DOCUMENT RESUME

ED 022 558 PS 001 251

By-Garfunkel, Frank

HEAD START EVALUATION AND RESEARCH CENTER, BOSTON UNIVERSITY. REPORT A-II, OBSERVATION OF TEACHERS AND TEACHING: STRATEGIES AND APPLICATIONS.

Boston Univ., Mass.
Spons Agency-Office of Economic Opportunity, Washington, D.C.
Pub Date 67
Note-32p.
EDRS Price MF-$0.25 HC-$1.36
Descriptors-BEHAVIOR RATING SCALES, CLASSROOM RESEARCH, *EVALUATION, MEASUREMENT TECHNIQUES, *OBSERVATION, PRESCHOOL TEACHERS, *TEACHER BEHAVIOR, TEACHER CHARACTERISTICS, TEACHER EVALUATION, *TEACHING STYLES, TEST RELIABILITY
Identifiers-*Head Start

There are reasons why teaching behavior should be assessed, including (1) upgrading teacher education, (2) gaining insights into the learning of both teachers and children, and (3) studying social interactions. Two means of assessing teacher ability are quantification of teacher behavior by the use of rating scales, behavioral categories, etc., and participant observation (PO). The first, assessment by instrument, confounds the effects of too many interacting variables for the instrument to reliably represent the effects of teacher behavior. In the PO method, very well qualified and trained people are the assessing instrument. Observer judgment and observer influence upon the classroom situation are present, but if the observer is well qualified and well trained, as he must be for the success of the method, the data obtained should be more reliable and more relevant. Filming the classroom situation can also be used and adds much to the assessment process. The PO approach was tested on selected Head Start and elementary school classes. The data analysis from this testing is incomplete. It has been found, however, from a combined PO and filming of suburban and inner-city (Hartford, Connecticut) elementary classes, that suburban classes are uniformly superior to inner-city classes. (WD)

U. S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

HEADSTART EVALUATION AND RESEARCH CENTER

Observation of Teachers and Teaching: Strategies and Applications[1,2]

Frank Garfunkel

Boston University

ABSTRACT

The rationale for participant observation call for a greater reliance on experience and training of observers and on systematic procedures for sample selection and inter-class comparisons, than on the development of a system for directly and reliably recording categories or signs of behavioral fragments. Variations in teaching and in observation must be analyzed as interdependent sources which both contribute meaningful descriptions of differences between classes. Recording samples of observed behaviors is essential for training and analysis.

Applications using teams of observers in Head Start and inner city and suburban elementary school classes are described and discussed with reference to methodology and data reduction. Films were made of a stratified sample of classes in order to anchor observational reports and ratings and for the purpose of providing primary data on stylistic variation across school location and grade level.

---

1 "The research reported herein was performed pursuant to a contract with the Office of Economic Opportunity, Executive Office of the President, Washington, D.C., 20506. The opinions expressed herein are those of the author and should not be construed as representing the opinions or policy of any agency of the United States Government."

2 Films for this project were made under the resourceful direction of Professor Alvin Fiering. Mr. Charles Kokaska performed extraordinarily in developing positive relationships with teachers and in supervising field observers. Miss Janet Hudson has indexed films and organized data with consummate skill.

ED025558 PS002251

OBSERVATION OF TEACHERS AND TEACHING: STRATEGIES AND APPLICATIONS[1]

Frank Garfunkel

Boston University

INTRODUCTION

All too often educational studies employ a single recording technique to abstract teacher behavior into data. The monolith is this singular strategy rather than the claims and procedures of any one school of observational thought. Such a criticism is not confined to educational research, but to any studies that focus on complex human behaviors for which there is no optimal methodology that is accepted by professional consensus as being the epitome of validity. Although a particular methodological approach - participant observation (Bruyn, 1966) - will be described, the discussion of perspective is crucial to its elaboration. The vehicle of inference for participant observation is "observer" with experience, training, and theory rather than rating scale, checklist or behavioral protocal. In order to comprehend the validity of any of these vehicles it is necessary to explore their potential diverse contributions and to carefully describe defects in instrumentation, methodology and substance.

Participant observation is not cast as the only or preferred approach, but rather as a necessary component of research activity that aims at inferring useful data from teacher behaviors. The fact that such a strategy does not result in easily reportable and grossly comparable data should not be a deterrrent to its use if there is reason to believe that the behavior being studied is so diverse and complex that descriptive problems are inherent because of this diversity and complexity. Social sciences (and other sciences, as well) always run the risk of reporting that which is easy to describe rather than that which is important to the phenomena being studied.

RATIONALE

Strategies for obtaining data on teacher variation cover a wide range of procedures. Quantification is variously based on rating scales, behavioral categories, checklists, interaction analyses and projective inferences. Reliability is more a question of definition of behavioral units than of their relevance to teacher effectiveness. The substance of the behavior that is designated by the observational model is a reflection of either the instrument maker's or the observer's bias. Whichever

---

1 "The research reported herein was performed pursuant to a contract with the Office of Economic Opportunity, Executive Office of the President, Washington, D.C., 20506. The opinions expressed herein are those of the author and should not be construed as representing the opinions or policy of any agency of the United States Government."

is the case, there is always a presumption about educational goals and effective implementation. This is just as true of rating scales as of direct measurement which must make a prior decision about what is to be observed. It is not clear that any extant system is based on a theory which would systematically direct us to study particular behavioral categories.

When explicit attempts are made to empirically judge effectiveness by observing changes in children during the time they are with a particular teacher, and, furthermore, to select units of teacher behavior because of their relation to change, there are snarls because teacher effects are engulfed by developmental and social class effects and also, and perhaps more significantly, the behaviors that are most directly affected by teachers are not easily defined or measured. Achievement tests give an abstraction of intellectual behavior which may very well be invariant to teacher effects, especially when compared to intelligence and social class variance. This is not to imply that there are no teacher effects, but only that given the instruments and variables conventionally used, for practical purposes, they are not measurable, at least with the samples of teachers and children that have been used in teacher effectiveness research. This is an important "at least" for, as has been pointed out in psychotherapy research, demonstrated effectiveness of a particular therapist or procedure is very much a function of the diagnosis and severity of the patient. It is possible and probably that teacher effectiveness studies must take into careful consideration the age, sex, and educational-intellectual status of students. The teacher variable will probably prove to be more demonstrably effective for disadvantaged, disturbed, retarded and generally disabled children than for normal children because the variability in criteria is, to a large degree, accounted for by independent variables that are constructively and methodologically highly correlated. There is a confounding between the research problem - are teachers differentially effective? - and the measurement problem, that is largely unresolved.

While admitting that the ultimate criteria of teacher effectiveness are changes in children, it does not necessarily follow that the important teacher variable (or variables) should be derived by regressing changes (in children) against a myriad of input variables (teacher behaviors). For this to be the recommended procedure it would have to be established that the criteria are desirable and that they are meaningfully linked to teacher behaviors, neither of which is definitively so. Research on teaching is faced with a forbidding gap between teaching and learning which is partly a function of the autonomy of teachers and partly of the nature and limitations of teaching and measurement technology.

Failure to develop a predictive system for determining effectiveness has been accompanied by (and partly by default led to) the development of authoritative systems whereby one or more professionals describe what makes an effective teacher. Items, scales or categories are abstracted so that they can be used by a more or less skilled observer, to obtain data on the purported effectiveness of a sample of teachers. Behavioral units can be quite global, encompassing such broad areas as permissiveness, warmth, creativity or control, or they can be extremely specific relatively nonjudgemental, such as recording the number of times or amount of time that particular behaviors and interactions take place. Global assessment depends on trained and experienced observers while specific assessment depends on trained but not necessarily experienced observers (experience referring to teaching and training referring to observer training).

The construct validity of any more or less global or specfic system will depend on not only the substance of categories or items, but on other disiderata as well. In fact, substance might very well be of least significiance in light of situational and procedural varibilities that are often erroneously assumed to be relatively constant. Given the fact that teachers vary, it does not necessarily follow that procedures are directly comparable, operational goals are the same, samples of children in different classes require the same approach, curricular and time of day variations are insignificant or cultural forces or particular schools are not predisposing. When the burden is on the instrument (rather than the observer) it is difficult or impossible to correct for confounding that is implicit in each of these sources of variation. Given instruments will only be effective to the extent that these intervening variables are not only controlled for (presumably by randomization or manipulation) but are measured and, it follows, whose distributions are adquately represented in the given sample of classes. This suggests that either studies of teaching should concentrate on intensive surveys of relatively homogenious clusters of classes that differ on few but potent dimensions, or that large scale studies include manipulation of curricular, sampling of children, in-service training and supervision. This is to say that there is too much noise in the system for any single instrument to validly assess teacher effectiveness. This is just as true if the instrument is based on a construct as it is if it has been empirically derived.

Another rather imposing source of variation is the observer both-the procedures by which he is trained and those that he uses in the course of his observations. It is not only that different people see different things, but that the conditions of training, visiting classes, feedback, and articulation cannot be assumed to be constant. The use of a single instrument will not insure comparable data unless either the observational process is continuously standardized, the instrument has built in features which suppress observer and observing contamination, additional data is collected to provide for necessary nominal distinctions, or the variability in phenomena being observed dominated observer variability in a direction consonant with the purpose of the data gathering process.

It follows that no single strategy is inherently superior to another one but that there are situatipnal, temporal, economic, and personnel considerations which will suggest that one approach will be more valid than another. The reduction of teaching behavior is desirable because inference is based on more clearly understood judgements. However, reduction can lead to spurious and often misleading data, if it is not accompanied by compatible reduction of other relevant behaviors of teachers, children, and schools. Furthermore, the sin quo non of reduction is that the transformation be reversible. If reduction leads to a collection of irreversible bits that cannot be associated with the child's and teacher's other (and more global) behaviors, then studies of teaching will leave the domain of education and enter some other (possibly meaningful) domain. There are obviously impelling reasons why teaching should be validly assessed, not the least of which is upgrading teacher education, gaining insights into learning of both teachers and children and studying social interactions. If reductionism leads away from these by so abstracting and fragmenting behavior then it is likely that it will contribute much more to behavioral analysis than to change.

The greater the reduction to highly reliable bits of teacher behavior, the more likely it is that accurate predictions will be made of correspondingly reduced to bits of child behavior. Therefore, if the research goal is to get such correspondance, disregarding its relevance for teaching and learning, then maximal reduction is to be desired. But the reduction process, in general, ignores relevance and only accidentally

provides indices for units of behavior that are clinically meaningful. Human behavior has not been structured (theoretically) as an accumulation of behavioral bits that go together in an orderly and linear model. It is not at all clear that these bits have any useful meaning by themselves. It is a pragmatic question that can be dealt with only in terms of specified applications which become the guage of usefulness. The research decision to concentrate on any given units is germane not only to methodological considerations--how is the unit best measured?--but to the theoretical connection between teacher and learner. This connection can be conceptualized as being mapped by any level of abstraction or generality. The crucial question arises when clinical requirements demand reversibility--that results under any system of inquiry be useful as feedback in order to affect behavior other than that which is under a microscope. There is just as much need for transfer from datum to person as there is from skill to ability. Without this transfer both systems would be sterile.

Transfer is implicit in a well ordered and predictable system where reversibility (from behavior to abstraction to behavior) is generated from an object (intra) and across objects (inter). An individual's within variability over abilities is reflective of sampling variation across individuals and time and vice versa. The Stanford-Binet IQ is reversible (for middle class children) not because we can go directly back to the individual from the IQ, but because we can go the sample and then, in a meaningful way, back to the individual. "Meaningful way" refers to the well ordered system whereby a probablistic statement can be made about the individual's future academic behavior with regards to the group. Without this characteristic test scores or observational data become one way streets that make no useful connections.

Classroom observation is up against the reversibility dilemma no matter how abstract or reliable are the protocols. When data are obtained they may fit into a regression analysis but they cannot be transformed back to the class either directly or indirectly because of the lack of order in the system, either horizontally or vertically. Because of this,films (or Kineoscopic tapes) are needed to provide a mechanical vehicle for reversibility in the absence of a theoretical or empirical vehicle. Admittedly this only provides for the reversibility; it is not established. But at least the possibility exists. At the same time the vehicle for transfer is present--various techniques can be applied to the same sample of classrooms. Variability of multiple dimensions and strategies can be put to the crude, but immediate test of viewer (film) variability. Direct comparisons can be made between direct recordings of behavioral bits, ratings of qualities, and authoratative judgements. And, most significantly, teachers can be confronted simultaneously with data and behavior. For the present, films would appear to be necessary for the development of any form of observational analysis--without films even carefully obtained data will be lost to a specific, non-transferable and irreversible "black box" process.

The fact that the introduction of the photographer or the observer transforms the situation is not without theoretical interest. If non-reactive procedures can be used in educational studies, as was done by Sexton (1961) and as is recommended by Webb, et al (1966), they are to be desired unless the reactive effects are theoretically important in the reconstruction of phenomena being observed. There is reason to

believe that the principle characteristic of teaching is that it is not observed and that feedback is not existent and, in fact, impossible. Education is essentially a nonreactive system which is unaffected by contemporary social movements, recent scientific advances and critical reappraisal of current practices. Authorities set up the models and pontificate but teachers and principals run the show in autonomous conclaves. This autonomy is personal rather than professional. Textbook and examination conformity is obviated by variability along indeterminate and self-defeating lines. The model of classroom observer (or photographer) is one that involves more than an invasion into the classroom for the convenience of research. It is a different and more viable model that permits (but does not insure) a continual reappraisal of curriculum and behavior. The study of unobservable teachers is a paradox without resolution. Teaching conceived as art, science, or some combination of the two is untenable unless it can be researched on the one hand, or experienced on the other. Given the present state of research technology, the falling tree in the forest does not make a sound unless there is someone (or something) to hear it.

Orchestras need listeners, recorders and critics less they exist in an incestuous vacuum. The reinforcement of teachers consists of a bundle of meretricious acts and words which contribute more to a religion than a profession and more to a mystical epistomology than to a vital language that has some relation to behavior. Therefore, the criticism that the observer changes the situation is accepted and encouraged. That the necessary research vehicle is just as essential to pedagogy is not a coincidence. The claim can be made (even if it cannot be rigorously supported) th t any social scientific techniques should have direct payoff to the individual or groups being observed and manipulated. Using film to study teaching is an example of this claim.

Disregarding the technique used to record behavior, observational studies are usually confronted by comparisons of teaching that depend on values rather than behavior. If comparisons are to be made between teachers who lecture and those who lead discussions in varying subject fields, any system of measurement will break down unless it is either assumed that one approach is inherently better than the other (values) or that the different behaviors are irrelevant to the measurement of effeciveness which is to assume that goals transcent methodology. There are several ways around this dilemma. The curriculum and/or methodology can be stipulated (Belleck, et al 1966) and teaching can be thusly compared. Unless teachers have opportunities for participation in several manipulations there will be teacher-method confounding. Manipulation can be contrived (with or without teacher involvement) or they can be unobstrusive (and thus really not manipulations) by selecting sequences of comparable behaviors that already exist. In either case and disregarding the observational and recording technique, there is some control so that "everything being equal" is not a completely empty phrase.

If manipulations of the first or second kind are impossible to accomplish, adjustments must be made either by restricting the field of study or by using an "instrument" that allows for diverse methods, curricular and samples. Such an "instrument" might be a series of conditional scales which are selected by the observer depending on the curriculum and techniques being used. Comparisons could be made on those scales that were selected a sufficient number of times. The "instrument" could also be a highly trained and experienced team of observers who have necessary skills to compare somewhat dissimilar teaching situations. To assume, as is often done, that the observer who has the task of selecting and judging, will be more subjective than a series of protocols that cannot deal with the complexities of teaching variance, necessarily involves the tautology that such an observer is definitively subjective, and direct be-

havioral recording and rating scales are definitively objective. This fallacy is an inheritance of the so called "objective" test which is presumed to be objective because of its format, not because of its item selection, mode of inquiry or reactive effects. Admittedly, the scoring process is less subject to the biases of the scorer and the paper and pencil standardization conditions of test administration are relatively constant, but this does not provide sufficient conditions for objectivity. Reliability is an aspect of what might be referred to as internal objectivity[2] but it is not necessarily primary. It is necessary to consider the effect of the instrument on not only the subject but the educational process, the selection of items, the mode of item presentation and the problems inherent in the transformation of behavior to data. The high reliability of "objective" tests is not without a price in external subjectivity. The assumption that reliability is generic to validity has already been challenged with regard to "objective" testing and it can be similarly challenged with regard to "objective" recording of teaching behavior.

The argument is the same. The selection of items and modes of presentation involves gross subjectivity even though recording and scoring processes (which can be one and the same) are highly reliable. This is not to say that essay tests and the use of the observer-as-instrument necessarily insure external objectivity but only that they provide an alternative strategy which can more directly get at higher level processes. Thinking, reasoning, problem solving and creativity may be vague but they come closer to the expressed goals of education than memorizing, recalling, and educated guessing. Similarly, the assessment of humane, creative, elaborative, insightful, and intelligent teaching is more directly to the point than counting the amount and number of times teachers and students ask questions, make statements, make demands, and are silent. This is not to preclude that specifically defined behaviors can be important indicators of generalized functions but only to gain perspective about their limitations and the value of alternative "subjective" strategies to approach a more profound objectivity than is to be had by using "objective" methods exclusively.

The question of reaction is not a trivial methodological issue that can be relegated to vagaries of research. The teacher who is "counted" and the observer who is counting are part of the system and will respond in some way to this procedure as opposed to an alternative one. The reductionism involved in "counting" reduces not only behavior, but the work and status of the observer and, therefore, of observational process. This is not a polemic for eliminating "counting" but rather an argument for questioning any reactive procedure, not because it is reactive, but because of the quality and force of the reaction it might evoke.

GENERAL STATEMENT OF PROCEDURES

We address ourselves specifically to the problem of evaluating and describing the potential effectiveness of teaching in a diverse sample of classrooms and schools (or centers). Amount of observation will depend on sample variability and O sophistication. In order to obtain approximations of these parameters the design calls for

---

2 This followed Campbell and Stanley's (1963) distinction between internal and external validity.

multiple O's making multiple observations of classes over an extended period of time. O's will have had teaching experience and will participate in seminars prior to and throughout the PO. Training will consist of a variety of experiences aimed at facilitating inter-O communication, becoming familiar with a behavioral model and developing observation sensitivity. Seminars and workshops prior to PO will be used to screen out unsuitable candidates. O's will participate in an observational seminar where they collectivly observe groups of children in classes and discuss at length, teaching and learning as they view it. O's will observe each other teaching children and discuss varieties of approaches and values.

Films will be utilized in the observational seminar in order to allow for review of discussed behaviors at any time. These films should show diverse teachers doing similar tasks and similar teachers, or a given teacher, functioning in varying ways. It is desirable for O's to view different teachers with the same group of children.

O's will keep careful logs of observed behaviors which will provide detailed accounts of teacher, child and interactional behaviors. Analytical reports will be written, utilizing the log as sources of evidence. Finally, O's will write interpretive summaries of teachers and classes, describing their estimation of effectiveness and indicating teaching characterisitics that are critical for their assessment. Procedures for writing these reports are set forth in greater detail in the appendix to this report.

Scales representing important and adequately variable dimensions of teaching and child behavior will be constructed in such a way as to relate the observed behavior to the behavioral theory. O's will Q-sort classes on each of these scales--rating all classes on one scale at a time thus minimizing associational biases. O's will underline and label logged behavioral recording according to a notation that related scales to specific recorded behaviors. Scaled judgements can then be supported by molar sequences of observed and recorded behaviors.

MODEL

Although participant observation (PO) varies as to the specific procedures used, it is always based on the principle that although the observer (O) will adapt preconceived structural outlines and dimensional scales on the course of his summary, he is the instrument for inferring data, rather than any outlines or scales. There must be enough intensity and duration in the involvement with the phenomena being studied for its unique structure and process to be indentifiable. The amount of contact is a function of the kind and degree of distinctions between individuals and agencies that are required. Once the target system is defined O has the responsibility of determining a traffic pattern for himself which will lead to an understanding of relationships and direction. Hypotheses are constructed by relating a presumed general theory of behavior to the behaviors of the system. PO methodology is independent of the theory or of the working hypotheses--but some articulated theory is necessary.

O is presumed to be experienced and trained although specifications for both depend on task requirements. Training can be presumed from the previous experience of O or it can take place prior to and during PO. Reliability will depend on the perspective

and sensitivity of O and multiple O's can be used to provide anchoring if diverse situations are to be observed. O will observe and become involved (interviews, utilization of unobtrusive data, manipulation) to an extent necessary to test hypotheses about prediced outcomes and structural relationships. Guidelines for participation must be drawn up, prior to observation, with the cooperation of individuals involved.

Biases of O must be continuously dealt with but this will depend on whether they are a legitimate source of error. Where O bias will produce variation equal to or greater than phenomenolpgical variation , it is necessary to articulate and hypothesize about bias x behavior interaction in a manner suggested in general terms by Mydral (1953). Where bias is of minial important (as in many cultural anthropological studies) it need be only articulated.

Just as in any data gathering process, inferences are only as strong as the instruments that are used. PO depends on high quality O's who can demonstrate their perspicacity by being able to predict interactions and circumstances and to relate observed behavior to given theoretical models. Proof of quality can either be left to the readers of final reports or it can be currently brought into relief by using multiple O's with parallel systems. The test of effectiveness or precision is clearly not a reliability coefficient or an "F" ratio. Any such statistical test works smoothly once the data is obtained and disregarding the validity of the data. PO emphasizes letting meaning speak for itself in much the same way the Skinnerians proclaim that data should be directly recorded and then speak for itself.

The assumption of PO is that there are O's and methodologies which can be used to obtain data that reveals more about observed processes than about O's. Methodologies can be designed to efficiently utilize O's with given degrees of competing biases and with specified goals with reference to designated behavioral systems. This is to say that design will have to be adapted for known variations in O's, goals and systems.

PO is not clearly defined methodology that is uniformly used in the social sciences. The practice of having an O look closely at a segment of interpersonal (or individual) behavior is simpleminded and elementary. Where more clearly defined procedures are appropriate they should certainly be used. The designations of adequate O's is difficult and perhaps, often impossible. It might appear that PO is a regression to pre-scientific methodology, where uncontrolled judgements are combined with unknown weights. But it is even less scientific to use "powerful" instruments to perform tasks for which they are unsuited. The decision to use PO is made in light of the complexities of teaching, the difficulties of obtaining comparable samples of behavior, the problems of irreversibility, the tenuousness of child behavioral criteria and the obscurity and ineffectiveness and inappropriateness of personality measurement for obtaining adequate measurements of teacher characteristics. This could lead to the abandonment of such research or, as in the case of PO, to the adaption of relatively crude processes which can, albeit subjectively, deal with those obstacles. Developments in audio-visual technology will make it possible to give more substance to the inferences of O's and to provide reasonably direct documentation of classroom processes that can be exposed to more varified procedures.

APPLICATIONS IN HEAD START AND ELEMENTARY SCHOOL CLASSES

Applications of modified participant observation approaches were made on selected Head Start and Elementary School classes in connection with two projects, which were taking place concurrently. The first involved twenty Head Start classes which were being evaluated by the Boston University Head Start Evaluation and Research Center as a part of its participation in the National Evaluation Program. The second was with Project Concern, an experimental study of the effects of suburban education on inner-city children in and around Hartford, Connecticut. Since the data on both of these projects, with regards to the tested and observed performances of individual children, has not been made available, this report is necessarily incomplete. Procedures for observing classes and obtaining data will be described in some detail, and preliminary descriptive statements will be made with regards to dimensionality of scales that were used in each investigation and agreement between raters on a variety of scale ratings. In addition, for the Project Concern application, the division of classes into inner-city defacto segregated and suburban unite with one, two or three bussed negro children in them permits a straight forward comparison over location of classes.

Although the general principles behind participant observation, as developed by Bruyn (1966) were followed in the development and carrying out of procedures, the sustained and intensive contact of observers with classrooms and schools was not followed partly by choice, because of the kinds of variation that were of most interest, and partly by necessity. Future studies will provide for considerably more contact between observers and the institutions they are observing in order to realize the depth which is only being approximated by procedures to be reported herein.

The aims of these studies were twofold: 1) to study the relationship between selected characteristics of teacher style and changes in mental abilities, academic achievement, personal-social development and creativity of children in selected classrooms; 2) to describe, through cross-sectional procedures, teaching situations which Head Start children are exposed to and those to which they will most probably be exposed to if they attend inner-city or suburban elementary classrooms.

PROCEDURES

Both applications called for the recruiting and training of observers who had extensive experience both as teachers and as observers of preschool and elementary school classes. Initial training sessions involved observation of classes and discussion of an all-inclusive categorical model of classroom procedures (Appendix C). This model was not for the purpose of providing a checklist or of focusing observers' attention on particular variables so much as it was for directing their attention to all possible contingencies and teaching situations. The model included listings, under the general heading instruction, of materials, lessons, motivation, evaluation, and achievement. A second section under the general heading of controls included form, quantity, tone, consistency and student pressure. Facilities listed characteristics, and implications for teaching. Student interaction included opportunity characteristics. A last category, teacher-student interaction included humor, address, feelings, reinforcement. This model was meant to be a vehicle which would serve to provoke discussion and generate questions about varieties of teaching experiences. In addition, an exhaustive list of variables associated with teachers, students and curriculum was

constructed, through the deliberations of observers, in order to sensitize them to differences between independent, intervening and dependent variables (Appendix F). It is critical to note that the models developed from observational seminars and were, therefore, the produce of the efforts of observers. They were not handed listings of categories and variables which had been developed externally and which would have been, therefore, imposed upon them.

Observers were asked to keep detailed notes on their observations without regard to a particular model, but with specific regard to what they considered to be the most important characteristics of the classrooms they were observing. These notes were to be transformed into process reports which were to be concluded by analytical reports and summary interpretations (Appendix D).

Scales were developed for both studies by observers after carefully and deductively describing contrasting characteristics of teaching situations which observers judged as being relatively unique. (See appendices A&E). The scales, are, therefore, a reflection of differences seen by observers, rather than the basis for making distinctions. This meant that this approach to studying teaching involved a concomitant study of observer variation, and that these two separate focuses were mutually interdependent.

The burden of responsibility was clearly on observers rather than scales and it called for an inferential process which would be only as defensible as the perceptiveness and intelligence of the observers permitted. This process structures a systematic approach to dealing with subjective impressions of observers who are required to defend these impressions in the face of careful scrutiny by other observers and by senior members of the project staff. The process assumes that each observer has enough experience and insight to be able to produce salient reports and interpretations of teaching variation. Resulting inference must attend to both sources of variation--teaching and observing--in order to adequately describe stylistic variation within stylistic categories.

In order to provide a superstructure for teaching and observing variations, films of selected classes were developed. In covering a wide range of activities, these films have and will continue to provide referent behaviors for the reports and ratings of observers. Extensive use of these films has been and is continuing to be made in order to clarify reductions of behavior that were made by observers.

OBSERVATION OF HEAD START CLASSES

Of the twenty sample classes used in the National Evaluation program, nineteen were observed sufficiently by two or more observers to produce reports and ratings on a series of scales which were constructed by observers during the course of their observations.

Eight scales were used in rating nineteen teachers by six observers, with each

teacher being rated by two, three, or four separate observers.

The scales were as follows:

1. Attitude towards teaching situation.
2. Teachers differentiation of children and activities.
3. Predominant emphasis of curriculum.
4. Purposefulness of classroom behavior.
5. Control of materials and interactions.
6. Communication-responsiveness.
7. Work-play continuum.
8. Overall rating.

The detailed statements about each of these scales were given to each observer and can be found in Appendix A.

Rater agreement on the ten scales varied between 80% and 90% and on the overall rating the agreement was 92%. Interscale correlations varied between .60 and .90. Variation between classes appear to be sufficient to allow for maximal rater agreement as well as the probable inflation of scale inter-correlation.

Observers were instructed to sort all teachers on each scale, rather than rating each teacher on all scales, in order to minimize halo effects.

Since four of the six observers had training and experience in early childhood education and, consequently held a point of view which valued highly differentiated programs with a considerable amount of freedom for individual children, resulting ratings are necessarily a reflection of this point of view and are, therefore limited in their generality. Observational teams that participate in such a strategy should represent a wide spectrum of points of view with at least two observers representing each major variation. Similarly, it is essential to obtain samples of classes where competence and style are relatively independent so that their respective sources of variance can be partialed out.

PROJECT CONCERN: Comparisons of inner-city and suburban classes.

Project Concern is a large scale interventional project which provides for educational placement and supportive services for 250 inner-city children. The inner-city children are all residents of Hartford, Connecticut, and the experimental intervention consists of placement in surrounding middle class suburban schools. A randomly selected control group of 250 children is being studied concurrently in order to test hypotheses regarding the differential effects of inner-city and suburban schools on children. A summary of the theoretical framework and the experimental design of Project Concern can be found in Appendix B.

The Boston University Head Start Evaluation and Research Center has been involved in observing and filming a random sample of classes that contain experimental and control children. Observations have also been made on a sample of Head Start classes so that educational continuity between Head Start and elementary school could be ascertained. Filming took place within a careful observational survey design so that

the validity of the filming process could be evaluated.

From thirty-nine schools involved in Project Concern, thirty-eight classes were selected for the observational and film survey. Ten of these classes were filmed over a five-month period. The extent to which filmed behaviors of particular classes represent those classes, as well as the extent to which the film classes are representative of all classes, is presently under careful consideration. Findings thus far are that independent observers can go from films to reports and from reports to films with equal facility and that ratings of films are in almost complete agreement with observer ratings made of filmed classes at other times during the year.

Both the observational and film survey included kindergarten, first, second, third, and fifth grades in both inner-city and suburban schools Inner-city classes were selected randomly (stratified on grade) from the total pool of control classes. Suburban classes were selected randomly from two communities that had greatest participation in the project and that represented more and less cooperative coomunities with regards to Project Concern.

The observational team consisted of five observers with widely different backgrounds and points of view. They were trained, respectively, in preschool education, elementary education, elementary and special education, secondary and special education, and elementary education and counseling. Each observer was randomly assigned a sample of classes in both inner city and suburban schools. They were required to make at least two extended observations, separated in time by at least one week, and, preferably, three or four separate observations. In addition, each observer was required to observe classes of two other observers at least once and, preferably, twice each.

Observers wrote process and interpretive reports and rated each class on ten scales that had been derived by the observational team from preliminary observations of the total sample of classes. A sorting technique was used so that a given rater would focus on inter-class variability over each scale, rather than within class variability on all scales.

Scale derived areas follows:

1. Involvement and interest of children
2. Purposeful behavior of class
3. Source of direction of academic activities
4. Nature of control over behavior
5. Effectiveness of behavioral controls
6. Quality of presentation of subject and materials
7. Differentiation of instruction
8. Teacher reaction to classroom situation
9. Reinforcement of behavior of children
10. Nature of reinforcement

With the exception of scales five and nine, there appears to be a general factor

which differentiated teaching in observed classes. Intercorrelations between scales ranged between .60 and .80 and the internal consistency of the scales is well documented across all observers by scale--total score correlations of .80 and .90 with the exception of the two scales mentioned. Rater agreement on individual scales, with the exception of scale 9 varied between .50 and .60 and rater agreement on the cumulative mean rating that was made by each observer on each teacher was correlated .65.

There are important differences between raters as is reflected by their respective interscale correlation matrices. For two of the raters, the interscale correlations were generally between .40 and .60, while two of the other observers had interscale correlations between .75 and .85. Subsequent data analyses which are aimed at establishing differential effects within suburban and inner-city classes will treat observer score matrices separately in order to access the validity of different observational points of view with respect to predicting change in diverse educational settings.

Data obtained from scales was unequivical in showing suburban classes to be uniformly superior to inner-city classes. Seventy-five percent of the suburban classes were above the median and seventy percent of the inner city classes were below the median which was highly statistically significant on "t" test.

Differences between inner-city and suburban classes were statistically significant on all scales except 5, effectiveness of control; 9, reinforcement of behavior; and 10, the nature of reinforcement.

Thus, observational ratings clearly distinguish inner-city and suburban classes on selected scales and on mean rating over all scales. However, 30% of the classes overlap, five suburban classes being below the median and six inner-city classes being above the median.

These observational data will be used in order to modify the prediction of change in inner-city and suburban classes in order to determine whether high quality (as here defined) classes in inner-city schools are associated with changes in children in high quality classes in suburban schools and, similarly, whether low quality instruction in the suburbs is associated with low quality instruction in the inner-city.

DISCUSSION

This carefully structured observational survey demonstrated the degree and kind of difference that is manifest between inner-city and suburban classes. This is backed up by a film survey of selected classes, kindergarten through five, in inner-city and suburban schools. There is a close correspondence between filmed behaviors and those that are reported in the data analysis of the scales used by observers. In both cases it is apparent that inner-city schools are characterized by relatively uninvolved children, classes with extremely restricted purposes and teachers who tend to pervasively control materials and children. This control is often expressed as coercion and threats and is accompanied by a rather pedestrian presentation of materials with relatively little differentiation of instruction. Inner-city teachers appear to enjoy their teaching less than suburban teachers. These differences are quite apparent in the films, which are presently being prepared for showings at several national conventions.

Inner City and suburban classrooms will be displayed simultaneously on two adjacent screens in order to bring these comparisons into relief. Films have been subjected to detailed analyses in order to refine scaler differences. Films of the inner city and suburban classes have been combined with films of Head Start classes in order to specifically and objectively present a cross sectional longitudinal comparison of the experiences that children have in preschool, kindergarten and through the grades. The films vividly portray the contrast between selected Head Start and selected elementary school classes.

All filmed sequences have been coded according to a curricular scalematic devised by Garfunkel (1967) which identifies activities according to curricular classification (activity, substantitive or routine), substantitive or activity category (construction, performance,play gratification, language, social science, snacks, clean up or rest), process focus (mechanistic routine, skill, perceptual, cognitive or social) and control (teacher or child dominated). Each sequence is also rated on the scales developed by observers. This allows for matching of contrasting curricular and stylistic sequence across and within location (inner-city-suburban) and grade level (Head Start and Kindergarten through Grade Five). Furthermore, it provides a basis for comparing filmed sequences on ten classes to observed, recorded and rated behaviors in 38 classes which were selected by using systematic and random sampling procedures. The validity of the films is, therefore, based both on techniques and methods of selecting classes and filming then, and analytically, by obtaining comparable data on films of a limited sample of classes and anecdotal reports and ratings on a representative sample of both inner-city and suburban classes.

Preliminary findings from these studies document wide variations across Head Start inner-city and suburban classes. The obvious next step is to follow children who have been exposed to certain styles of teaching and to compare their responses to elementary schools that offer similar and contrasting classroom environments. This can serve as a control for predicting how high and low changes on various measurement procedures will respond to continuous and discontinuous learning environments. Of particular interest will be the interactions between Head Start and elementary school stylistic variations on selected measures of achievement and social-emotional behaviors.

REFERENCES

Bellack, A.A., Kliebard, H.M., Hyman, R.T., and Smith, F.L.Jr., The Language of the Classroom, New York: Teachers College Press, 1961.

Bruyn S.T., The Human Perspective in Sociology: the Methodology of Participant Observation. Englewood Cliffs, New Jersey: Prentice Hall, 1966.

Garfunkel, F., Observational Strategies for Obtaining Data on Children and Teachers in Head Start Classes, (OSOD), Boston: Boston University Headstart Evaluation and Research Center Final report to Institute for Educational Development, 1967.

Mahan, T.W. Project Concern: An Interim Report on an Educational Exploration. Hartford, Connecticut: Project Concern, 1967.

Myrdal, G. An American Dilemma. New York: Harper, 1944.

Sexton, P.C. Education and Income. New York, 1961.

Webb, E.J., Campbell, D.T., Schwartz, R.D., and Sechrest, L. Unobtrusive Measures-Nonreactive research in the Social Sciences. Chicago: Rand McNally, 1966.

Appendix A April 1, 1967

HEADSTART EVALUATION AND RESEARCH CENTER

Boston University

Scales for Rating Participant
Observational Reports of
Headstart Classes

1. Attitude towards teaching situation

This scale is specifically aimed at a judgement of whether the teacher enjoys the teaching situation and not whether she is a good teacher or whether the observer likes her. At the high end of this scale such adjectives as happy, pleased, exhilerated, joyful, and so forth.. At the low end of the scale, unhappy, miserable, sad, pained, and so forth. The judgement revolves around what the observer sees in the behavior of the teacher and not a projection by the observer as to whether O would be happy doing the things that the teacher is doing. This, as well as other judgements, will depend upon evidence that is collected in the course of observations, and it should be possible to sight that evidence. Therefore, it is theoretically assumed that the total behavioral protocal is reducible in such a way as to provide bits of evidence to support each scaler judgement. Without such reducibility, the judgement becomes simply a "gut reaction." While admitting that the "gut reaction" is an important part of perception and judgement, the process of collecting evidence and making judgements should force the observer to look deeply into his reaction and to make essentially two judgements: the first one being whether or not he can make a rating, and the second being conditional on an affirmative response to this. The condition of being able to make the rating will always depend upon the articulation of evidence to support a given judgement.

2. Teacher's differentiatiation of children and activities

At the end of the scale we have a teacher who runs a class that has a high rating of individual instruction and who does not make demands upon groups of children to do the same things at the same time. High differentiation would involve either one of two strategies: A.) where there is a special plan for each child depending upon his abilities and attitudes and b.) where each child is allowed to go his own way and to seek out his own kind of activity and activity level. Low differentiation would be evident by a preponderance of classroom activities which involve all children. It does not follow from this that this scale will necessarily correlate with good teaching or poor teaching, but that it represents a style of teaching with respect to dealing with individual children of groups of children.

3. Predominent Emphasis of Curriculum

This is essentially a nominal scale which calls for a judgement on the part of the observer as to which of the categories suggest the principal manifest goals of the activity being observed. The extent to which these categories are ordinally related depends upon a presumed value system with regards to desired goals of

preschool teaching. The categories to be used in this scale are taking directions, cognitive, perceptual, social emotional and a fourth category, unclear, which indicates that no single emphasis can be inferred from observed activity. The judgement, of which category a given sequence of behavior belongs to, will depend upon the behavioral priority system that operates for a given class. For example, if a given lesson or period appears to be dominated by cognitive training but if the behavior of the children cause changes in plans and redefinition of the program, then cognitive would be viewed as being a secondary goal, and the kind of activities which cognitive training give way to would be the primary designation. It is essential that we observe classes closely and long enough so that we can make inferences about what the goals, in fact, are, rather than what they are said to be. Freeplay periods might be dominated by something like the learning of routines and/or language training. Perceptual training might very well be dominated by social/emotional considerations if the behavior of the children causes the teacher to shift the emphasis for individual children, from time to time. As has been stated for the other scales, it will be necessary for observers to present evidence for manifest goals and to distinguish between the nominal categories of this scale and overall judgement of effectiveness. A good deal of work will have to be done on this scale so that it presents the observer with a series of branching scales with alternative categories, but with a theoretical connection between the different branches.

4. Purposefulness of classroom behavior

An affirmative response to this scale will depend upon clear evidence of direction and continuity. One would expect to find a considerable amount of observer disagreement over this scale because this is particularly subject to whether or not the observer is in harmony with the teacher and is able to see the underlying goals of the class as it evolves. In order to rate a teacher as being purposeful and the class as being purposeful, it will be necessary to show evidence for continuity and direction; and similarly, inorder for a teacher to be rated as being not purposeful, it will be necessary to point out discontinuity and to show many apparent shifts in direction during the course of observation.

5. Control of Materials

The question here is not so much whether it is the child or the teacher but, rather, whether the child has a say in either the gross selection of activities or materials or in their use after they are selected, or whether the teacher dominates both selection and use.

6. Communication--Responsiveness

This question is directed at the class and raises the issue of whether, whatever is going on in the class, there is great responsiveness to it on the part of the children or are they largely unresponsive or indifferent and, if anything, following through on routines rather than being responsive to activities and to the teacher. Responsiveness is indicated by a large amount of verbal and non-verbal communication, but it does not indicate that this communication is constructive or destructive or that it is good or bad.

7. Work and Play

At the high end of the scale, work and play are undifferentiated and the teacher makes little attempt to label or construct activities as being work or play, but, rather, they tend to meld together. At the low end of the scale there is a clear distinction--certain activities are presented as play activities and others are presented as work activities.

8. This scale is for a total "gut reaction" to the teacher, class, and children; and it asks the observer to indicate, without any great demand for evidence, that he thinks a given teacher is more or less effective.

All of these scales are intended to get at ordinal distinctions between a specified sample of teachers that a given observer has been assigned. All judgements are necessarily comparative, and they will depend upon what observer has seen as a part of the observational task. It is the job of the designer of the sample to make sure that each observer has a fair distribution of teacher variability in his sample and, furthermore, that this variability is not highly skewed. This means that the assignment of a sample of classes to a given observer must be proceeded by enough observation to provide evidence for gross variability within a given sample of teacher. Samples for observers should have relatively homogeneous variance.

Appendix B
October 10, 1966

PROJECT CONCERN

T. W. Mahan, Project Director

Brief Summary of Theoretical Framework

PROJECT CONCERN, although directly related to the problem of de facto segregation, is not essentially an experiment in integration; rather, it is an experiment in educational intervention designed to counteract the limited influence of urban education on the disadvantaged. Research has described the "cumulative deficit" which the child from the low socio-economic environment tends to exhibit in his school performance--a phenomenon which is dramatically accentuated among the non-white poor--and has underlined the profound task involved in reversing the trend. A review of the literature quickly communicates the impression that the problem goes beyond special teaching techniques, enriched materials, and better programming.

PROJECT CONCERN will be evaluated by measured changes in pupil behavior. Nonetheless, it is important to outline, at least in skeletal fashion, the theoretical base from which these changes are predicted. Basically, the research stems from a conviction that changes in stimuli, environment and other input data can result in changes in response or output behavior. However, it also felt that cognitive patterns for copying with formal learning situations and the affective responses which accompany these patterns have been well crystallized at the time of school entrance. This results in the use of traditional response paterns which, for the disadvantaged, are frequently ineffective for school goals. To counteract this established tendency it seems best to present the subject with an intense and pervasive experience in a radically different environment so that new responses can be provoked. This is the first stage of PROJECT CONCERN--to create some dissonance within the pupil in terms of his usual perception of himself in relation to school and to take advantage of this period of flux by reinforcing positive behaviors and attitudes.

The second aspect of the intervention model is tied to the influence of peers as a basis for the development of role fulfilling behaviors. By placing a limited number of inner city youth (about 10% of the classroom population) in a suburban classroom these same youth will be constantly in contact with models of behavior more in keeping with school values. By limiting the impact of models which reinforce the current, ineffective behavior and emphasizing the impact of different, but reasonable consistent models, it is hoped that some "shaping" of the pupils' learning styles will take place in the direction of increased academic performance.

As a catalyst to prevent too much dissonance which might create a withdrawal and/or rejection reaction, significant adult figures who share much of the child's heritage but also exhibit the desired characteristics in terms of attitudes toward school and learning are provided in the supportive team. The effectiveness of this additional factor in the change process is a focus of the research design and, hopefully, evidence will be available at the termination of the project to determine the differential impact of the learning environment as separated from the impact of adult identification figures.

In essence, PROJECT CONCERN focuses around the change in perception, already to a large extent stereotyped, which can be accomplished by a confrontation with experiences highly charged with novelty but also in a context of interpersonal support. It is predicted that changes will take place and that they will take place in the direction of the models which the suburban youth present to the bussed pupils.

EXPERIMENTAL DESIGN

PROJECT CONCERN is designed to determine the relative effectiveness of a radically different educational environment as a preventive and corrective intervention in the education of urban youth from the inner city. The theoretical rationale for the position has been discussed above, but the pragmatic aspects must be mentioned briefly here. The "vacant seat" for pupil assignment has resulted in considerable variability in the placement with some classes having only one experimental S while others have four. This in turn has created a situation which results in the experimental Ss being spread across thirty-three (33) schools while control Ss are drawn from six (6) schools. Hopefully, this diversity will have a self-cancelling effect which will underline the impact of the experimental variable - the treatment procedure. In this same regard, it is also important to stress that the Experimental Ss not receiving external supportive services are all placed in one school system (6 schools) and that generalizations from their performance must be made with that fact clearly in mind.

Nonetheless the design seems adequate to examine the relative impact of four (4) methodologies on the learning attitudes and motivations of inner city youth. These methodologies, in order of their predicted effectiveness, are as follows:

1) Placement in a suburban system with supportive team assistance.
2) Placement in a suburban system without supportive team assistance.
3) Placement in an inner city school with supportive team assistance.
4) Placement in an inner city school without supportive team assistance.

Ss assigned to treatment procedures one (1) and two (2) above are considered to be Experimental Ss since they are subject to the impact of the major variable under study: placement in a radically different educational environment. Ss assigned to treatment procedures three (3) and four (4) above are classified as controls. As described above all Ss were drawn from the same population in a random fashion. Schematically, the design is as follows:

	Experimental Groups				Control Groups			
Grade	With Support		Without Support		With Support		Without Support	
	N	Schools	N	Schools	N	Schools	N	Schools
Kdg.	32	8	14	3	--	--	50	1
1	38	9	5	2	12	2	40	2
2	47	9	2	2	12	2	40	2
3	30	7	7	3	12	2	40	2
4	25	6	9	4	12	2	40	2
5	41	9	6	2	--	--	40	1

The criterion variables which will serve as basis for evaluating the effect of the treatment variables (suburban school placement and supportive team assistance) can be grouped into four (4) general headings:

a) Mental Ability

1. Wechsler Intelligence Scale for Children
2. Primary Mental Abilities

b) Academic Achievement

1. Reading
2. Listening
3. Arithmetic

c) Personal-Social Development

1. Sociometric Status
2. Test Anxiety
3. Attitudes
4. Teacher Ratings
5. School Attendance
6. Vocational Aspiration

d) Creativity

1. Picture Completion
2. Circles

These data will be collected at four points: September, 1966, as a base; May, 1967, to evaluate effects after one year; September, 1967, to assess loss during the summer; May, 1968, to evaluate effects after two years. The basic statistical tests to be used will be analyses of variance and covariance. All data will be analyzed for the interaction of the following variables with the primary variables: age, sex, grade placement, school system, and where the N permits, school.

In addition, case study materials reported on a weekly basis by teachers will be utilized in an attempt to discover patterns of growth and development. Along with this approach there will be data collected which will indicate parental involvement and attitude as well as neighborhood reaction to a child's placement in the suburbs. It is anticipated that there will be significantly greater growth for the Experimental Ss as a group, but it is also hoped that evidence as to most productive and effective intervention for pupils with differing characteristics may be revealed by careful manipulation of the results.

The techniques described above will be employed on the total samples. However, it is expected that smaller samples drawn from these samples will be used to study other areas such as speech improvement, frustration tolerance, and personality variables. The major outcomes of the Project will be evaluated from this design framework by means of the following specific hypotheses stated here as predictions. For operational purposes, a "statistically significant difference" shall be defined as a deviation of such magnitude that its likelihood of occurring by chance does not exceed one in twenty.

1) Experimental Ss will have significantly greater gain scores than control Ss in:

   a) all measures of mental ability
   b) all measures of academic achievement
   c) all measures of cognitive flexibility (creativity)

2) Experimental Ss will show significantly greater decrease than control Ss in measures of:

   a) general anxiety
   b) test anxiety

3) Experimental Ss will not differ significantly from control Ss in sociometric measures of:

   a) acceptance by classroom peers
   b) acceptance by neighborhood peers

4) Analyses of teacher report data on Experimental Ss will show a pattern of sequential responses which follows the following trend for Ss who show significant gains in academic performance: uncritical acceptance by the teacher; more realistic appraisal by the teacher, but with a tendency to emphasize assets; a tendency to recall and report successes and achievements; attainment of a plateau in terms of reporting pupil behavior as being relatively unexceptional and consistent.

Appendix C

Category	Examples
INSTRUCTIONS	
A. Materials	
Characteristics and amount	teacher prepared, commercial student prepared
Content--specifically, the amount, nature, or characteristics of topics related to urban environments or problems.	
B. Lessons	
Interpretation	by teacher or student and the amount. "What would have happened if there were no Civil War."
Deviations within lessons	Does the teacher allow students to introduce or follow issues that may lead away from lessons?
Spontaneity	Does T. allow asides, immediate student reactions, etc. during lessons?
Opportunity for Participation	Does T. call on all students? Do faster ones dominate? Are slow ones encouraged and given a chance?
Individual Participation	Amount of individual reading, board work, participation.
*What are the project student's reactions during recitation? How much participation, attention, cooperation?	
C. Motivation	
Origin	teacher, children, a combination through some form of theme.
Pursuit	Does T follow children's ideas, accounts even fantasies?
Characteristics	What is discussed? How is the environment utilized?
D. Evaluation-Achievement	
Type	tests, oral statements, displays of students works. (Are project students works displayed?)

Category	Examples
TEACHER-STUDENT INTERACTION	
A. Humor	Does T utilize humor to include students as opposed to ridicule.
B. Address	How does she address individuals or the class? "Boys and girls." "Students." "Children" Last names--first names.
C. Feelings	Does she express or discuss her own feelings and attempts to elicit those of the students?
D. Reward-Punishment	How does she express her favor or disfavor. "I'm proud of you." "I like obedient children."

*Examples of specific interaction with project students.

Category	Examples
DISCIPLINE	
A. Form	
Verbal-direct	"Sit down." "Don't do that."
Verbal-indirect	"Please write the word." "Why don't you put your books away."
Auditory	Clapping the hands, striking the piano
Visual	The evil eye
Physical	Holding, touching, etc.
B. Amount	How many discipline instances during any one visit.
C. Tone	Must the class be completely silent. How much noise is allowed.
D. Consistency	Is the teacher consistent with her rules and enforcing them?
E. Student pressure	Are there occasions when students discipline or assist the teacher in this area by bringing pressure upon others. "Sssh, be quiet."
PHYSICAL ORGANIZATION OF CLASSROOM	
A. Characteristics	Straight rows, tables, clusters of two and three desks.
B. Room divisions	Are their study areas, work areas, hobby areas, reading areas, etc.
C. Interaction	Does room organization assist teacher-student and student-student interaction.
STUDENT INTERACTION	
A. Opportunity	Does the seating, lessons, and assignments allow or encourage interaction. Learning groups, work groups, teacher's assistants.
B. Characteristics	Describe interactions. Students selecting one another to write spelling words on board, or to clean the desks, etc.

*Degree of project student's "mix." Do they choose others, are they aggressive, moderate, or retiring in their interactions.

APPENDIX D

April 1, 1967

HEADSTART EVALUATION AND RESEARCH CENTER

Boston University
School of Education
Boston, Massachusetts

CLASSROOM OBSERVATION AND WRITTEN REPORTS:

INSTRUCTIONS FOR OBSERVERS

INTRODUCTION

In order for us to most effectively use your observations of classrooms, it will be necessary for us to have several kinds of reports which will reflect, in a variety of ways, the teacher and child behaviors which you have observed in the classes assigned to you. These reports must be detailed enough and must include sufficient affect so that other readers can read a series of reports and rate them in ways similar to the ways in which you will be requested to rank and rate the various classes that you are observing. This does not call for the suppression of your biases, but rather the ready admission of them and explicit attempts to distinguish between those behaviors which you take a liking to as differentiated from those behaviors which you think are of high quality. This means that you have not only to observe and report what you see, but also to assimilate what you see into the working model that is represented by your ideas, feelings, and experiences. We shall bring together the various models of the several observers into an integrated framework which is controlled partially by the outline which was distributed and, further, by a series of scales which will be presented to you after you have concluded your observations.

The process of abstracting from classroom behaviors to your observations, and then to your written reports and then, still further, to a series of relevant scales is a difficult one which will depend on the kinds and degrees of differences that are found between the various classes that you observe. Difficulty is, at the same time, a function of the differences that exist within any one class over a period of time. The process that is being constructed will give a more or less clear indication of whether classes are describably and meaningfully different and, to a lesser extent, the degree of differences between these classes. The reliability of the process will depend upon the clarity and comprehensiveness of the written reports. It is necessary both to be able to carefully describe the classes that we see as well as to make some clear statements about how equivical or unequivical the system of measurement is when it is put to a fair test. In this case the tests will include the observations of classes by different observers as well as the ratings of the classes by individuals who have not seen them, but who have access to the written reports.

OUTLINE FOR CLASSROOM OBSERVATION

This outline, which was distributed to each observer, is not to be used as a checklist or as an observational guide. Rather, it should be used in the following way; observers should read and reread it carefully so that they are quite familiar with the various categories and sub-categories that describe a more or less all inclusive listing of behavioral possibilities in classroom situations. The outline does not represent a mutually exclusive system nor does it cover the detail which would bring it so much closer to the classroom situation. Observers should be quite familiar with it, but they should not actively use it during the course of their observations. After completing process reports, they should refer back to the outline in order to sensitize them to the kinds of information they are getting and the behaviors and situations which they should attend to on future visits to the class. The outline will be referred to again when the summary report is discussed below.

PROCESS REPORTS

These should include a detailed statement of everything that is observed in the classroom including the behaviors of the teacher and children, the physical characteristics of the classroom, the materials that are used and any other observations which are pertinent to discussing the class. These reports are to be thought of as the total of the observer/class interaction and they should not exclude the observer and his feelings from the report.

Observers will differ in the way in which they construct this process report, but the end result should be pretty much the same. Some of you will take notes as you are observing the class, others should write out a detailed report immediately after you leave the class, still others might develop a system for sketching out their observations so that they can then be transcribed into a running commentary describing what was seen and how it was seen.

These process reports are the raw materials for everything that follows and a single report should be made out for every observation of the class. Therefore, each observer will have at least two and preferably three process reports on each class that they observe.

It is hoped that these reports will not simply be a rather dry chronological listing of everything that happens but that they will include appropriate adjectives and interpretations that are a part of the observational process. The total interpretation of a given teacher and classroom will come in a later report. What we are interested in here are the more minute interpretations of the specific behaviors that are observed. Although we are not specifically attending to fragmentary quantitative questions such as how many times a given child is reprimanded or how often the teacher talks opposed to how often the children talk. But we should be quite aware of duration and quantity and appropriate notes should be made about persistent kinds of behaviors that take place.

The process reports will be used in two ways: in the first place they will be used by independent readers who will make judgements about the classes from reading these reports; in the second place, they will be used to document the findings of this survey and relevant parts of these reports will be abstracted and integrated

into a total report of all classes. In both uses of the process reports it is necessary to have writing that is provocative and comprehensive and that projects the reader into the classroom so that he gets a feeling for what is taking place and how it is taking place.

ANALYTICAL REPORTS

There should be an analytical report for each visit to a classroom. This report represents the observer's explanation and synthesis of what he has seen. It can draw uponthe material from the process report but it is not an observational report as such but rather a critical appraisal of the classroom for the period of time that was observed. If there is no substantial difference between several process reports, it is possible to combine several of these into one analytical report. However, in general, there will be a separate analytical report for each process report.

The analytical report should refer back to the outline and should assess which parts of the outline are most relevant for the class under consideration, and what kinds of information are not readily obtainable either because of the structure of the class or because of the accident of having observed a particular kind of class or a particular segment of the curricular.

SUMMARY INTERPRETATION

There will be one summary interpretation for each teacher that you observe. This will draw upon the several process reports and analytical reports and it should integrate all of the material that you have in your possession. This summary report should have two sections to it: first, an open-ended judgemental and inferential report describing the essential of the observed behavior of the period of two or three observational periods. It should be completely openended (projective) in that you are free to draw on any material that you have in any of the visits and you should underline freely as you see fit. The second part of the summary report should closely follow the outline and should comment on each of its major sections. If there are many omissions here then it should be clear that you have not observed the class either a sufficient number of times or sufficiently long enough on any one time. We continually have to address ourselves to the question of whether we have observed behaviors which make any particular class comparable to other classes.

Classroom observation is continually plagued by the lack of comparability of data. In one class a teacher may do a large amount of talking and it might be considered to be extremely important in assessing her effectiveness. Another teacher may also do a lot of talking but it might be trivial compared to other behaviors which she displays in her work with children. This means that the problem of describing and evaluating teachers has to consider more and less effective behaviors as well as behaviors which are not applicable in an assessment of effectiveness.

Somewhere along the line, we must make judgements which stem from our descriptions and which say something meaningful about the degree and kind of impact a particular teacher might have. We must obtain a sufficient amount of material on teachers to make judgements about how effective they are with respect to the teaching of academic subject matter, of providing an environment for individual self-determination, and encouraging appropriate inter-personal relationships between the teacher and the children and between the children.

APPENDIX E

May 1, 1967

HEADSTART EVALUATION AND RESEARCH CENTER
Boston University

Scales for Rating Elementary School Classes*

1. Involvement and interest of children

   Indifference, Apathy ——— Curiosity, Absorption

2. Purposeful behavior of class

   Aimless Wandering ——— Direct Responsive

3. Source of direction of academic activities

   Teacher ——— Child

4. Nature of control over behavior (Teacher)

   Coercion, Threat ——— Trust, Respect

5. Effectiveness of behavioral controls (Teacher)

   None, Class out of control ——— Complete, Class well controlled

6. Quality of presentation of subject or materials (Teacher)

   Pedestrian, Routine ——— Creativity Variety, Innovation

7. Differentation of instruction (Teacher)

   Monolithic, Uniformity ——— Highly differentiated, Individually discriminate

8. Teacher reaction to classroom situation (Teacher)

   Unhappy ——— Indifferent ——— Happy, Involvement with children Obvious enjoyment
   Hostile

9. Reinforcement of behavior of children (Teacher)

   Not apparent ——— Frequent

10. Nature of reinforcement (Teacher)

    Negative, Punitive, Threatening ——— Bribery ——— Positive, Approval, Encouragement

*All teachers observed by a given rater are to be sorted into five categories so that two-thirds of the teachers are in categories 2,3, and 4; one-third are to be 1 and 5. Category 1 is the left hand side of each scale and category 1 to the right hand side. Category 3 is an intermediate category.

APPENDIX F

VARIABLES FOR OBSERVATIONAL SCHEDULES
(WITH SELECTED REFERENCES)

Variable Types

Object	Independent Characteristics	Dependent Behavioral	Dependent Curricular
Pupil	School background Placement procedure, 2 Diagnostic Information: 2, Aptitude Achievement, Personality Family-Home	Problem solving, 2 Motivation, 2 Attention, 1 Curiosity, 1 Activity, 2 Origination Mobility Participation Disruption Individuality, 2 Pupil-Pupil Interaction, 2 Sociometric variables	Grouping, 2 Getting help, 1 Independent Activitity, 1,2
Teacher	Education Experience Age Sex Certification(s) Professional Organizations and Journals Attitudes, 3	Preparation, 2 Direction, 2 Presentation Variety,1 Sequence, 2 Verbal-Nonverbal Management-Discipline Empathy-Support-Humor,1 Evaluation-Criticism, 2 Reinforcement-Rewards, 1	Use of curriculum guide, 2 Textbooks, workbooks, 2 Teacher-prepared materials, 2 Evaluation-Reports, 2
Pupil-Teacher Interaction	Not applicable	Direction-Initiative, 1 Social Organization-Teacher or pupil centered, 1 Delegation of responsibility	Differentiation, 1,2 As related to content and procedure
Classroom	Demographic Location-Type of community Size Equipment Supervision-reported Level	Supervision-observed Climate, 1 Routines, 2 Discussion, 2 Competition, 2 Order-Disorder, 2	Content, 1: Academic-Vocational-Crafts-Social-physical and recreational Subject or project Consultants-music, art, physical education, 2

1. Classroom Observation Code Digest (Cornell, Lindrall, Sarpe, 1952)
2. Schedule for observing special class for mentally retarded children (Blatt, 1963)
3. Minnesota Teacher Attitude Inventory (Cook, Leeds, and Collis, 1951)