ED 241 524 SP 024 011

AUTHOR Weisman, Richard M., Ed.; Casini, Barbara P., Ed.
TITLE Three Views on Improving Basic Skills Instruction.
INSTITUTION Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Aug 80
NOTE 43p.; Papers presented at the 1978-79 Tri-State Conference on Improving Basic Skills Instruction.
PUB TYPE Collected Works - Conference Proceedings (021) -- Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Basic Skills; *Classroom Environment; *Classroom Research; Classroom Techniques; Cognitive Processes; Elementary Secondary Education; *Instructional Improvement; Student Behavior; *Teacher Effectiveness; Teaching Methods; Time on Task

ABSTRACT

Three researchers, addressing the problem of instructional improvement, identify sound research findings and cite problems associated with the transfer of these findings into classroom practice. Donald M. Medley, in "An Overview of Research on Classroom Teaching," identifies three variables which consistently differentiate between effective and ineffective teachers: learning environment, use of pupil time, and quality of instruction. Inconsistencies between these research findings, educational theory, and common sense are noted. In "Implications of Research for Adaptive Teacher Preparations," Robert S. Soar separates four domains of the learning environment--emotional climate, student behavior, learning tasks, and thinking processes. Soar's research indicates conflicts with accepted educational practice and theory. In "Using Feedback to Change Teacher Behavior," Frederick J. McDonald addresses the issues of transferring research results into practical applications. He asserts that it is the researchers' responsibility to develop a more simplified system for conceptualizing teacher performance, observing teacher behavior, and providing feedback to change teacher behavior. Several suggestions are offered to enhance the effectiveness of inservice training for teachers. (JD)

Three Views on Improving Basic Skills Instruction

Papers presented at the 1978-79 Tri-State Conference on Improving Basic Skills Instruction

Edited by
Richard M. Weisman and Barbara P. Casini

August 1980

Basic Skills Component
Research for Better Schools, Inc.
444 North Third Street
Philadelphia, Pennsylvania 19123

The work upon which this publication is based was funded by the National Institute of Education, Department of Education. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Contents

Introduction

Our schools are challenged to provide basic skills education that meets the needs of both individuals and society. Over the past two decades we have not been as successful as we would like in meeting this challenge. The reasons for low achievement are numerous and complex. It is the conviction of Research for Better Schools, Inc., (RBS) that instructional effectiveness would be strengthened if we could transfer more effectively the findings of research into classroom practice. However, there are many questions that must be attended to if such a transfer is to take place.

In 1978 the Basic Skills Component of RBS sponsored a Tri-State Conference on Improving Basic Skills Instruction to explore major issues related to a research-based approach to staff development and the improvement of basic skills instruction. Three distinguished researchers were featured speakers at the conference: Dr. Donald M. Medley of the University of Virginia, Dr. Robert S. Soar of the University of Florida, and Dr. Frederick J. McDonald of Educational Testing Service. Their presentations constitute some outstanding work on basic skills instruction which is relevant to this day.

Medley likens the practice of teaching today to medical practice a century ago, with respect to the reluctance of physicians to base their treatments on scientific research rather than on theory, experience, and common sense. Based

on a critical review of a large volume of research in teacher effectiveness, Medley has identified three variables which consistently differentiate between effective and ineffective teachers: learning environment, use of pupil time, and quality of instruction. He also draws attention to interesting inconsistencies between these research findings, educational theory, and common sense. Medley's findings have important implications for improving instruction. Utilization of the research findings is discussed in terms of current conceptions of competent teaching. Research, according to Medley, must have a central function if teaching, like practicing medicine, is to become more science than intuition.

Soar develops, from some of his own longitudinal research, a framework for conceptualizing teacher effectiveness which relates to one of Medley's variables, classroom learning environment. The framework separates four domains of learning environment - emotional climate, student behavior, learning task, and thinking processes. Within each domain a different balance between freedom and structure is functional for optimal student learning. Like Medley, Soar's research indicates conflicts with accepted educational practice and theory. For example, the unidimensionality of accepted concepts is questioned along with common assumptions that the relationships between process variables and outcomes are linear. Soar's framework suggests a basis for development of a more effective classroom management system.

McDonald addresses the issue of transferring research results into practical applications. He has identified a pattern of effective teaching behaviors based on student achievement gains identified in Phase II of the California Beginning Teacher Evaluation Study which he co-directed. The challenge of using these data to change teacher behavior is illustrated by an inservice program which he implemented in a Trenton elementary school. According to McDonald, it is the researcher's responsibility to meet this challenge by developing a more simplified and meaningful system for conceptualizing teacher performance, observing teacher behavior, and providing feedback to change teacher behavior. Several suggestions are offered to enhance the effectiveness of inservice training for teachers.

In addressing the problem of instructional improvement, Medley, Soar, and McDonald, have identified sound research findings and cited problems and needs associated with the transfer of these findings into classroom practice. Although many issues remain unresolved, these researchers have made important contributions to our understanding of the value of research for improving basic skills achievement. Researchers, developers, and educators need to give serious attention to these works.

David C. Helms
Director, Basic Skills Component

An Overview of Research on Classroom Teaching

Donald M. Medley
University of Virginia

In an editorial in a recent issue of *Science*, Dr. Lewis Thomas of the Sloan Kettering Institute remarked on the great reluctance the medical profession showed around the end of the last century in accepting two "catastrophic" findings of nineteenth-century research. These days, the contributions of research to the practice of medicine are so widely known and so generally accepted that it is difficult to imagine how different things were almost a century ago. The two findings that revolutionized medicine were, first, that a large proportion of sick people got well regardless of anything the physician did; and, second, that almost everything in the extensive armamentarium of therapy available to practitioners in those days was worthless and had no real effect on patients at all.

The repertory of treatments available to, and used by, nineteenth-century physicians was vast, and included all kinds of medicines and remedies as well as procedures involving the application of electric currents and leeches; most of these remedies had either firm theoretical bases or long experience to back them up. The evidence that none of these things had any real efficacy—that what the witch doctor, the snake oil vendor, or the qualified physician prescribed were all equally beneficial—was available for quite a few years before the medical profession accepted it. Small wonder. But the evidence was inescapable, and the profession was forced finally to accept the fact that, until that time, it had survived mainly by taking the credit for the spontaneous remissions and by disavowing the blame for the failures. It was, of course, this accumulation of successful clinical experience that made it difficult for physicians to accept the findings of the research.

When Dr. Thomas got his own training at the Harvard Medical School in the 1930s, he tells us, medicine was still in a state of "therapeutic nihilism" in which physicians were not trained to treat patients, but only to diagnose their ills and make

accurate prognostications - to tell patients what their chances of recovering were, how long it would take, and so on. It was not until the 1940s, when research produced penicillin, that the present situation came about. Today's physicians have available to them a number of effective treatments that have not only lengthened our life span but also freed us from the many disabling effects of illness. In retrospect, it seems remarkable that the profession survived the intervening years at all, and even more remarkable that the positive image of the profession was unaffected.

There seem to be some striking similarities between the practice of medicine as it was a century ago and the practice of teaching today. The teacher of today has a large armamentarium of things to do that theory and long experience indicate are effective in helping most pupils learn. Few of these practices are backed up by any sound research evidence showing that they actually produce learning that otherwise would not have taken place, but they are nevertheless firmly entrenched in use.

The possibility that the practice of teaching now, like the practice of medicine then, survives by taking credit for what is learned by the more apt pupils—those who would learn as much without the teacher's interference—and by trying to avoid taking blame for the failure of less gifted pupils to learn is a very real possibility indeed. It is certainly compatible with much of the research findings to date, which tend to be weak and inconsistent at best. This unpleasant possibility leads one to ask: How long will it take the profession of education to face this possibility and to mount the massive research effort needed to begin building a sound basis in research for the practice of teaching? Is there any point in waiting for "catastrophic findings" like those that shook medicine so long ago? It seems to me that there are strong current pressures on the profession (manifest in the demands for accountability and in PL 94-142, among other places), the likes of which the medical profession has never encountered. Certainly the public has never demanded that the physician cure every patient, as the public is, in effect, demanding that teachers do. Pressures like these suggest to me that the teaching profession may not have the time the physicians had to set their house in order before the public gets wise. The first malpractice suits in education are already beginning to appear, and there will be more. It will be wise to be able to base our defense on evidence that current practice reflects the best research available; the more extensive, and the sounder, that research base is, the better off we will be. I understand it is the purpose of this Conference to move in that direction—to base staff development on current research knowledge.

This is the context in which I would like to share with you the results of an examination of the research base for the practice of teaching that I completed recently.[1] By research in the practice of education, I mean research designed to find out how a teacher should behave in the classroom in order to be effective in helping pupils in that class learn better than they could without that teacher's help. I am excluding research in the knowledge base of education, in human growth and development, in how children learn, in the subject matters or disciplines taught, and so on. Knowledge of anatomy and physiology seems important to the practice of medicine, but the possession of this knowledge does not in itself qualify a person to practice the profession. What I have in mind is research into the procedures, the behaviors a teacher must perform in order to capitalize on such knowledge—what a teacher must *do* to be effective.

Research in teacher effectiveness (as I shall call it) has been going on for almost a hundred years now, that is, for as long as any other kind of educational research. Not everything that has been called research in teacher effectiveness by its perpetrators fits my definition of the term, however. As I use the term, research in teacher effectiveness refers to efforts to study the behaviors that make the teacher effective.

The earliest attempts at research in teacher effectiveness sought to identify characteristics of effective teachers by asking students and former students to

[1] Medley, D. M. *Teacher competence and teacher effectiveness: A review of process-product research.* Washington, D.C.: American Association of Colleges for Teacher Education, August 1977.

describe characteristics of the most and least effective teachers they knew. In later studies, ratings of teachers judged by their supervisors to be most or least effective were analyzed to discover how the two groups differed. It was not until around 1960 that a type of research called "process-product" began to appear. A process-product study is one in which objective measurement of various dimensions of behavior in a teacher's classroom are correlated with measures of pupil gains in achievement. This is the kind of research whose findings I plan to summarize here.

In the literature search on which my review of the process-product research was based, we examined almost four hundred studies, which reported thousands of correlations between teacher behaviors and pupil learning. Most of the correlations were small, and many of them conflicted with correlations reported in other studies. Many of the studies did not conform to what I regarded as minimal standards of quality for process-product research in design, in instrumentation, or in other respects. Under the assumption that a poorly designed study is more likely to yield incorrect findings than a well-designed one, I decided to disregard all studies that failed to meet certain criteria of quality, expecting to eliminate many of the inconsistencies between findings of different studies.

But even a well-designed study can yield spurious findings: idiosyncrasies in the schools, classes, or teachers used in any one investigation can lead to correlations that would not appear in other settings, or with other teachers. Such correlations are likely to be small; it is rare to encounter a large correlation that is the result of such chance conditions. Therefore, to avoid being misled by spurious correlations, I disregarded all correlations smaller than .39.

These limitations on my review eliminated all findings from 95 percent of the studies; after the dust settled, I was left with some six hundred correlations from just fourteen studies I considered reliable enough to report. This meant, of course, that I took the risk of overlooking or missing a substantial number of important but smaller relationships. What I have to share with you, then, is not a complete set of findings, but only the strongest, most dependable findings of this research.

Let me say a word or two about the monograph in which the findings are reported. The goal I set myself was to put the reader in direct contact with the findings without interposing any interpretations of my own. The 613 relationships between classroom behavior and outcomes are presented in 43 tables, organized so that consistencies and inconsistencies between them are readily apparent; in this way, the reader may make his own interpretation. Anyone interested in using the findings should study these tables himself and draw his own conclusions. The only rule I would like to enforce is that the conclusions must be based on *all* of the relevant findings. To pick and choose only the results that agree with one's preconceptions is to defeat the purpose of the monograph and to invalidate the conclusions.

In order to give you some idea of the nature of the findings, I will present a brief summary; it is important to bear in mind that in doing so I cannot avoid mixing in certain interpretations of my own, which may differ from any interpretations you may make. I repeat, the raw results are available to anyone who cares to examine them.

What I consider the most striking finding is that, once the results of this research were screened in the way I have described, much consistency in the findings of different studies was revealed. A considerable number of relationships were verified in two or more independent studies—in studies done by different people, in different parts of the country, working in different populations of pupils and teachers. These are the relationships that interest me, mainly because, since each such relationship was also large in size, its existence may be regarded as well established. In other words, the likelihood that further research would fail to confirm any of them is very slight.

These dependable results were all found in classes of pupils mainly from homes of low socioeconomic status in grade three or below. Because federal funding strategy in recent years has given high priority to research in the teaching of disadvantaged pupils during their first few years in school, a critical mass of data about this particular group has been accumulated. It is unfortunate that we do not have comparable

amounts of data about classes of nondisadvantaged pupils in these same grades, or about classes of pupils of any kind in the higher grades. There seems little reason to doubt that if comparable support were given to research in these kinds of classes, comparable numbers of reliable conclusions would be available.

It is important to remember (in case I forget to remind you) that, when I speak of effective or ineffective teachers from now on, I mean teachers of classes made up mainly of disadvantaged pupils in their first few years in school.

The dependable relationships seem to me to fall into a systematic and consistent pattern of differences between effective and ineffective teachers of disadvantaged pupils in the first three grades. These teachers differ, first, in the kind of *classroom learning environment* they create and maintain; second, in their *use of pupil time*; and, third, in the *quality of instruction* they provide.

Table 1
Learning Environment in an Effective Teacher's Classroom

Classroom Behavior	Frequency of Behavior	Number of Studies
Disruptive pupil behavior	low	5
Criticism	low	2
Permissive behavior	low	3
Time on management	low	3
Praise	high	3

Table 1 shows how the environment in an effective teacher's class differs from that in an ineffective teacher's class. Process variables, or classroom behaviors, are shown at the left; the relative frequency of each behavior in the more effective teacher's class is shown at the right. The numbers indicate the number of different studies reporting each relationship. The effective teacher's classroom tends to be more orderly and less permissive than the ineffective teacher's classroom, a more supportive and less hostile place, and one in which less class time is used to maintain order. Clearly, the effective teacher maintains order more skillfully and in a positive, nonthreatening way.

Table 2
Use of Pupil Time in an Effective Teacher's Classroom

Classroom Behavior	Frequency of Occurrence	Number of Studies
Time in academic activities	high	4
Time in large group with teacher	high	2
Time in independent small groups	low	2
Time in seatwork	low	4

Table 2 shows the findings related to the use of pupil time. The effective teacher's pupils spend more time in task-oriented or "academic" activities, and in a large group led by the teacher. The amount of time pupils spend in independent activities, that is, working as individuals or in small groups without the teacher, is

greatest in the classes of less effective teachers; effective teachers use these activities relatively infrequently. The implications are that the more time a pupil spends on the content being taught, the more the pupil learns, and that the way the effective teacher usually keeps pupils engaged with content is by organizing them in large groups under her or his control.

Table 3
Quality of Instruction in an Effective Teacher's Classroom

Classroom Behavior	Frequency of Occurrence	Number of Studies
Low cognitive level questions	high	4
High cognitive level questions	low	3
Amplification, discussion of pupil answers	low	3
Pupil questions	low	3
Feedback on pupil questions	low	2
Attention to pupils during seatwork	high	2

Table 3 shows results related to what I have called the quality of instruction. During discussion periods, effective teachers ask more low-level questions and fewer high-level questions than do ineffective teachers; their pupils ask fewer questions, and they receive shorter shrift from the effective teacher. During periods when pupils are working independently, the effective teacher pays closer attention to what they are doing than does the ineffective teacher, even though (as we saw in Table 2) they spend less time in such activities.

There you have it. These are the differences that research in classroom teaching has clearly established between teachers of disadvantaged pupils in the first three grades who are learning most and teachers whose pupils are learning least. Let me mention in passing that others who have reviewed this research using different procedures have reached substantially the same conclusions.

Our first finding was not surprising. We found that pupils learn best in classrooms that are orderly and supportive, and are kept that way with a minimum of fuss and bother on the teacher's part. This is certainly obvious; one may be inclined to question whether we needed any research to tell us this.

Our second major finding, that effective teachers keep their pupils engaged in learning-related activities a greater part of the time, also agrees with common sense. But large-group instruction is something teacher educators teach their students to avoid, particularly at these low grade levels.

Our third finding, that discussion in classes of effective teachers is low-level and teacher centered, and that the effective teacher's pupils ask few questions and get short answers, seems also to contradict what many teacher educators train their students to do. These findings seem to many of us to contradict what everyone knows. A colleague of mine, Harold Mitzel, used to say that the real purpose of research is to enable us to distinguish between the things we know that are so and the things we know that are not so. That this observation is much more profound than it may at first appear is manifest in a tendency we all have to accept those research results that agree with our own preconceptions—the ones, we say, that make sense—and to reject those that do not. Results that upset long-held beliefs (sometimes called prejudices) are suspect; the usual response is to question the validity of the research that produced them.

Let me assure you there were no differences in the soundness, validity, or any other aspects of quality among the research that yielded any of the findings I have reported. They all come from the same fourteen different studies, and each study and each correlation passed a set of severe quality tests before it was included.

Before I turn to the question of how research findings can be used in the improvement of instruction, let me try to anticipate and answer some questions that have doubtless occurred to you.

The first question is: What kinds of changes in pupils were measured as a basis for deciding which teachers were more or less effective? The primary measures used in all fourteen studies were adjusted mean gains of pupils on standardized tests of reading and arithmetic. In addition, some studies also used measures of pupils' attitudes toward school or toward themselves. Any large correlations between classroom behavior and such measures (.39 or larger) were also reported in the study.

An attempt was made to measure pupil gains separately on items of low and high complexity, in view of the possibility that different patterns of behavior may be more related to low-level outcomes than were related to high-level gains. No such differences were found, nor were any important differences found between patterns of teacher behavior related to student reading gains and patterns of teacher behavior related to student arithmetic gains.

The second question is related to the first: Did teachers' efforts to achieve high cognitive gains have side effects on pupil attitudes or their self-concepts? To answer this question, we examined all instances in which the same teacher behavior was found to correlate with both affective and cognitive gains. There were ninety such pairs, three-fourths of which were of like sign, and one-fourth of which were of opposite sign. That is, in three cases out of four, a behavior associated with high gains in reading or arithmetic was also associated with high affective outcomes. Pupils in a class in which they are learning to read and do arithmetic tend to like school and to grow in self-esteem.

The third question is: Does this same pattern of behavior characterize effective teachers of classes of nondisadvantaged pupils in the first three grades, and effective teachers in the higher grades? As I mentioned earlier, because funding agencies assigned priority to research in teaching the disadavantaged, particularly in the elementary grades, most of the research that survived our criteria was done in such classes. The monograph reports a number of correlations in classes of these other types, but not many of them have been verified.

As far as they go, these results suggest that, in the higher grades, effective teachers maintain the same kind of learning environment identified earlier, but that the quality of instruction offered by teachers in the higher grades differs from the instruction in classrooms of effective teachers of disadvantaged pupils in the first three grades—particularly with respect to the kinds of discussions they conduct. Evidence related to the use of pupil time in the upper grades was too sparse to comment on.

In order to find out whether effective teachers of nondisadvantaged pupils behaved the same way as effective teachers of disadvantaged pupils in the first three grades, we examined all pairs of correlations between the same behavior and the same outcome, one obtained in nondisadvantaged classes and one in disadvantaged classes. Eighty-four such pairs were found, of which 38 percent were of like sign and 62 percent of opposite sign. This means that, in two out of three cases, the behavior of the effective teacher of disadvantaged pupils was the same as that of the ineffective teacher of nondisadvantaged pupils in the same grade range. This strongly suggests that opposite teaching strategies are most effective with the two kinds of pupils. If this is true, a teacher teaching an integrated class—one with pupils from both high- and low-SES homes—may have a problem. Almost anything he or she does that will be effective for half the class will be ineffective for the other half. These results are not, of course, as dependable as those reported above.

As I have suggested, the only satisfactory way to find out what the research really says about effective classroom teaching is either to study the 43 tables and the

613 correlations in the monograph or to go to the original studies. What I have presented represents my own attempt to summarize these findings as a basis for considering how they may contribute to efforts to improve instruction through staff development or inservice teacher education. I would like to conclude these remarks by making some comments and raising some issues related to research utilization.

One issue I want to discuss is the professional development of teachers: What are the objectives of staff development? If professional development is seen as a matter of getting all teachers to behave or to teach in the same way--in some way regarded as the "best" or most effective way of teaching—then the first question to ask is whether the teaching style revealed in the three exhibits is that "best" way. Although this style may not be the best of all possible teaching styles, it seems to me to be better than the one many teachers are using. If more teachers of disadvantaged pupils learned to teach the way the most effective teachers now in the schools teach, these disadvantaged pupils should improve substantially. There would still be room for improvement, but a real gain should be apparent if staff development concentrated on helping the least effective teachers improve their skills in environmental maintenance, in constructive use of pupil time, and in quality of instruction. A considerable amount of knowledge is available to us about the techniques for achieving these goals--positive ways of maintaining discipline, involving pupils, using large-group instruction. If teachers do not acquire these skills in preservice training, it may be because preservice teacher educators do not assign high priorities to such skills; certainly it is not because nothing is known about them.

Another view of professional development seems to be based on the idea that instruction can best be improved by helping teachers enlarge their repertoires of skills or competencies. There is no presumption that one particular style of teaching is best for all teachers; it is assumed that equally effective teachers may behave quite differently. The teacher is expected to select from a set of alternatives—all of which he or she has mastered--the ones best suited to his or her individuality. From this view, the research findings may be said to have identified those ways of behaving that are most likely to prove useful to any teacher who has the kind of pupils represented. Among all the various skills a teacher may acquire, skills related to using large-group instruction, keeping order in a supportive way, and asking low-level questions are most likely to prove useful. Other skills are also recognized as potentially useful and may form part of the training opportunities available, but those identified in the research would receive highest priority. If this view were adopted, the immediate impact on pupil achievement might not be as great as that expected under the view described earlier; it does, however, offer the opportunity for all teachers, not just the less effective ones, to grow.

A third view of professional development regards teacher effectiveness as dependent not only on how the teacher behaves—what he or she does—but also on when or for what purpose he or she does it. From this view, the behaviors identified in the research are recognized as important ones for teachers to acquire. But how much a teacher's effectiveness increases as a result of acquiring these skills depends on how much wisdom or good judgment the teacher shows in employing the skills. The role of the research findings is to identify the skills to be developed; there is also a need for information from another kind of research regarding when and for what purpose a skill should be used.

There are, of course, other ways of conceptualizing competent teaching, but these three seem to be the most useful. No matter which of the three is adopted in the development of a program of teacher evaluation or staff development, research findings have a central function.

Earlier in this discussion, I told how reluctantly medical practitioners and medical educators made the transition from a phase in which the practice of medicine was an art to the present phase, which Dr. Thomas describes as a mixture of science and technology. The transition came when they recognized research as the only source of dependable knowledge about the effectiveness of treatment. The reluctance was due to the great confidence they had in the lore of the profession,

much of which was contradicted by research results. And even after they came to accept the research, it was many years before the research began to pay off.

It is useful to draw a parallel with the practice of teaching. Today teaching is based on much the same kind of lore that nineteenth-century medical practice was based on. Research has not yet turned up any catastrophic findings, although there are growing doubts about the efficacy of the methods we use, and I have heard it suggested that we have survived this far by taking credit for what some pupils have learned in spite of, rather than because of, our teaching. Must we wait for research to destroy what we have before we begin to listen to what it is telling us? At the rate things are going, this may take another hundred years.

The plans of this group, as I understand them, seem much more sensible: to begin listening to what the research is telling us now, to begin incorporating research findings—incomplete though they are-- now, so that teaching can change, gradually rather than abruptly, from an art to a mixture of science and technology. The change must come; let's be part of the change rather than part of the resistance to it.

Implications of Research for Adaptive Teacher Preparation

Robert S. Soar
University of Florida

For those of you who have not been in an elementary school for a while, I would like to tell a favorite story of mine that may recall to you what elementary schools are like, since all of the data I will report are from that source.

It had been a perfectly terrible day in a first grade classroom in a big northern city. The weather was so bad that the children could not go out, and you know what it gets like by the end of the day when that happens. The teacher had helped with an endless parade of coats, hats, boots, mufflers, buttons, and all the rest, and she had reached the last child, a grubby little girl with stringy hair and knees that were not quite clean, and a pair of boots that were impossibly tight. The teacher had struggled and struggled and finally got one of the boots on while the little girl stood impassively, and just as the teacher finished the first, the little girl spoke up and said, "These are not my boots, you know."

The teacher ripped off the boot that she had just put on, and then the little girl continued, still impassively, "They're my sister's, but Mother told me I could wear them today." So the teacher figuratively shrugged her shoulders and started to work on the same boot again, and just as she got it on the second time, the little girl spoke up again and said, "But the mittens are mine." The teacher stopped, this time cautiously, and asked, "Oh? And where are the mittens?"

"In the toes of the boots."

I want to draw on the four past studies that my wife and colleague, Ruth Soar, and I have done, and to talk about parallels across those studies. A fifth study, our final year in Follow Through, produced results that really do not fit in with the results of the other four, and for that reason I do not plan to talk about them. I also do not plan to say much about some current work of ours that you may be familiar with, and I suppose I ought to explain why. Our current work is a reanalysis of some of our data in a way that, we think, legitimately lets us look at relationships within classrooms, as

well as relationships between classroom means. We want to be able to answer questions like, "Do children who are high in anxiety respond differently to a disorderly classroom than children who are low in anxiety?" A different procedure of analysis from the one that has been used previously is necessary. This is what we are doing now.

The four studies I plan to discuss begin with one that was finished in 1966, for which data collection began in 1962. The study was carried out in fifty-five classrooms, grades 3 through 6, in the Columbia, South Carolina, area.[1] The students at pretest were a grade level advanced, so these were not the lower-grade/lower-SES groups that are typical of the more recent results.

The second study was our first year in Follow Through,[2] and the results I want to talk about were from twenty first grades, scattered all over the country, for which we had pupil data, primarily low SES though not entirely so. They included six programs in Follow Through, which ran the gamut from the implementation of the British Infant School in this country to the Becker-Englemann program, one of the more tightly structured contingency-management programs.

The third study was of fifty-nine fifth grade classrooms from the North Florida region,[3] roughly a third of them from center-city Jacksonville, about a third from a semirural county south of Gainesville, and the remaining third scattered through a series of exceedingly remote rural counties north and west of Gainesville—so remote, we realized toward the end of the study, that at that time they were out of reach of commercial television.

The final study was a sample of twenty-two first grade classrooms, all in the city of Gainesville.[4]

These latter two samples spanned the socioeconomic range as widely as we could manage, but they were somewhat below average in achievement and probably also in socioeconomic status.

I would like to organize the results in terms of a paradigm that has slowly emerged for the two of us. Most of the results are in the two publications included in the handouts for this Conference,[5] but they are not organized as I will present them now. The organization has been most helpful to us in thinking about the results and perhaps in thinking about teaching in general.

I would like to make a first distinction between emotional climate, on the one hand, and teacher management or control of what occurs in the classroom, on the other. Separating those two is critical, it seems to us, because it is fairly easy to find classrooms in which the four combinations of the extremes of those two dimensions can be found. First, there are classrooms that are very warm and friendly, but that show very little order. This is a fun-and-games classroom where we all have a good time, but not much work gets done. It may even be chaotic, but it is a friendly kind of chaos. Then there are the contingency-management classrooms in which teachers use positive affect very skillfully, and therefore are able, I think, to control students more closely than any other possible way—at least at this grade level. The classroom is very warm, but the control is exceedingly close.

---

[1] Soar, R. S. *An integrative approach to classroom learning* (NIMH project numbers 5-R11 MH01096 and 7-R11 MH02045). Philadelphia, Pennsylvania: Temple University, 1966. (ERIC Document Reproduction Service No. ED 033 749)

[2] Soar, R. S., & Soar, R. M. An empirical analysis of selected Follow Through programs: An example of a process approach to evaluation. In I. J. Gordon (Ed.), *Early childhood education*. Chicago: National Society for the Study of Education, 1972.

[3] Soar, R. S., & Soar, R. M. *Classroom behavior, pupil characteristics, and pupil growth for the school year and for the summer*. Gainesville, Florida: Institute for Development of Human Resources, University of Florida, 1973.

[4] Ibid.

[5] Soar, R. S. *Group learning environments for the early school years*. In *UPDATE: The first ten years of life* (Proceedings from the Conference Celebrating the Tenth Anniversary of the Institute for Development of Human Resources, College of Education, University of Florida, Gainesville, March 29-31, 1976). Gainesville: Division of Continuing Education, University of Florida, 1976. Soar, R. S., & Soar, R. M. An attempt to identify measures of teacher effectiveness from four studies . . . *Journal of Teacher Education*, 1976, 27(3), 261-267.

There are also classrooms where teachers use negative affect as a means of control. They run a taut ship, so to speak. And finally, there are, unfortunately, classroms in which the teachers spend the day screaming at the students lots of negative affect but never get enough order established to teach.

These are the four extremes of control and of emotional climate. In that first study we found we could identify teachers who fitted those four extremes with very little trouble.

Let me point out still another way. Permissiveness is typically described, at least in the literature we have read as a style of management in which the teacher is very warm and supportive and shares decision making with pupils. The assumption, then, is that warmth and freedom go hand in hand, but the data say they are two independent dimensions.

Parenthetically, our hunch is that one of the problems we as educators have confronted, both in research and in thinking about teaching, is that many of the concepts we use probably are not really concepts at all, but muddles of unrelated dimensions. As the computer people say, garbage in, garbage out. If you start with a concept that is garbage, you end up with garbage, and it does not make much difference what you do in between.

I want to go on to talk about the relation of the emotional climate dimension to gain. There was only one surprise here for us. Negative affect related just as you would expect it to, strongly negatively with outcomes. But the surprise was that in none of those four studies did positive affect relate positively to any outcome.

When I went back and looked at Donald Medley's review,[6] I found that positive affect divided about half and half—relating to outcomes as often negatively as positively. A considerable fraction of the positive relationships comes from the same Follow Through final report of ours that I said I distrust at the beginning of this talk, so the data are, at best, mixed in Medley's review. To counterbalance that, in some reanalysis of our data, positive affect related negatively to pupil achievement gain, and strongly enough to take seriously. This may be a fluke, but at least it raises a real question about whether one of the educator's sacred cows—the belief that the classroom ought to have lots of positive affect in it—is really true.

Another aspect of the data that surprised us initially was the finding that negative affect was more destructive for the low SES child than for the high SES child. We had not expected that. We thought the low SES child would have had fairly frequent experience with negative affect in his or her environment; if you believe in adaptation theory, you would expect the child to have adapted to it so that he or she would be relatively untouched by it, while the tender middle-class child would be very easily bruised and upset by negative affect. So it surprised us that the data indicated quite the opposite.

Afterward, when we had thought it over a bit, it made more sense. We remembered the number of years in which our daughter, who is one of the tender types, came home upset afternoon after afternoon. Ruth would regularly spend an hour or two in the course of the evening trying to undo the harm that had been done that day in school. But the lower-class child who has a parent working, perhaps at two jobs in order to keep the family housed and fed, is considerably less likely to have that sort of support available to him or her.

It is more likely, then, depending on what the classroom is like, that the lower-class child either makes it or not, whereas the middle-class child may have a degree of outside support that is just not available to the low SES child. That is a guess on our part, of course; your interpretation is as good as ours. But the interaction is very clear, and it is also present in the Brophy-Evertson data,[7] so there is some degree of replication.

---

[6] Medley, D. M. *Teacher competence and teacher effectiveness. A review of process-product research.* Washington D.C.: American Association of Colleges for Teacher Education, 1977.

[7] Brophy, J. E., & Evertson, C. M. *The Texas Teacher Effectiveness Project: Presentation of non linear relationships and summary discussion* (Report No. 74-6). Austin: Research and Development Center, University of Texas, 1974.

What the data suggest, then, is that an affectively neutral classroom is a desirable situation, and that is probably somewhat different from the usual expectation. What is most clear, however, is that an absence of negative affect is critically important.

Let me move on to the management and control dimensions. I would like to break them down, in turn, into three areas that have evolved for us. I want to present our conclusions from last July. They have changed a little for us since then, but I will not pursue that unless there is some particular reason to do so.

I would like to distinguish three domains of management: behavior, the learning task, and the thinking process. Management of behavior refers to the nonsubstantive activities of the child in the classroom—freedom of movement; freedom of children to socialize, to talk to each other, to subgroup, to move around; the noise level that is permissible — things other than task activity, that is. Management of learning tasks, as a second domain, has to do with the choice and conduct of the learning task and the amount of freedom and self-determination that the child has in that domain, in contrast to the tasks being set and monitored by the teacher.

In relation to thought processes, it makes some sense, I think, that, within a task set by the teacher, children may have the opportunity to explore ideas of interest to them, or they may be boxed in to low cognitive-level activities. So freedom and support for pursuing a variety of ideas or for high cognitive-level activities are represented in this third dimension. This is fuzzy, and we are not entirely sure it ought to be separated from the second dimension, but it seems to us that it may, at least provisionally, be useful to do so.

The results for management of behavior parallel what Don Medley spoke about yesterday, but I guess I would take it just a little bit further. The results of each of the studies seem to indicate that the less freedom of physical activity the children have, the more learning takes place—the less physical activity, the more learning. There is no evidence of nonlinearity here. It may simply mean that the teachers in whose classrooms we have collected data had the wisdom not to control behavior more closely than was functional. I think research goes a long way before it betters the wisdom of skillful practitioners, and this may be such a case.

There are a couple of interesting interactions here, however. Classrooms where control of behavior is low—that is, where there is a good bit of misbehavior—show interactions with pupils who are anxious (as I mentioned earlier) and also show interactions with pupils who have high pretest standing. High pretest children are more affected by classrooms where disorder is common than low pretest children are. Again, that is the opposite of what we would have expected, but that is what the data say.

For management of learning tasks, I would be more comfortable to draw some qualifications around the notion of direct instruction that seems to be represented in Don Medley's review and, I think, even more clearly, in some of Barak Rosenshine's writing.[8] I am really not entirely clear where Rosenshine stands on this currently, but early in the game he seemed to equate direct instruction with something like the Becker-Englemann program, a closely structured contingency-management program, using programmed instruction. It is so tightly organized that a person who knows can tell you that on day 53 the students will be on this particular lesson. What our evidence suggests is that learning proceeds best if the learning task is limited to some degree, but if the students also have some degree of freedom in it. That is where the difference between this dimension and the dimension of control of behavior seems important to us. Again, the behavior control is entirely positively related. The closer the control of behavior, the more learning. But this is not true for management of learning tasks. The best learning seems to happen if the children have a degree of structure, a degree of focus, a degree of organization, but within that context some

[8]Rosenshine, B. Classroom instruction. In N. L. Gage (Ed.), *The psychology of teaching methods*. Chicago: National Society for the Study of Education, 1976.

freedom of choice, some freedom to go in their own directions. And I am not sure that this is present in Rosenshine's idea of direct instruction.

I think the distinction between these two dimensions usually is not made. In our data, teachers in open classrooms tended to free both the learning task and also the behavior of students, so that those classrooms were sometimes chaotic. It is not hard to understand why not much learning happened there, because there was a lot of distraction present. On the other hand, the contingency management classrooms control both the behavior and the learning task very closely.

In two of our sets of data, the two dimensions correlate in the high 70s in one data set and in the high 80s in the other set, despite the fact that the data were collected by two different observers on two different observation instruments, with quite different theoretical bases. This suggests that, if the typical teacher closes down one, he or she closes down both—if he or she frees one, he or she frees both. But the data say that such control or freedom is not functional. What is functional is to control the behavior, but to free, to a degree, the choice and conduct of the learning task.

There are a number of linear relationships for factors, composite measures of behavior, that reflect some freedom and some structuring. For example, one of them is a pattern in which the children are structured into seatwork. They have an assignment, but when they finish the assignment they are free to pick some task of their own or do something else, as long as they do it within the established behavioral limits. The teacher is not involved at all; he or she is working with another group.

Incidentally, this situation is typical of all of the factors that have linear relationships and appear under this heading. They represent activities on the part of the child that have set structures but do not reflect direct monitoring by the teacher. The children are exercising self-control and have some choice of direction and freedom, but they are not monitored by the teacher in any of these factors.

The implication of Rosenshine's notion of direct instruction, and of some of the conclusions that the Far West Laboratory research[9] suggests about academically engaged time, is that monitoring by the teacher is the key, but these results suggest this is not necessary if the teacher has an effective management system.

For the data on thinking, our two first grade samples both show strong negative relationships between the amount of interaction involving teachers and students that occurs at a high cognitive level and the amount of pupil learning, even on a high cognitive-level outcome measure. At first glance, this does not seem to make any sense. How can students learn high cognitive-level tasks if they are not taught at a high cognitive level? A provisional explanation, one that seems reasonable to us, is that this is not saying that no high cognitive-level interaction occurs in these classrooms, but rather that it is relative. There is simply too much going on in some classrooms, and this is nonfunctional.

Those are the first grade classrooms. There is also an interaction in the North Florida sample indicating that high cognitive-level interaction is more destructive for low SES kids than for high SES kids. I guess this is not entirely surprising.

There are two other data sets. The North Florida fifth grade sample shows no relationships between the amount of high cognitive-level interaction and gain. That is a fifth grade data set, but for the data set as a whole the students are about a grade level behind in pretest, so in a sense they are a fourth grade sample. That is a gross oversimplification, of course, but perhaps relevant. The South Carolina sample had grades 3 through 6, so the mean grade level at the beginning of the study would have been about 4.5, but they were a grade level advanced, which means they are at about grade level 5.5 in terms of pretest standing. In that sample, a factor that involved high cognitive level interaction along with the positive affect (and another thing or two) related positively to complex measures of gain. We have, then, lower grade/partially

---

[9]Fisher, C. W., Filby, N. N., Marliave, R., Cahen, L. S., Dishaw, M. M., Moore, J. E., & Berliner, D. G. *BTES: Beginning Teacher Evaluation Study* (Technical report V-1). San Francisco: Far West Laboratory for Educational Research and Development, 1978.

lower SES kids showing negative relationships between high cognitive-level interaction and complex gain; a fifth grade, but really fourth grade, sample of somewhat below average SES showing no relationship; and a sample at grade level 5.5, in a sense a higher-than-average SES/more-able-than-average group, showing positive relationships between these two measures.

If you put it all together, the suggestion is that a greater amount of high cognitive-level interaction is nonfunctional for the lower grade/lower ability students, but as you move up through the grade levels and the ability levels you pass through zero relationship and begin to get some indications of positive relationships at the higher level. This is a satisfying interpretation, even though, admittedly, a tenuous one.

There are a couple of other issues that may be worth raising here. One of the other distinctions made in a recent analysis was separating out a cluster of items that reflect high cognitive-level interaction, but are what Ruth characterized as "loose and sloppy." (You should know that most of the really perceptive interpretations are hers. She does the work and the thinking while I junket around the country talking about it.)

What does loose and sloppy mean? In the reading lesson the teacher asks a series of questions like, "Gee whiz, what do you think Jimmy would do next?" Any answer is right in response to a question like that, and there is no checking with the data by asking "What leads you to think this?" "What evidence is there?" None of that ever happens.

The other pattern ("hard-nosed") is one in which the question may go a step further: "What do you think would be a good thing for Jimmy to do?" Then we go back and look at the alternatives. We look at the consequences and make some evaluations about which of these would be a good thing to do. The difference is that we do our broad thinking, but then we go back to the evidence and relate it to the divergent ideas and evaluate them.

The earlier factors have not made that distinction. For this reason, we suspect that a good bit of the high cognitive-level interaction was loose and sloppy and that there needs to be a tie to the data for the interaction to be functional.

There is still another possibility suggested by one of the interactions in the Florida fifth grade data. Remember, no relationship was found overall between high cognitive-level interaction and gain in that sample. But there was a significant interaction there—several, in fact. One of interest at the moment is the finding that, if the teacher frequently chose the problem and also frequently engaged the students in high cognitive-level interaction, then this did not promote gain. Nor was gain associated if the teacher rarely did either of these. However, if the teacher frequently chose the problem, but did not engage the students in high cognitive-level interaction, there was more gain. And there was also more gain if the students were engaged in high cognitive-level interaction, but the teacher had less often chosen the problem. The presence of one or the other teacher activity was associated with gain; the presence of both or neither was not. We think that one way of making sense out of these findings is to suspect that, if the teacher does not pay close attention to selecting and monitoring the thinking process, students may have the opportunity to fit the problem to what they can cope with, whereas that, if the teacher is selecting and monitoring the task and also engaging the students in high cognitive-level interaction, the students may not be able to adapt the task to something they can cope with and may be forced to engaged in a task with which they cannot cope.

The pattern that emerges may be one of the teacher's being engaged in high cognitive-level interaction with three or four students in the classroom, the rest of the students sitting by —out of the interaction, unable to deal with it, not really following it, but with the process marching along, leaving them further and further behind. The situation would be rewarding to the teacher, so you can understand why it would go on. It would also be rewarding to the small number of students that are engaged in the interaction at a high cognitive level, but it is a failure for most of the classroom.

Taba and associates' work[10] may be another explanation for this untoward finding. It suggests that, unless the teacher has laid an adequate foundation in lower cognitive-level activities, the students are unable to sustain higher level thinking. Unless you have first gathered the facts, you cannot think with them. That is an oversimplified interpretation, of course. We are not ready to conclude that teachers ought not to ask high-level questions. But we do raise a red flag to the idea that all teachers ought to ask more high-level questions of all students. Ruth refers to that sort of idea as a universal prescription. I think the educational literature has universal prescriptions in it with some frequency—not stated as badly as I stated that one, of course—but the implication is that teachers ought to ask more broad questions, without any qualifications about where, or when, or with whom, or for what purpose.

We have talked about a series of linear relations, but we have also looked at nonlinear relations with some frequency in our data. This is a way of testing what Ruth calls the "more is better" fallacy. Any time you calculate a linear correlation, you are assuming that if some is good, more is better—without limit. Again, teachers probably have the wisdom to protect researchers from that error, but not always. Sometimes they have been pushed into ways that are, we suspect, erroneous.

Let's look at five figures based on data published in my paper for the conference *UPDATE: The First Ten Years of Life*.[11]

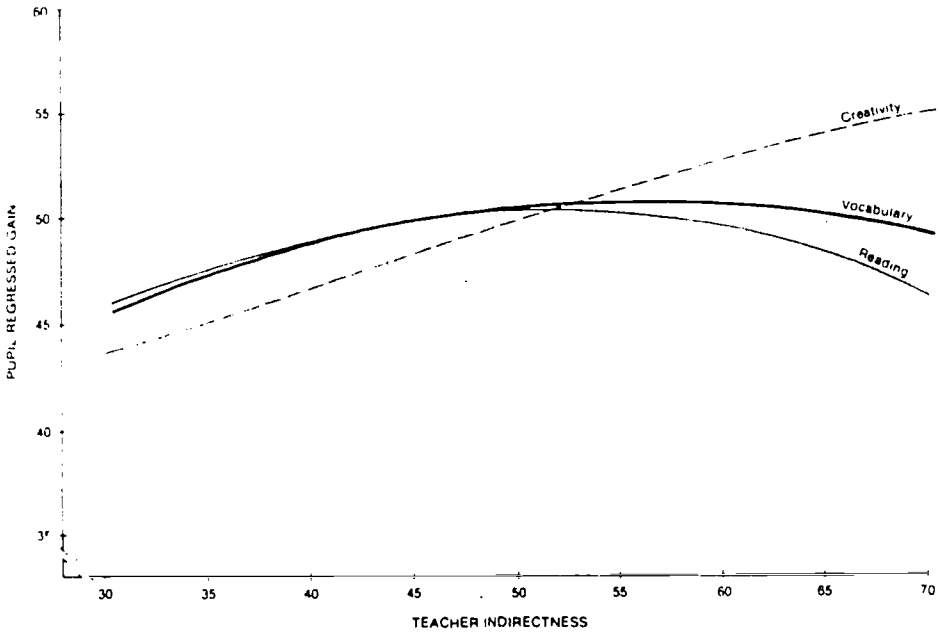

Figure 1. Teacher indirectness related to pupil growth.

The first one comes from the South Carolina study, the study with upper grade/upper ability students. (See Figure 1.) The measure on the base line is teacher indirectness from the Flanders system, but it is a complex measure that has a variety of kinds of indirectness represented in it. The measure on the vertical is pupil gain. We have three plots: one for vocabulary, one for reading from the Iowa Test of Basic Skills—both relatively high cognitive-level outcome measures of skills—and the third, essentially the straight line, is a measure of gain in creativity from Torrance's battery.

[10] Taba, H., Levine, S., & Elzey, F. F. *Thinking in elementary school children* (Coop. Res. Proj. No. 1574. OE. U.S. Department of H.E.W.) San Francisco: San Francisco State College, 1964.

[11] Soar, R. S. *Group learning environments for the early school years.* In *UPDATE: The first ten years of life* (Proceedings from the Conference Celebrating the Tenth Anniversary of the Institute for Development of Human Resources, College of Education, University of Florida, Gainesville, March 29-31, 1976). Gainesville: Division of Continuing Education, University of Florida, 1976.

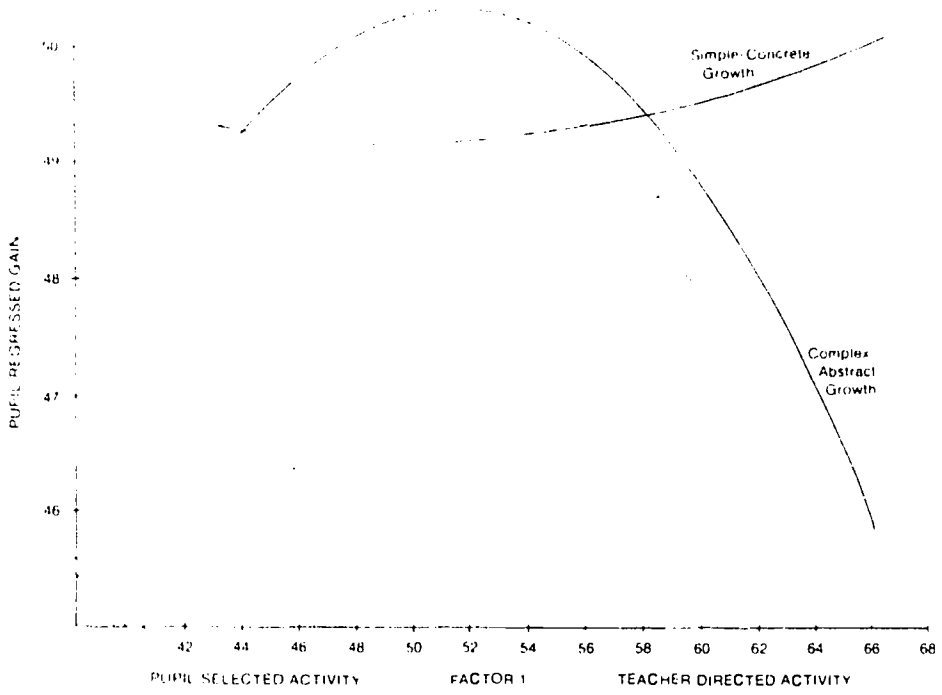

Figure 2. Relation between teacher practices observation record factor 1 and pupil growth

The basic message is that, if you line up classrooms with respect to the degree of indirectness the teacher uses, then as teacher indirectness increases, gain increases through the lower part of the range; but beyond some point, more is not better for two of the measures. For creativity, increasing indirectness is still useful to the most extreme classroom. I am not sure this would hold up, but there it is. Figure 2 has a measure on the base line taken from live observation in the classroom with the Teacher Practices Observation Record. (This is an instrument that looks at the classroom through the eyes of Dewey's experimentalism.) The factor is one that reflects the teacher's choosing and monitoring the activity in contrast to the pupils' having a good bit of choice in this process (a measure of management of learning tasks). Another relevant observation we probably can make is that this factor accounts for more differences between classrooms than any other. In fact, if there is any single dimension that differentiates classrooms more strongly than any other this is it: The extent to which the teacher is front-and-center managing and directing the learning process, in contrast to turning the students loose to work on their own.

The left-hand end of the dimension represents pupil freedom of choice; the right-hand end represents teacher control, teacher limit-setting. That figure alone raises some questions for me about the usefulness of generalizing the notion of direct instruction. The curve that angles upward is a measure of simple-concrete learning, mostly memory tasks. The one labeled "complex" requires information processing on the part of the students. One of the questions, for example, is "What does a teacher do?" The students are asked this at the end of the year, and the answers reflect the students' ability to abstract out of their experiences in the classroom those that are central to the business of teaching. Close teacher control is associated with sharp decreases in that sort of complex response.

Figure 3 is only worth talking about because it is coded from audiotape on the Reciprocal Category System, an extension of the Flanders system. The coders had never seen these classrooms, did not know what program they were in, and did not know anything about them except what they heard through earphones. This dimension is one that runs from pupil initiation at the left-hand end of the scale--the freedom the student has to speak up in the course of the interaction—to drill at the right-hand end of the scale, in which the student is boxed in completely.

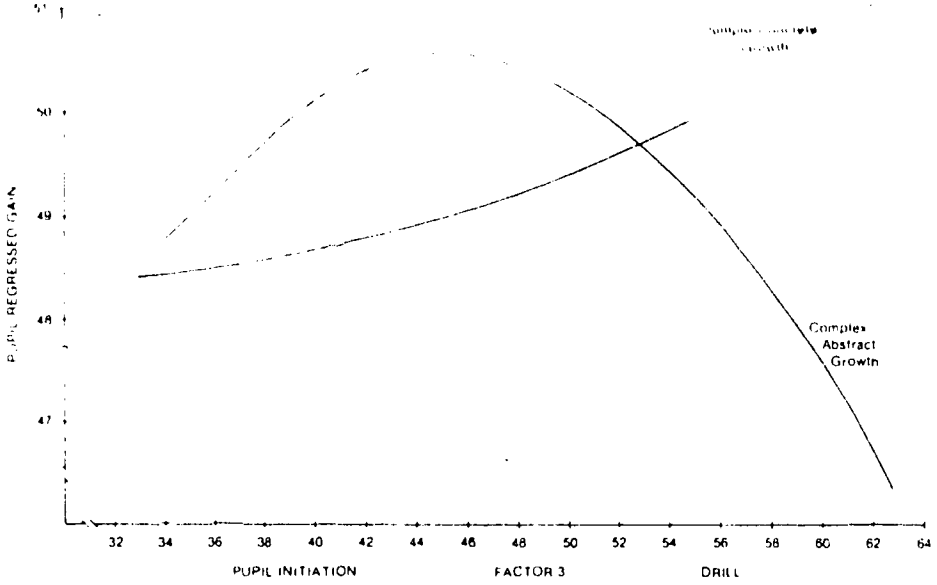

Figure 3. Relation between reciprocal category system factor 3 and pupil growth.

The curves from those two data sets were virtually identical. It is supportive to us that people who were never in the classroom produced results that were so similar to those produced by classroom observers who knew somewhat more about what was going on. Notice that the nonlinear relationship only holds for the complex measure in these cases.

Figure 4 comes from the Florida fifth grade study. The only significant difference in that figure is between spelling, on the one hand, and reading and vocabulary, on the other; spelling was a low cognitive-level measure, and reading and vocabulary were higher cognitive-level measures.

Figure 5 is from the Florida first grade study. The measure is the Metropolitan Readiness Test, and the curve that is high at the left is Numbers. It involves the child in making comparisons (greater than, less than) in counting and in solving word problems. The other curve, the one that peaks toward the right, is Word Meaning, but all the words in that vocabulary measure are nouns and are represented with pictures.

I do not think it is much of an overextension to say you could teach the correct responses to a pigeon if you used proper conditioning procedures. That may be a little too much, but not really. On the other hand, vocabulary in the Iowa Test of Basic Skills is made up of adjectives, adverbs, and a few conjunctions, with no nouns. They are all words that represent relationships between things for which the child could learn the meaning only by abstracting out of his own experience with language. So it seems to us that the vocabulary measure from the Iowa test is at a relatively high cognitive level. Obviously this is interpretive, and you may differ with it.

The baseline measure in that fifth figure is the amount of interaction that takes place at the level of translation, the next-to-lowest cognitive level. What those figures suggest is that, in the management of learning tasks, there is a balance between the teacher's setting the task and monitoring it, on the one hand, and some degree of pupil freedom of option, freedom of self-direction—"wiggle room" is Ruth's term for it—on the other. The greatest learning occurs when an appropriate balance is struck between control and freedom.

The minor theme against that major one is the extent to which these curves differ. With almost complete consistency they differ in the direction of the higher cognitive-level outcome measure growing best under a greater degree of freedom for the students. The higher the cognitive level of the learning task, the greater the freedom that is functional for the students.

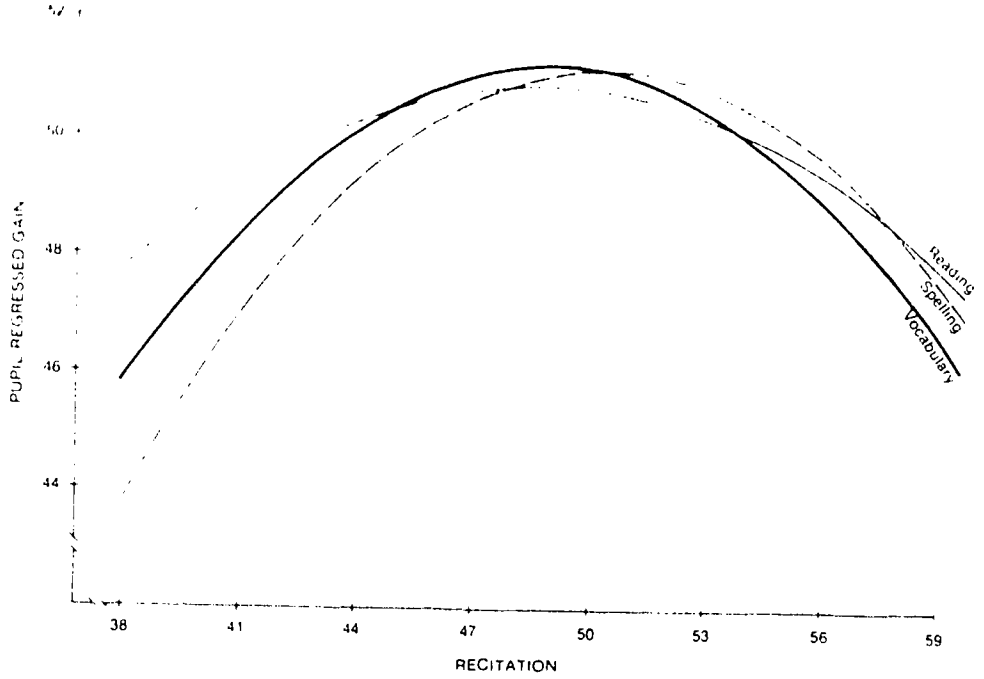

Figure 4 Relations between recitation and three achievement gain measures

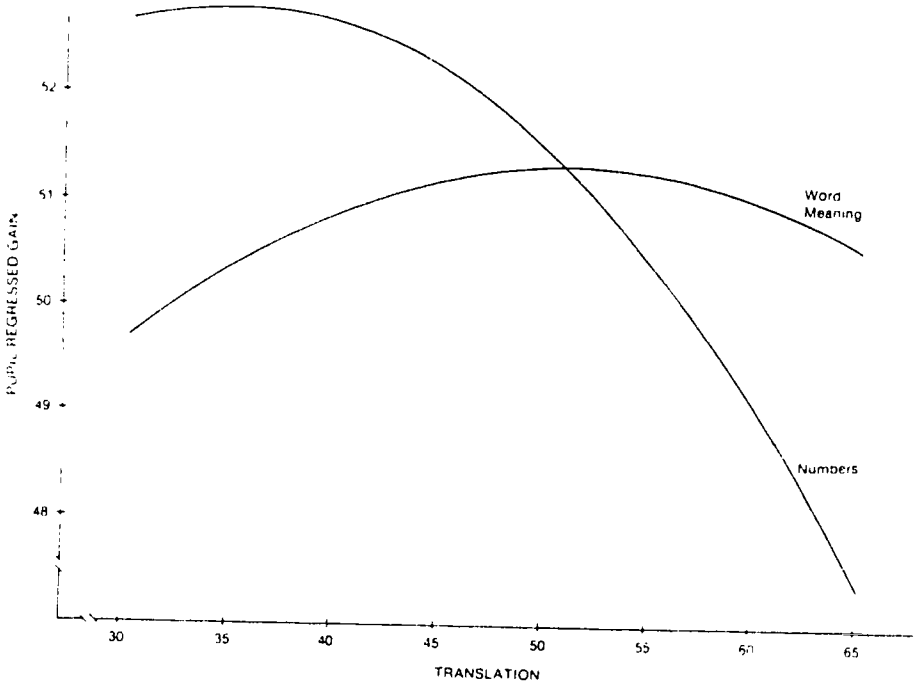

Figure 5. Pupil achievement gain in relation to teacher-pupil translation.

Again, that is a minor theme and I think it probably would be easy to overstate the extent to which the data support it, but there are suggestions of that running through all four of the data sets. There is also some support for it in one of our analyses of the second year's Follow Through data. We factored the items from the achievement battery and developed measures representing three different levels of cognitive complexity. In a comparison of programs, students in the Becker-Engelmann program were at the top or next to the top of the list in the simple memory

measures and in the discrete, atomistic, academic skills measures, such as phonics. But on the measures that required complex information processing, the Becker-Engelmann students were at the bottom of the list, or near the bottom of the list.

Jerry Cline spoke to us at Florida a couple of years ago, while he was director of the AIR project to reanalyze the Follow Through data. Among the results he presented was the finding that the Follow Through students in the Becker-Engelmann program were well ahead of non Follow Through students the first two years, but that in the third year of the program they fell behind the non Follow Through students. When he and his associates went back and analyzed the test results by item, they found that, by the third year of the program, the tests used to measure gain had altered their emphases from relatively low cognitive level things to items that reflected information processing, and once the items began to reflect manipulation of concepts, the students in that progam just simply stopped growing.

I guess Ruth and I conclude (oversimplifying a little) that direct instruction has limited usefulness. Imagine, if you will, a four-cell table in which the rows represent the socioeconomic status of the students: high economic status, low economic status. The columns are defined by the cognitive level of the outcome measure: high level outcomes, low level outcomes. In that four cell table, direct instruction is most useful for the cell that reflects low socioeconomic students and low cognitive outcome measures. It is exceedingly effective for that combination. But it seems to us that it is less effective as one moves out of that cell into any of the other three.

That statement oversimplifies the situation somewhat, but I think it conveys the sense of where we are. But, again, this is an interpretation. This clearly goes beyond the numbers, you ought either to accept it or not, as it seems reasonable to you.

In summary then, it seems to us to be useful to think of four major domains of the classroom climate for learning: emotional climate, freedom of pupil behavior, choice of learning task, and freedom of thinking processes. Our results suggest that an essentially neutral emotional climate is functional, but even more strongly they suggest that limiting the expression of negative affect is critical. Close limits to the behavior that is acceptable also seem useful.

But in the areas of choice and conduct of the learning task and of freedom of the thought processes, it seems to us that a balance between pupil freedom and teacher control is most functional for pupil learning -- a degree of task structure or focus, but within that structure, a degree of freedom for pupils to choose their own directions.

It seems to us that one of the difficulties of conceptualizing teaching effectiveness may have been the failure to distinguish the domains in which close limits to behavior are functional from those in which a balance of limits and freedom are functional. At the same time, the distinction seems to be accepted by the teachers with whom we have worked as reasonable and meaningful. We hope it may be a useful one for the teachers involved in this project.

Using Feedback to Change Teacher Behavior

Frederick J. McDonald
Educational Testing Service

One of the problems this Conference has confronted is the relation between research results and their use in the schools. I suggest to you that solving this problem requires translating results into practical activities. The next step is to do the developmental research to evaluate the effectiveness of these practical activities.

Research data typically describe teacher behaviors related to pupil outcomes. They are only a part of all the actions in which teachers engage. But teaching is conducted by human beings who are acting and talking and moving all the time, and we need to have a good picture of how those "effective performances" fit into a pattern or style of teaching. The usefulness of research results for practice requires a description of teaching styles and actions that embody the teaching performances found to be effective.

In my opinion, it is very difficult to go from the kinds of statements derived about effective teaching in the research to prescriptive rules for teaching. Something much more complicated is needed, and the researchers are usually not in a position to do that practical design work, or, unfortunately, are not always interested in doing it.

When the R&D centers and laboratories were created, one of the ideas was that the research centers would do the basic research and then the regional laboratories would take on the work of development. That concept worked very well in two examples. One was the development of IPI in Philadelphia and Pittsburgh, resulting in an innovative project that had been effectively managed from research to practical application. The other example occurred on the West Coast. At Stanford, we did the basic research related to microteaching, which was then taken over by the Far West

---

This version of Dr. McDonald's speech summarizes his presentation to the Conference, with accompanying tables and figures drawn from his previously published research.

Laboratory and developed into practical and effective minicourses for teachers. I think we are now at a point where that process needs to be engendered in the domain of teaching behavior.

It seems to me that one of the productive things that a regional laboratory can do with the state departments of education is to begin to build developmental applications of the findings that have come out of research data. I will talk about some of the difficulties in trying to go from the research data to application.

Let me begin by describing very briefly some data from Phase II of the Beginning Teacher Evaluation Study in California. The California study was conducted in a variety of schools, forty-five schools in eight school districts; the teachers and pupils comprised the range of socioeconomic groups that you find in California schools.

The results in Phase II were very similar to the results that Donald Medley described today. Three of us who were involved in the study—Jane Stallings, Jere Brophy, and I—have repeatedly checked notes, and in general the same conclusions have come out of our work. These conclusions also correspond with Robert Soar's results, except that he keeps doing curvilinear regressions and making life more complicated for all of us. But in general our work is comprehensive in terms of the variety of teachers involved, the variety of places where it has been done, and the general similarity of the results. These results cannot be dismissed on the usual methodological grounds that there were not enough teachers involved and that there was not enough variety in the kinds of schools and children.

The Beginning Teacher Evaluation Study was done for the California Commission for Teacher Preparation and Licensing. It was the second phase (the first being a design and planning stage) of a long-term study to gather data that would enable the Commission to write policies on teacher preparation and licensing.

In Phase II the Commission asked us to design a study that would tell them whether teaching performance (or actions), teacher aptitude, or teacher knowledge makes the largest difference in pupil achievement. We wanted to know the relative influence of those various categories of factors on pupil achievement. Information of this kind would enable the Commission to make decisions about policies on the admission of students to teacher-training programs, about the substance or content of those programs, and about performance training.

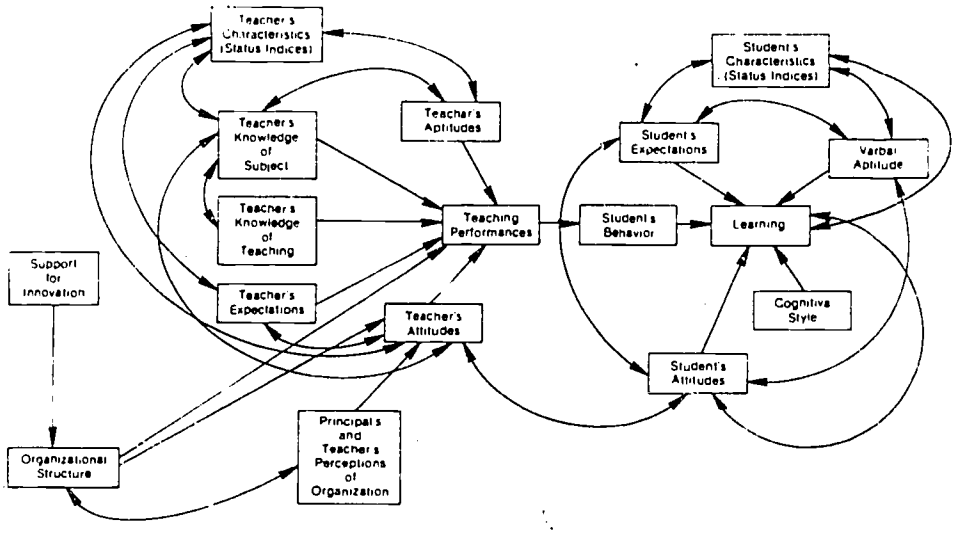

Figure 1. A structural model of the domain of variables influencing teaching performance and children's learning.

Figure 1 presents a model of the study as it was designed. The three boxes in the center are Teaching Performance, Students' Behavior, and Learning. The Commission decided what areas of learning were to be the criterion variables: reading,

decoding skills, comprehension skills (literal and inferential comprehension), attitudes toward reading, and applications of reading where application means ability to read materials other than what is typically found in school materials. Similarly, in mathematics the categories were concepts, computations, applications, and attitudes.

We built a comprehensive test battery. In reading, we wrote a decoding test that is quite different in many respects from what you find in the typical standardized test, and we designed the comprehension part of the test to measure both inferential and literal comprehension.

The tests were administered twice, in the fall and spring, to the students of some one hundred teachers. That testing provided one basic set of data. The other basic set of data was derived from observations of the teachers in the classrooms. We related what we observed to differences in pupil learning.

In addition, we looked at the organizational climate, support for innovation, and organizational structure in the school. Every principal was interviewed for an hour. We gathered data on teachers' attitudes toward what they were teaching, their expectations for pupil performance, what they knew about the subjects they were teaching, and their backgrounds. The teachers also took a battery of tests measuring a variety of aptitudes.

Let me say, without going into details, that in all of these organizational and attitudinal factors the two categories most related to differences in teaching performance were the teachers' attitudes, defined in terms of their aspirations about their work and satisfaction in it, and the aptitude measures. The strongest set of relations between performance and any of these factors was the set with the aptitude measures.

The pupils were observed during class at the same time the teachers were. We also gathered data about students' expectations, verbal aptitudes, cognitive styles, attitudes toward the subjects they were being taught, and their backgrounds.

We were primarily concerned with identifying the teachers who were most effective and least effective, defined in terms of differences in pupil gains on the measures of pupil learning. For each teacher, we correlated the scores of his or her pupils at the beginning of the year and at the end. A correlation tells if the pupils are rank-ordered in the same way at both times. Across all the classes, fall and spring scores correlated anywhere from .80 to .90. This strong relation means that the fall score is the best predictor of spring scores. It also means that only about 30 to 10 percent of the variance may be accounted for by other factors.

The interesting fact is that the correlations were quite dissimilar from teacher to teacher. When a correlation coefficient is very high, .90 or better, it means that the pupils were rank-ordered about the same in spring as they had been in the fall, but, of course, they probably had higher scores. For some teachers the correlation was much lower, .50 to .70, which means that the pupils were not ordered in the same way. Perhaps some who had been low made large gains, or some who had been high were achieving less well in the spring. We were curious about these differences. They suggested that there might be differences in the pattern of gains in the different classes. We used a simple method to study these patterns.

The correlation between the two sets of scores may be portrayed graphically, as in Figure 2. The points in the diagram represent the scores of pupils. Pupil 1, for example, had scores of 20 in the fall and 30 in the spring. Pupil 2 had scores of 50 in the fall and 60 in the spring. The other points represent the scores of other pupils in the class.

I have drawn a line through this array. There is a way of calculating the equation of this line from the scores. This line, called a regression line, provides basic information on the pattern of the scores.

We generated these lines for each teacher. If all the lines had been parallel, then the only difference among them would be their respective vertical heights, and the differences would indicate differences in the amount of the gains. But we knew that they probably would not be parallel.

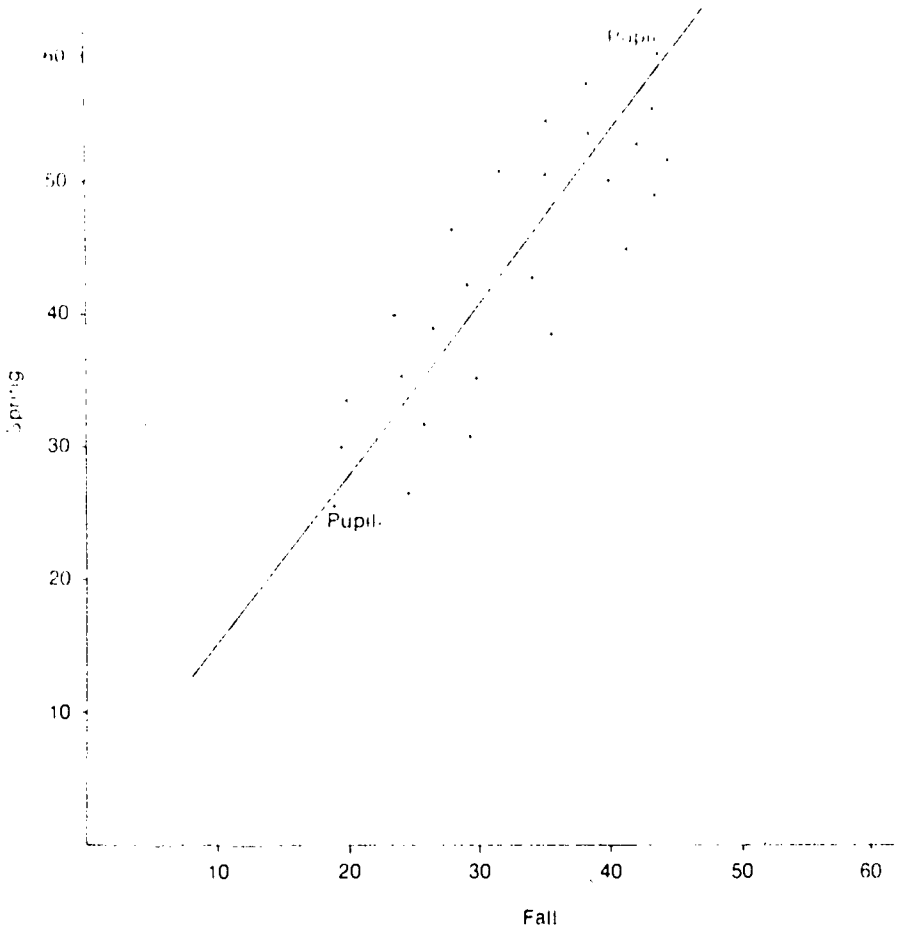

Figure 2. A graphic portrayal of the correlation between fall and spring scores in reading in one class.

Refer to Figure 3. If there is a perfect correlation between fall and spring scores, the regression line will be sloped at 45°. Different correlations produce lines with different slopes. We found three kinds of lines. We found lines that were steeply sloped, lines that had shallow slopes, and lines sloped at about 45° but elevated above the origin of the axes.

What is the relation of the slope of the line to pupil gain? When the slope is shallow, most pupils in the class are making some gain. In classes with a steeply sloped line, the gains are occurring primarily in the upper half of the class. When the slope is 45° but elevated, all students gain. These regression lines are describing those differences in the patterns of gains across the different classes. Given where the class started and where it ended, the regression line describes the pattern of change.

There is another way of demonstrating or portraying this same information. For each line there is a number representing its slope and a number representing where it crosses the Y axis. We took those two numbers for each line and plotted them for each teacher (see Figure 4). Along the horizontal axis are numbers for the slopes: to the left are numbers less than 1, representing shallow slopes; and to the right are numbers above 1, representing steep slopes. Along the vertical axis are numbers representing the intercepts.

You will notice that the differences among the teachers fall into a pattern. The same pattern was found in second grade reading, second grade mathematics, fifth grade reading, and fifth grade mathematics. That pattern, we think, is one you are likely to find whenever you do this type of analysis.

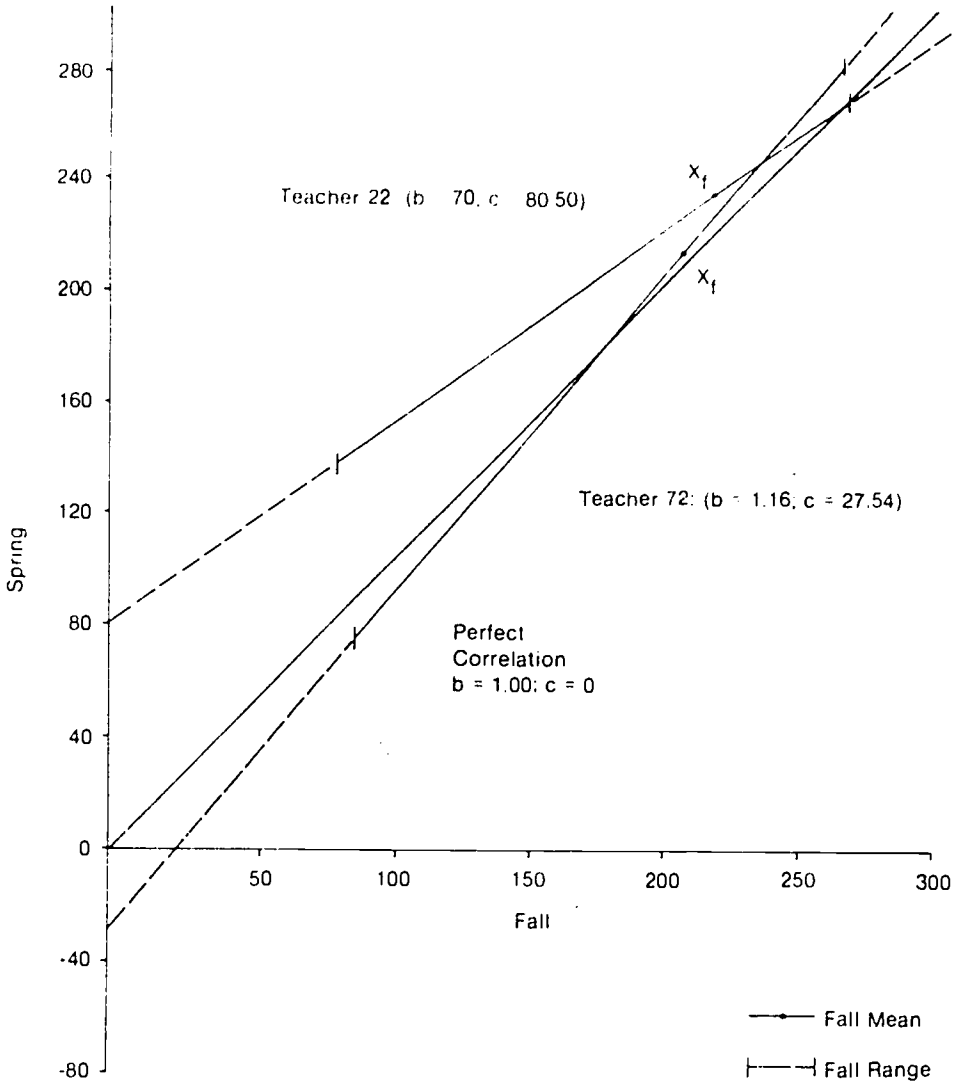

Figure 3. Regression lines for two groups, one in top group (Teacher 22), one in bottom group (Teacher 72).

The differences in the slopes are significant here. These numbers, as portrayed in Figure 4, represent differences in effectiveness. We would predict that a teacher with a higher slope would have students with less gain. A teacher, however, is not necessarily ineffective because the scores of his or her pupils produce a steeply sloped regression line. But there are obviously teachers who are more effective because they are helping all the children in their classes to make some improvement.

We next assigned a number (+1) to each of the ten teachers in the upper left in Figure 4 and another number (-1) to the bottom ten (lower right), and assigned a zero to the teachers in the middle. These numbers represented each teacher's location in one of these three groups. Then we took all the teacher performance data and said: Do differences in teaching performance predict the group in which a teacher is located?

It turned out that the multiple regression coefficient (an estimate of the predictive power) was extraordinarily high, a number that nobody usually sees in this kind of research. It meant that, outside of error of measurement, the differences in teaching performance accounted totally for the differences in location. These

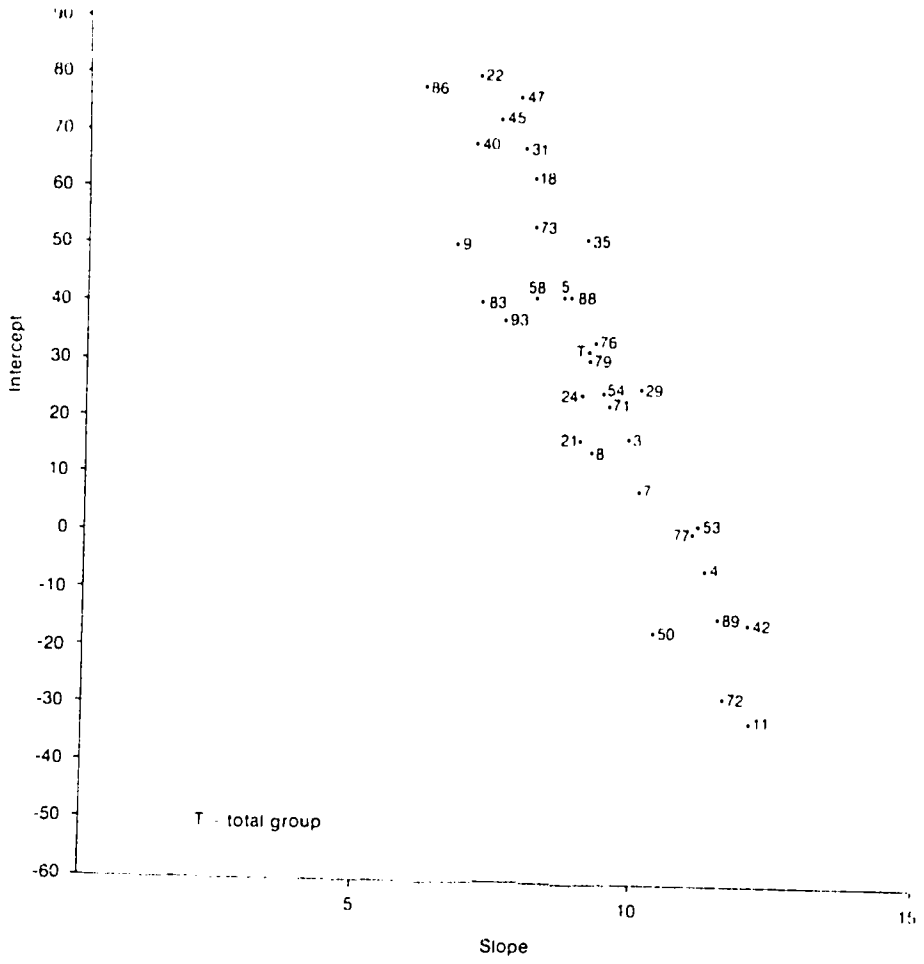

Figure 4. Slope (b) plotted against intercept (c) for each teacher (referred to by an identifying number) in grade 2 reading (N = 33).

differences in the pattern of gains in a classroom are accounted for by the way the teachers teach.

What were the teaching performances that distinguished between the two extreme groups? The pattern characteristic of the most effective teachers was very much the one that Dr. Medley described today, the pattern of direct instruction in which the teacher works with each child as frequently as possible, approximates the tutorial model as closely as possible, increases productive on-task time, and uses a relatively mild form of social control. Effective teachers used practically no whole-group instruction. In fact, in our study, whole-group instruction was usually a negative predictor of learning for second grade classes, though not for fifth grade classes.

The effective pattern of interaction with students was one in which the interaction was generally sustained and more substantively rich than the alternative pattern, which is essentially the spot type of questioning in which the teacher moves quickly from one pupil to the next.

One of the characteristics of those teachers in the lower group—who are not poor teachers, by the way—is that they spent more time organizing instruction than instructing. They also did more whole-class teaching and had less productive on-task behavior.

Table 1
Mean Difference, Standard Deviation of the Mean Difference, Slope and Intercept for the Top and Bottom Ten Classes Grade 2 Reading

Top 10 Classes

Teacher Number	X	df	Mean Diff.	SD_{s-f}	b	c
22	212.84	18	16.76	24.55	70	80 50
47	183.07	26	34.36	30.79	.77	77.06
45	181.38	21	26.44	27.56	.74	72.97
73	168.13	20	20.30	43.07	.80	53.72
40	159.38	17	20.33	36.75	.69	69.19
86	145.89	16	20.72	37.72	.60	79.25
31	142.89	23	35.59	31.62	.78	67.64
18	116.48	19	39.64	43.77	.80	62.55
35	116.15	10	41.14	33.20	.90	52.35
09	99.34	22	17.15	36.55	.66	51.37

Bottom 10 Classes

Teacher Number	X	df	Mean Diff.	SD_{s-f}	b	c
72	203.22	21	5.61	28.31	1.16	-27.54
77	194.21	18	20.63	26.62	1.10	1.35
50	174.58	11	1.46	35.55	1.03	-16.15
07	163.35	18	7.92	22.58	1.00	7.99
04	162.51	11	15.42	16.93	1.12	- 5.99
89	149.04	18	8.33	26.31	1.15	-14.49
11	141.96	10	2.13	35.80	1.21	-31.63
42	140.78	17	13.60	36.82	1.20	-14.70
08	129.76	22	2.81	31.06	.91	14.82
53	125.96	18	16.15	35.57	1.11	2.46

df = degrees of freedom
SD_{s-f} = standard deviation of mean difference
b = regression line
c = intercept

Let me quickly give you some idea of what these differences mean in terms of pupil achievement. Refer to Table 1. You may be thinking that all classes that started out high in achievement are in the top ten. Look at the line for Teacher 09. The reading test was a 300-item test. Teacher 09's class began with about a third of the items as their mean score. That is a low performance. The standard deviation of that mean was very large, 55, which means—subtracting 55 from 99 equals 44 - - that a substantial number of the children had a score of 44 or less. In fact, this was a class where the children were probably by and large illiterate at the beginning of the second grade. The mean gain for that class was 17 points.

If you read in the third column of Table 1, you begin to see how large the mean differences in performance were for these top classes. No number is smaller than 17 and the highest number is 41. Forty-one points is a substantial gain. In contrast, in the lower ten you see mean gains of 5, 20, 8, 15, 7, 2, 1, 16.

To make the contrast a little bit sharper I paired a class from the top group with a class from the bottom group, using their fall scores to match pairs (see Table 2). Note the first pair where initial scores are 212 and 203 (Teacher 22 and Teacher 27). These are high-scoring classes. A score over 200 means that on the average over two-thirds of the items were answered correctly. The scores were spread out about the same amount. But one class gained 16 points on the average and the other class gained 5.

Table 2
Top and Bottom Classes Paired by Order of Magnitude of Fall Means; Mean Differences and Standard Deviation of Fall Mean for Each Class
Grade 2 Reading

Teacher Number	$\bar{X}_f$	SD_f	Mean Diff.
1. 22 (T)	212.84	48.71	16.76
72 (B)	203.22	42.58	5.61
2. 47 (T)	183.07	49.74	34.36
77 (B)	194.21	33.94	20.63
3. 45 (T)	181.38	46.97	26.44
50 (B)	174.58	60.28	1.46
4. 73 (T)	168.13	53.40	20.30
7 (B)	163.35	55.21	7.92
5. 40 (T)	159.38	40.23	20.33
4 (B)	162.51	53.47	15.42
6. 86 (T)	145.89	42.86	20.72
89 (B)	149.04	42.08	8.33
7. 31 (T)	142.89	58.51	35.59
11 (B)	141.96	26.60	2.13
8. 18 (T)	116.48	55.21	39.64
42 (B)	140.78	45.53	13.60
9. 35 (T)	116.15	37.24	41.14
8 (B)	129.76	39.08	2.81
10. 9 (T)	99.34	55.21	17.15
53 (B)	125.96	42.01	16.15

T = top ten
B = bottom ten

Note the seventh pair (Teacher 31 and Teacher 11). They both start about the same point, 142 and 141. Their standard deviations are 58 and 26 respectively. The scores for Teacher 31's class are spread out and the scores for Teacher 11's class are compact. Look at the differences in the mean gains: 35 points versus 2.

As I said, none of these differences would be interesting if they were unrelated to the differences in teaching performances. Certainly for second grade reading and second grade math this relation was substantial.

A question of general interest always is: Were the teachers who were in the top group in reading also in the top group in mathematics? The answer is no. Only one teacher was in the bottom group in both reading and mathematics and four were in the top in both analyses. Teacher 22, who is at the top of the diagram in reading (refer back to Figure 4), is at the bottom in mathematics. Teachers are not necessarily equally effective across these subjects, which means that inservice training has to be given subject by subject, or area by area.

We used these data and, working with the Trenton State faculty, designed an inservice program for the teachers in a school in Trenton. The basic faculty of 16 teachers (this group did not include aides or Title I teachers) assembled for two hours of inservice instruction every Monday afternoon. The instruction had three compo-

nents. The first component was a simulation or presentation or modeling of a desired behavior or skill. Then the teachers practiced this teaching performance in their classrooms for two weeks after that, during which they were videotaped; on successive Mondays they discussed their videotapes or other activities with members of the Trenton State faculty. The program was a model-demonstration-inform-practice-feedback type of training program. It was modularized in the sense that each unit of this kind was devoted to a specific teaching skill. All the skills were related to classroom management and reading instruction.

We observed the teachers before the training began, during the training, and always for a short period after a module was finished. We have data on where they began, what they looked like during the training, and what they looked like at the end of the training.

At the same time we measured pupil performance in reading, in decoding and comprehension, three times in the year: In the fall, somewhere in the middle of the year, and in the spring. We were looking for two consequences: (1) Did the training have any effect on classroom teaching performance? (2) Was there any relation between the way the teachers taught and pupil gains in reading?

The school was an inner-city school. It had a very wide range of scores on the statewide assessment battery and on the City of Trenton tests. We did find that the decoding test that we had used in California was too easy in the earlier grades, so we had to make a more difficult form. That told us the time spent on teaching decoding in this school was probably effective.

The observational system that we used gathered data on a wide variety of aspects of teaching—it was the RAMOS system developed by Bob and Kate Calfee at Stanford. Our observations took place every day except Friday, for obvious reasons, and we observed these same teachers over a two-year period. We observed both the beginning and the end of the reading period. We varied the day on which we observed. We have what amounts to an almost continuous record of teaching performances during the second half of the year.

The observations were made in the second half of the year for two reasons. During the first half of the year the inservice program was being developed. And, by the final half of the year the school had been reorganized and teachers were familiar with pupils.

Refer to Table 3. Across the top are numbers that represent the modules. There were five modules during the first year. When you read down a column you are reading numbers representing *percent of time* spent in this particular activity by the end of a certain module. Find the label, Role: A (Assess/Diagnose). under *Pre* and opposite *B* is the number 5 percent. These are baseline data; teachers were observed in the role, Assess/Diagnose, 5 percent of the time before training. Reading to the right, you see that at the end of the first module this number went up to 8 percent, at the end of the second module to 14 percent, and by the end of the fifth module it was up to 20 percent.

The first three modules were not designed to train the teachers on this role. The training was to help them keep children more on task. The last two modules, however, were on assessing and diagnosing, and it is after these modules that the largest changes occur.

This particular change, because of the nature of the number of teachers involved, is not statistically significant; but, given the small sample size, we paid attention to consistent trends. We were not using a control group. We were using the "subject" as his or her own control; that is, we analyzed where a teacher began and how much he or she changed. When the change is not statistically significant, but shows a trend, it means that some, but not all, teachers changed.

You may wonder why the percentages for Assess/Diagnose increased across the first three modules. This change appears to be a side effect of this training, which was on classroom management directed to improving the pupils' on-task, productive behavior. Perhaps that training led to more assessing and diagnosing as a way of monitoring pupils' task performance.

Table 3
The Effects of the Training Modules on Teaching Performance: Frequency by Module: Phase I
(Decimal numbers represent percentages of time. Only selected codes appear in these tables.)

	Pre	Post 1	2	3	4	5
Role A (Assess, Diagnose)						
B	05	08	14	09	16	20
B•1	05		13	11	15	21
B•1•2	08			11	15	21
B•1•2•3	10				14	20
B•1•2•3•4	11					20
Role D (Discipline)						
B	02	01	03	00	00	00
B•1	01		03	00	00	00
B•1•2	01			00	00	00
B•1•2•3	02				00	00
B•1•2•3•4	01					00
Role N (Instruction)						
B	55	45	24	46	36	32
B•1	46		25*	45	36	30
B•1•2	40			45	36	30
B•1•2•3	41				36	32
B•1•2•3•4	40					32
Role F (Facilitate)						
B	16	24	24	25	22	25
B•1	24		25	24	24	27
B•1•2	25			24	24	27
B•1•2•3	23				23	27
B•1•2•3•4	23					27
Role M (Manage)						
B	17	21	24	13	11	12
B•1	20		23	13	11*	12*
B•1•2	20			13*	11*	12*
B•1•2•3	18				11*	12*
B•1•2•3•4	16					12
Role T•S (Independent and Supervise Staff)						
B	00	00	02	03	03	00
B•1			02	02	04	00
B•1•2	01			02	04	00
B•1•2•3	01			02	04	00
B•1•2•3•4	02					00

	Pre	Post 1	2	3	4	5
Role A•N (Assess Diagnose and Instruction)						
B	59	53	38*	55	51	52
B•1	51		38	56	50	53
B•1•2	47			56	50	51
B•1•2•3	50				51	52
B•1•2•3•4	51					52
Role A•N•F (Assess Diagnose and Instruction and Facilitate)						
B	78	76	62	80	74	78
B•1	75		63	80	74	78
B•1•2	72			80*	74	78
B•1•2•3	74				74	79
B•1•2•3•4	74					79
Mobility S (Stationary)						
B	62	70	52	48	71	75
B•1	58		53	49	73	71*
B•1•2	58			49	73	71*
B•1•2•3	56				74	73*
B•1•2•3•4	60					73*
Mobility L•M (Moving)						
B	38	30	45	51	29	24
B•1	41		44	50	27	29*
B•1•2	41			50	27	28
B•1•2•3	43				26	27*
B•1•2•3•4	39					27
Mobility Number of Moves (Number of times teacher changes groups)						
B	23	16	24	18	12	18
B•1	20		23	18	12	19
B•1•2	22			18	12	19
B•1•2•3	21				12*	18
B•1•2•3•4	19					18

*Statistically significant at the .05 level
B = baseline
B•1 = after first module, etc

Let me describe one result of pupils' attending behavior to you. We did track, by scanning back and forth, whether children were on-task or off-task, whether they were engaged productively or unproductively in that task.

We sorted the on-task behavior in terms of whether the child was close to the teacher, at a middle distance, or on the periphery. There is an invariant pattern across these 16 teachers (and I would be very surprised not to find it everywhere) of a high on-task behavior close to the teacher, less in the middle, and much off-task behavior on the periphery.

We also found that there are individual differences across teachers. For some, the rates of off-task behavior were high across all three levels; for others, low. My research assistant said, "I bet when there is on-task behavior in the periphery, there is somebody out there." I told her to sort the data according to whether there was or was not another person in the groups remote from the teacher. No differences were found. If there had been 50 percent off-task behavior when the teacher was alone, there was 50 percent off-task behavior when she or he had help from an aide or student teacher.

The teaching performance required is, in part, a vigilance or scanning behavior. Whatever ways the teachers scan, this aspect of management seems to be stable. We plotted off-task behavior over days and it is consistent across days. The problem to solve is to find ways of improving teachers' scanning behavior.

Teachers whose classes were more on task made greater gains. But this variable only accounted for about 4 to 10 percent of the variance in final scores. The teaching performances related to on-task, productive behavior on which training was provided are the simplest forms of management skills; they are primarily monitoring performances. Other factors, such as arousing interest, may have more powerful effects on sustaining attention. In making this comment, I am pointing to a conclusion: The more specific the skill on which training is given, the less likely there is to be a large pupil effect.

We now turn to another effect of training as it was given in Trenton. Refer to Role N in Table 3. Look at what happened to instruction and you begin to see one of the real problems in training. Instruction occurred about 55 percent of the time before training. By the end of the fifth module it was down to 32 percent of the time.

These different roles are ones a teacher cannot take simultaneously (because of the way they are defined). You are assessing and diagnosing or you are instructing. If one goes up, the other inevitably goes down.

Facilitate (Role F in Table 3) went up 8 percent and then stayed constant across the entire time. Facilitating means the teacher goes around and works with the child, teaching the child something, monitoring, giving corrective feedback, and so on.

We tried to encourage the teachers to move around so that they could do a better job of monitoring students' on-task performance. The first three modules were designed to make them aware of students' off-task behavior, and to organize their classrooms so they could move around more easily and monitor each student's activity and give more feedback or help, as required. Apparently these teachers were giving more help because facilitating did go up, but they were not moving around very much. (See Mobility S and Mobility L+M in Table 3.)

There was no formal training on comprehension skills during the first year. There was, however, an increase in the use of comprehension skills (refer to Table 3) and it occurred after the fifth module. This change may have been a consequence of doing more assessing and diagnosing for comprehension.

Interpreting skills did not go up at all; they stayed the same. Teacher questions (QT) increased after the third module and especially after the fifth module.

We see two desirable changes in these data—increases in assessing and diagnosing and teachers' questions. But these two changes were accompanied by a decrease in instructional time.

What you also see here is a picture of changes that occur as a function of the specific training and then drop out. If you use the modularized-type program, geared to training on specific skills, one of the effects will be that teachers will learn to use the next skill, again at the expense of something else. The most difficult problem is to try to figure out how to modify the *total style* without always losing something in the process.

What was the effect of this training on pupil learning? Refer to Table 4. In the vertical column on the left are the teaching performance variables. We asked three kinds of questions. Was the mean level (**M**) in teacher performance related to pupil gain? Did variance (**S**)—how teachers stood with respect to each other on a particular variable—relate to pupil gain? The third question is the most interesting of all: Did the teachers' rate of change (**B**) on a variable relate to pupil gains?

These variables are repeated in three groups. In the first group, the code letters of the teaching performance variables are each preceded by an **M**. In this group the *mean* of the teachers' performances is related to pupil gains. In the second group, each performance variable is preceded by an **S**, which represents the *variance* of the performance scores, or how spread out they are. In the third group, each performance variable is preceded by a **B**, which stands for the *rate of change* of the teachers on those variables.

Two columns are labeled *Comprehension* and *Decoding*. In these columns numbers are calculated for a statistic (F). If this number is large enough to be statistically significant, it means that the teaching performance variable is significantly

Table 4
F-Values for the Regression Analysis on RAMOS Variables: Phase I

	RAMOS Variables	Levels 0, 1 Comprehension df = (1, 9) F	Level 1 Decoding df = (1, 7) F
M = mean	**M**-L+S	.0534	.4074
	M-A	.7032	.2624
	M-N	1.2690	1.5929
	M-F	.2156	.2363
	M-M	.1755	3.4924
	M-L+M	.6644	.2941
	M-XX	2.0401	(5.2997) -
	M-CIV	2.8281	.3741
	M-QT	3.5680	(4.6128) +
S = variance	**S**-L+S	1.2952	.0016
	S-A	2.3215	.0263
	S-N	.0148	.0164
	S-F	2.4546	.4888
	S-M	.0502	.2460
	S-L+M	.2168	.0687
	S-XX	3.0539	2.9991
	S-CIVCJ	2.6881	.0947
	S-QT	.1694	.4190
B = slope	**B**-L+S	.7123	.2867
	B-A	*5.8272 -	(4.5154) -
	B-N	*5.4298 +	*6.3303 +
	B-F	.2052	.0602
	B-M	.0358	.2998
	B-L+M	.0001	3.5063
	B-XX	.0040	3.2755
	B-CIVCJ	.2050	.9405
	B-QT	.1376	.0949

*Statistically significant at .05 level
+ or - indicates the direction of the relation
Parentheses indicate numbers approaching statistical significance.

L+S = group size (smaller groups)
A = assessing and diagnosing role
N = instructing role
F = facilitating role
M = managing role
L+M = moving around
XX = no feedback
CIV = asking comprehension questions
CIVCJ = asking comprehension questions
QT = teacher questions

correlated with pupil gains in reading comprehension or decoding. These correlations may be either positive or negative; positive means the larger the scores on the performance variable, the larger the gains; negative means the larger the scores on the performance variable, the smaller the pupil gains. The significant correlations are

Table 5
The Effects of the Training Modules on Teaching Performance: Frequency by Module: Phase II

(Decimal numbers represent percentages of time. Only selected codes appear in these tables.)

	Pre	Post 6	Post 7	Post 8
Role A (Assess Diagnose)				
B	06	07	07	08
B•6	07		07	08
B•6•7	07			08
Role D (Discipline)				
B	01	01	01	00
B•6	01		01	00
B•6•7	01			00
Role N (Instruction)				
B	49	44	47	38
B•6	45		48	39
B•6•7	47			39
Role F (Facilitate)				
B	23	27	28	38
B•6	24		27	37
B•6•7	27			37
Role M (Manage)				
B	11	11	06	05
B•6	12		06	05
B•6•7	08			05
Role T•S (Independent and Supervise Staff)				
B	02	02	02	02
B•6	02		02	02
B•6•7	02			02
Feedback Sign •C (Positive Corrective)				
B	02	03	04	05
B•6	02		04	05
B•6•7	03			05
Feedback Sign T (Negative Task Specific)				
B	03	01	01	01
B•6	02		01	01
B•6•7	01			01
Feedback Sign BB (Both Positive and Negative Both Task Specific and Undifferentiated)				
B	01	02	04	14
B•6	01		04	13
B•6•7	03			13

	Pre	Post 6	Post 7	Post 8
Feedback Sign XX (No Feedback)				
B	07	05	08	07
B•6	06		08	07
B•6•7	07			07
General Skills				
B	38	45	34	37
B•6	42		33	37
B•6•7	38			37
Phonics Skills				
B	32	28	28	22
B•6	30		27	21
B•6•7	29			21
Vocabulary Skills				
B	21	23	23	27
B•6	22		23	27
B•6•7	22			27
Grammar Skills				
B	06	04	07	05
B•6	04		07	05
B•6•7	06			05
Comprehension Skills				
B	35	37	38	36
B•6	37		37	36
B•6•7	37			36
Interpreting Skills				
B	15	18	24	30
B•6	17		25	31
B•6•7	21			31
Critical Judgment Skills				
B	00	01	01	01
B•6	00		01	01
B•6•7	01			01
Material BR (Basal Reader)				
B	06	05	01	01
B•6	05		01	01
B•6•7				

B baseline
B•6 after sixth module etc

marked with an asterisk. Some of these numbers were almost statistically significant and we have indicated them by placing them in parentheses. In my discussion I use both the significant and almost significant to interpret the results.

In Table 4, I am giving you a sample of data that look generally the same at different levels of instruction. The mean level of "no feedback" (**M**·XX) turned out, as you would expect, to be negatively related to pupil gain in decoding, and the mean level in teacher questions (**M**·QT) was positively related. There were no relations with comprehension. Nothing related to the differences in the distribution, the variance. In the group where rate of change (**B**) was the independent variable, a greater rate of

Table 6
F-Values for the Regression Analysis on RAMOS Variables: Phase II

Level 3

RAMOS Variables		Comprehension df = (1, 8) F	Decoding df = (1, 8) F
M = mean	**M**-L+S	.0250	1.1284
	M-A	1.1484	.1256
	M-N	.3413	3.3839
	M-F	.0441	1.7119
	M-M	2.2265	.6307
	M-L+M	1.7356	.4107
	M-XX	.0336	.4719
	M-CIVCJ	4.6384**	.5418
	M-QT	.0022	1.5817
S = variance	**S**-L+S	.2989	1.3615
	S-A	.3505	.8948
	S-N	1.8104	.0761
	S-F	2.0024	1.4472
	S-M	.9363	.0395
	S-L+M	3.2636	.001
	S-XX	.0074	.1106
	S-CIVCJ	.0925	.6374
	S-QT	15.8450*	.1120
B = slope	**B**-L+S	.2833	.5707
	B-A	.4851	.0053
	B-N	.1780	1.4547
	B-F	.1031	5.0617**
	B-M	.2181	.2821
	B-L+M	2.5481	.3299
	B-XX	.0657	.0315
	B-CIVCJ	1.1449	.7359
	B-QT	.0812	5.4050*

*$p_{.05} = 5.32$
**Approaching significance

L+S = group size (smaller groups)
A = assessing and diagnosing role
N = instructing role
F = facilitating role
M = managing role
L+M = moving around
XX = no feedback
CIVCJ = asking comprehension questions
QT = teacher questions

improvement in assessing and diagnosing (**B**-A) was negatively related to pupil gain, whereas a greater rate of change in providing instruction (**B**-N) was positively related to pupil gains. And that relation appeared for both comprehension and decoding.

You are probably saying, "Since they produced all that increase in assessing and diagnosing, I bet they started off the second year doing more assessing and

Table 7
Probability Values for Teacher Regressions on RAMOS Variables Over Time Modules 6, 7, 8

Teacher Number	Size L+S	Role A	Role N	Role F	Role M	Mobility L+M	Feedback Sign XX	Skills CIVCJ	Material QT
01	.0001(-)	.8085(-)	.3397(-)	.0155(+)	.2747(-)	.7919(-)	.6270(+)	.4454(-)	.2251(+)
02	.0001(+)	.3342(-)	.6001(+)	.8263(+)	.9946(-)	.1450(-)	.8321(-)	.6500(+)	.2840(+)
03	.1142(+)	.4328(+)	.0135(-)	.0132(+)	.0061(-)	.1666(+)	.2075(+)	.3927(+)	.0718(+)*
05	.8654(+)	.0462(-)	.8787(+)	.2545(+)	.5267(-)	.0634(+)*	.3866(-)	.0047(+)	.7501(+)
06	.0214(+)	.5146(-)	.1451(-)	.0134(+)	.4867(-)	.4289(-)	.4303(-)	.2864(+)	.1190(-)
07	.1977(+)	.8519(-)	.5809(-)	.0709(+)*	.8862(-)	.6216(+)	.8983(+)	.0147(+)	.5864(-)
08	.2458(-)	.1238(+)	.0269(-)	.0022(+)	.8842(-)	.0664(+)*	.0607(+)*	.5712(-)	.8271(-)
09	.7440(+)	.4720(-)	.4032(+)	.8428(+)	.0296(-)	.4777(-)	.1271(-)	.3277(+)	.0502(+)
10	.0477(+)	.8625(-)	.6894(-)	.2692(+)	.0385(-)	.6771(-)	.8547(+)	.2749(+)	.9730(-)
11	.0206(-)	.1502(-)	.1740(+)	.0776(+)*	.0593(-)*	.7155(-)	.2282(+)	.6171(+)	.6502(+)
12	.8311(-)	.4800(+)	.0136(-)	.2569(+)	.8171(-)	.3875(-)	.2463(+)	.3132(-)	.1910(-)
13	.1027(+)	.2707(+)	.0195(-)	.0871(+)*	.5417(-)	.4374(+)	.1676(+)	.0013(+)	.6321(-)
14	.9229(+)	.1586(+)	.4355(-)	.9795(-)	.7510(-)	.8469(-)	.1182(+)	.7107(+)	.7047(+)
15	.0048(+)	.7719(-)	.9318(+)	.4937(+)	.1821(-)	.1687(-)	.7557(+)	.8640(-)	.3614(-)
16	.5389(+)	.4141(+)	.7804(-)	.1222(+)	.3230(-)	.5206(-)	.1349(+)	.1709(+)	.4993(-)
18	.9495(-)	.0275(-)	.9089(+)	.7433(-)	.6917(-)	.1183(+)	.6320(+)	.3775(+)	.1989(+)
19	.9095(-)	.2984(+)	.5995(-)	.8658(-)	.4085(-)	.5152(-)	.3324(+)	.7092(+)	.5757(+)
23	.0473(-)	.2457(+)	.9653(-)	.6659(+)	.1516(-)	.2076(-)	.7931(+)	.8982(+)	.7955(-)

[ ] Significant p-value
*Approaching significance

L+S = group size (smaller groups)
A = assessing and diagnosing role
N = instructing role
F = facilitating role
M = managing role
L+M = moving around

diagnosing." In Table 5, as you can see, the teachers are back where they were at the beginning of the first year (compare Table 5 with Table 3, Role A). Note also that instruction (N) showed a decline (refer to Table 5).

A principle that comes out of social learning theory is applicable here: Modeling and demonstrations are the most effective ways to get more rapid acquisition of a skill, but in order to maintain the behavior over time you need to supply continuous reinforcement for using the skill. I think one of the problems in a training system is that during initial training you may have a modeling and feedback system that facilitates acquisition, but the maintenance of the behavior is ignored after this stage of training.

In the second year we trained the teachers on comprehension skills. Modules six, seven, and eight are designed to increase comprehension skills. Facilitating (F) goes up significantly. The amount of feedback (BB) goes up; it started out low and increased significantly.

The level of use of comprehension skills is almost identical to what it was the year before (compare Table 3 and Table 5). The interesting result is that the training worked very effectively in terms of producing interpreting skills, a change from 15 percent to 30 percent. Critical judgment skills also changed, but we probably should not take that change too seriously because the absolute amount is so small.

What these data show, of course, is that some of the training is effective. If you think of the training as having three projected consequences—increasing on-task

Table 8
Regression Weights for Teachers from Regression of Student Achievement Measures: Phase II

Level 3

Teacher Number	Regression Weight
1	-5.0724
2	4.7807
3	7.6048
5	9.3363
6	-8.8921
7	-1.9145
9	0.4733
11	-0.2742
14	-3.6651
15	0 7646
23	-3.141

behavior and productive work, assessing and diagnosing, and fostering reading comprehension--I think we were one-third successful. Other results were mixed.

Table 6 presents for the second year data similar to that presented in Table 4 for the first year. Variable CIVCJ is the code for comprehension skills; that is, the scores on this variable represent how frequently the teachers are asking comprehension-type questions. It is significantly related to pupil gain in comprehension.

Obviously the modules stimulated greater use of comprehension skills, and teachers who had higher scores on these skills produced greater gains in pupil learning.

The variance in teachers' questions (**S**-QT), a measure of how widely spread out the performance scores are, relates significantly to gains in comprehension. And those teachers who ask more questions over time (see **B**-QT) and those who increase their rate in facilitating (see **B**-F) have pupils who made significant gains in decoding skills.

This rate of change in the teachers is not something you see very often in reports on research on teaching. We were able to produce these data because we had enough observations to do a regression analysis across time. I suspect that the quantitative measure is picking up a kind of aptitude for learning teaching skills, and that leads me to think that the capacity of some teachers to profit from training more than others is underlying some of these data. This hunch is supported by some other data.

Refer to Table 7. In the left-hand column are code numbers identifying the teachers. Across the top are those nine basic categories of teacher variables that should have been affected by the training and should have had an influence on pupil learning. In the boxes are numbers representing significant changes on those variables for each teacher.

There are two ways to read this table. One way is to look across horizontally and see in how many ways a teacher changed or did not change. Teacher 01, for example, modified her or his group size and changed her or his facilitating role. Teacher 03 has changed in three respects, one positive and two negative. Teacher 14 did not change at all.

On some variables we produced positive changes in some teachers and negative changes in others, and it is really only in facilitating (F) that we have four positive changes. We have three positive changes in comprehension (CIVCJ). You

will remember that those two variables were related to significant pupil gains. (Refer to Table 6.)

These data, of course, convince me that the problem of changing teachers' styles and performances is extraordinarily complex. We do not know how to change this behavior, these performances, these styles, so that each teacher is genuinely effective.

I emphasize that we keep finding differences in teachers related to pupil gains. I would like to show you some numbers and call your attention to their significance.

Refer to Table 8. These numbers were derived from an analysis of how much effect a teacher has on pupil gains. O is the base point; you see numbers that go up as high as +4 and +6, and you see numbers in the negative direction that go as low as -6 and -7. These differences mean that some teachers are producing far greater gains than were predicted by the pupils' scores in the fall and in the winter, and some teachers are producing far fewer gains than were expected by the initial scores.

Can research techniques like those used in our studies be adapted to develop inservice programs? A practical program may go something like this. You gather test data from the fall and winter, and in that period of time you observe the teachers. Then, on the basis of pupil scores, you identify those teachers who obviously need help because they are producing fewer gains than were predicted for their pupils. I would use the observational data to identify those aspects of their teaching that are likely to be related to less gain, and then design an inservice training program to help them.

I also was asked to comment on feedback to teachers. The information we gather in our research is too rich to summarize quickly, so it is difficult to use these data to help teachers. We keep trying to develop a system that will reduce the number of teaching performance variables we have to look at and one that will produce a continuous record over a period of a week or two weeks that can be used to talk to teachers about their performance skills.

The problem is a research problem. We researchers have to find out which variables (teaching performances) make a difference in pupils' learning. We have made extraordinary progress on this problem in the past five years. We are now at a point where we can be fairly specific about the variables or factors that make a difference in learning in some subjects. But what we and you need to do is to refine and simplify the systems for observing teachers and giving them information that will help them change.

If you are going to use teacher performance data as feedback, you have to have a system that quickly gives teachers information on their teaching. You probably ask, "What about the videotape?" One of the problems with the video camera is that it is a very limited eyeball; it does not swivel as well as your head. It is always controlled by the camera operator. It has a very narrow range of vision and a very poor ear. It is a very limited observer. It probably can be used, but should be used selectively. It works better with high school teachers, because they do not move around as much as elementary teachers (with all due respect to high school teachers).

Another problem is how you talk to teachers about this information, whether live observation or videotapes. How do you translate numbers into actions that are meaningful to teachers? What I would do, now that I have learned a little bit, is to begin with pupil data and try to devise what could or should be done by the teacher.

I would use the regression lines and analyses for each class as a way of estimating a teacher's effectiveness. I would use those statistical procedures as a diagnostic tool to start talking about how the teacher interacts with pupils, what he or she has planned for them, and so on; and then I would use observational data as a part of the information to see how the teacher actually copes with teaching problems.

I am still baffled by what I think is the real underlying problem. I have now come to think of these teaching styles as essentially a form of coping behavior. As a psychologist I recognize how difficult it is to change coping behavior because it has functional value.

I believe we need a much more personal approach in inservice training, one in

which we study the teacher's perceptions, beliefs, and expectations, as well as their performances and knowledge. We also need to learn how to design inservice programs that modify teaching styles rather than isolated teaching behaviors. We need to study how teachers learn, and, as we do, I expect our inservice program will look quite different from the traditional ones and will be more effective.

Biographies of Speakers

Frederick J. McDonald is a consultant at Fordham University. He has a Ph.D. in educational psychology from Stanford University, where he was a faculty member and director of numerous research projects. He has also been associated with Educational Testing Service, Johns Hopkins University, the University of Texas Research and Development Center in Teacher Education, and New York University. Dr. McDonald's major interest is teacher education and evaluation, which he has researched extensively over the past fifteen years. He has published widely on these topics.

Donald M. Medley is Professor, Department of Research Methodology, Curry Memorial School of Education, University of Virginia. He received his doctorate from the University of Minnesota. He has devoted most of his professional life to research in teaching and teacher education, including eleven years at the Office of Research and Evaluation in the Division of Teacher Education of the City University of New York, five years as head of the Teacher Behavior Research Group at the Educational Testing Service, and nine years at the University of Virginia. He is the author of numerous journal articles and other publications, most recently of Teacher Competence and Teacher Effectiveness, a monograph commissioned and published by the American Association of Colleges for Teacher Education.

Robert S. Soar is Professor, Foundations of Education Department, University of Florida. He holds a Ph.D. from the University of Minnesota and has taught in the past at Temple University, the University of South Carolina, Vanderbilt University, and the University of Minnesota. Dr. Soar's major research interest is in the measurement of classroom behavior and teacher effectiveness, and he is the author of many papers on these topics.