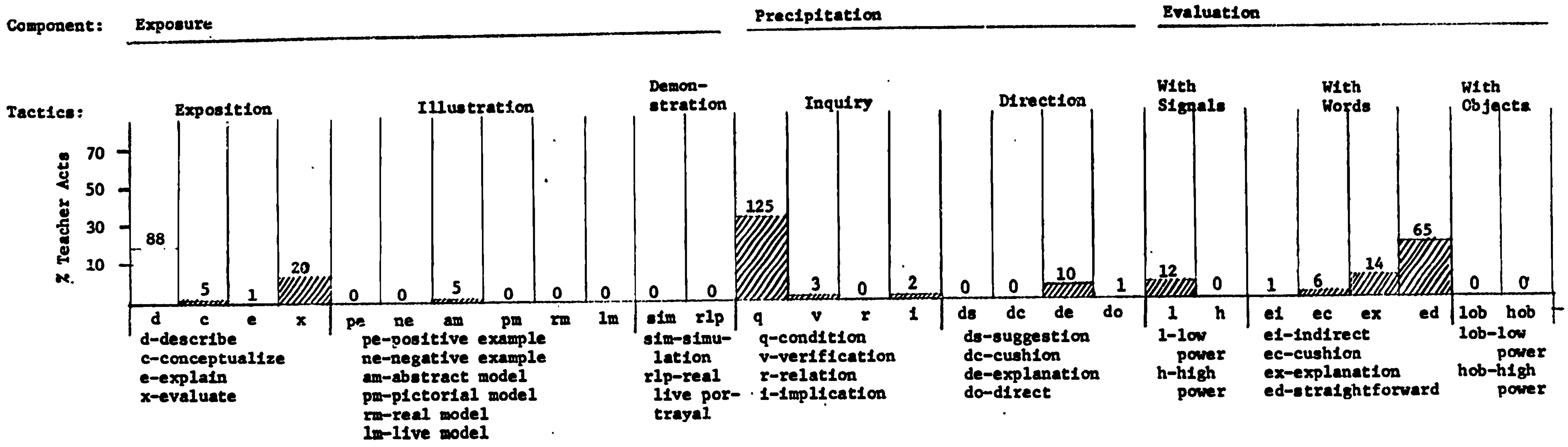


Figure 9. A graphic representation of the proportion of developmental acts used by a teacher, classified according to the instructional moves they represent. The move analysis is the fourth level of analysis in classifying TEACHING OPERATIONS. The move analysis is simply a further breakdown of the tactic analysis described in Figure 8 in that it describes in even greater detail how a teacher sends a message to a learner.

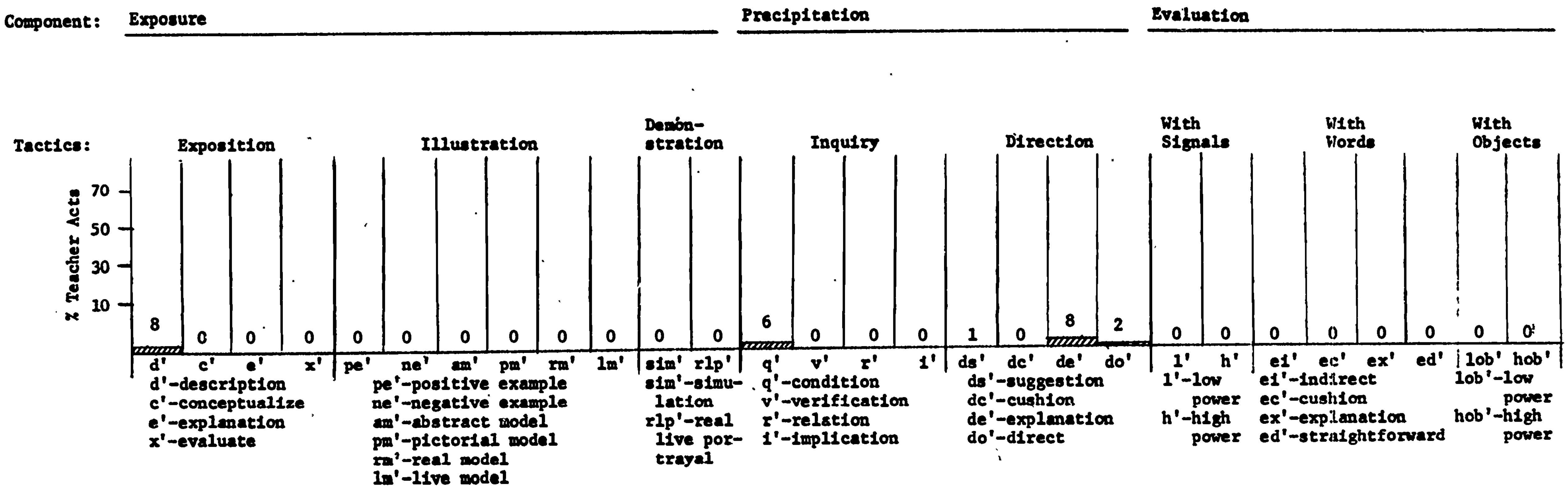
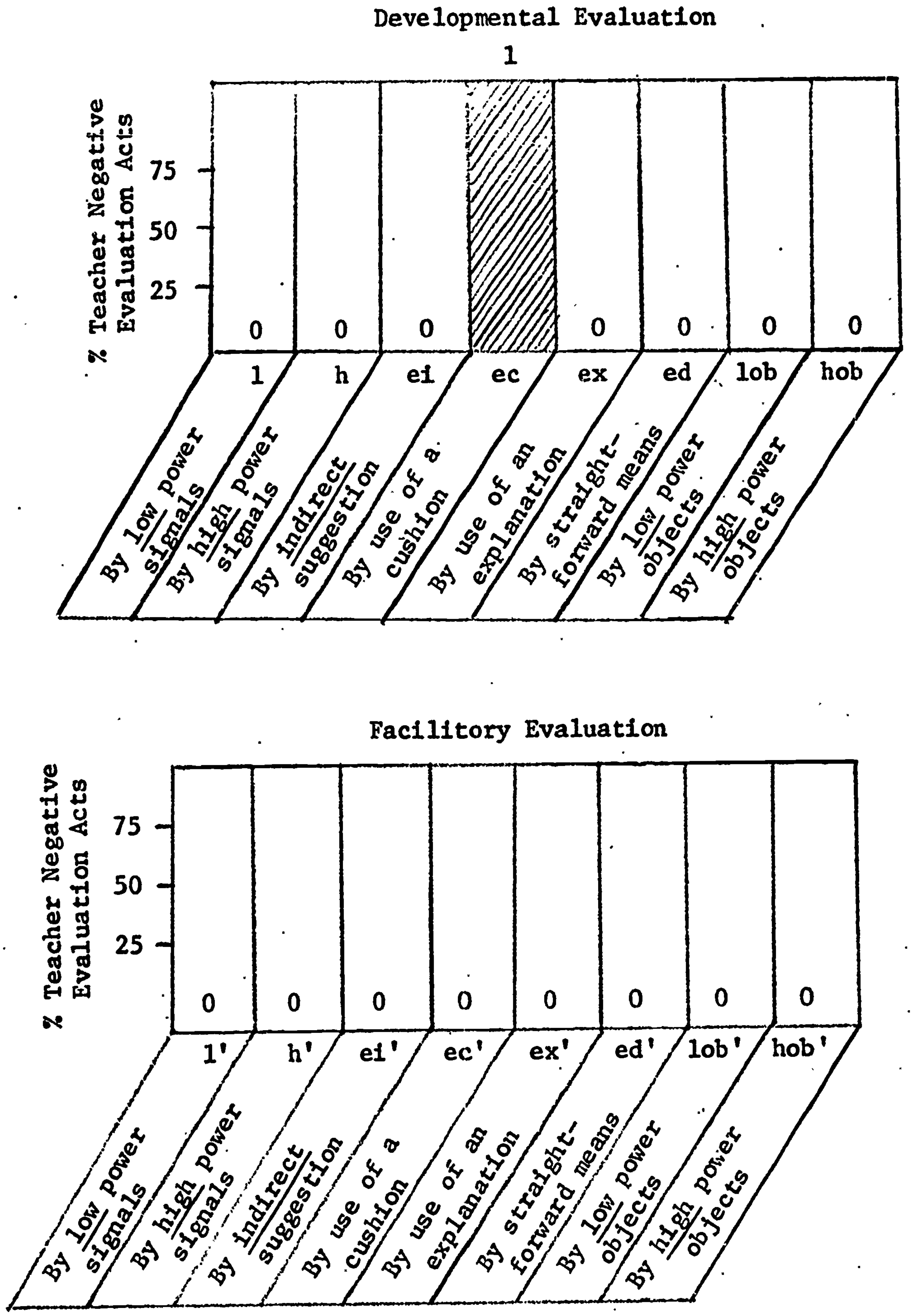


Figure 10. A graphic representation of the proportion of facilitory or management acts used by a teacher, classified according to the instructional moves they represent.

PROPORTION OF CENSORSHIP MOVES IN RELATION TO ALL INSTANCES OF CENSORSHIP OF LEARNER BEHAVIOR



Developmental Evaluation

Facilitory Evaluation

Figure 11. A graphic representation of the various censorship moves used by a teacher. The Developmental Evaluation graph represents the proportion of negative evaluative moves used to evaluate a learner's academic performance; the Facilitory Evaluation graph represents the proportion of negative evaluative moves used to censor or discipline a learner who is either out-of-focus, i.e., a learner who is not in the same FOCUS as the teacher, or in-focus, but behaving inappropriately.

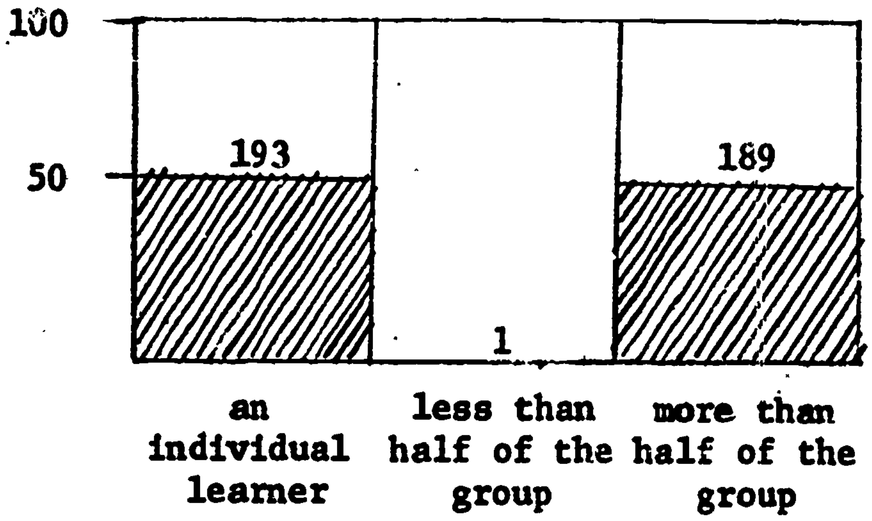
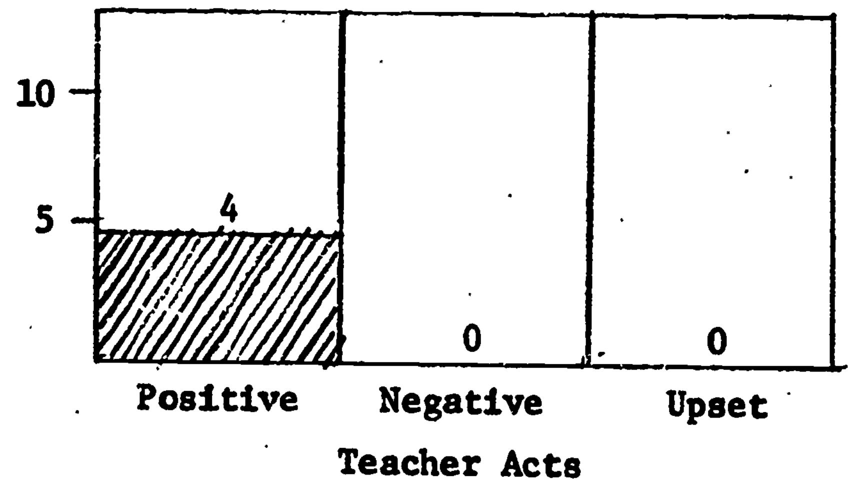


Figure 12. A graphic representation of the proportion of all teacher acts by recipient (the target audience the teacher sends the message to).



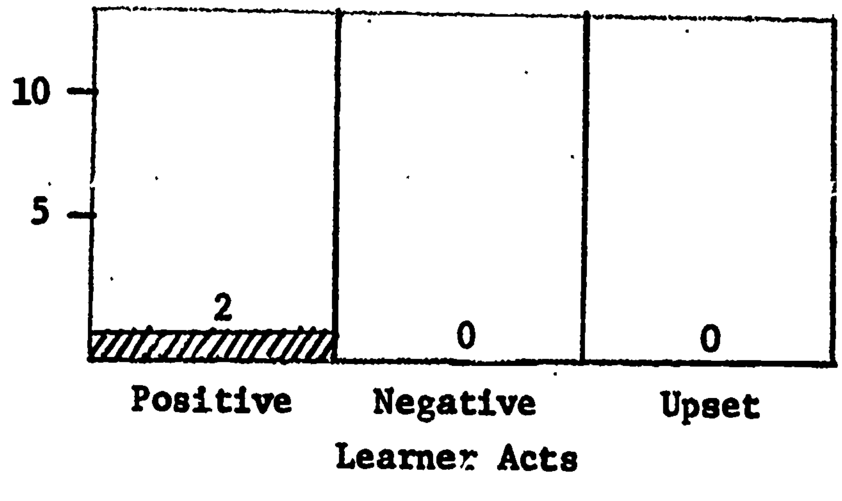


Figure 13. A graphic representation of the instances of affect in the classroom for both teacher and learners.