DOCUMENT RESUME

ED 343 864 SP 033 633

AUTHOR Holdzkom, David; And Others
TITLE A Longitudinal Study of Teacher Performance Evaluation.
PUB DATE Mar 89
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 1989).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Evaluation Research; Evaluators; Longitudinal Studies; Program Development; *Program Evaluation; Research Methodology; State Programs; Statistical Analysis; *Teacher Evaluation
IDENTIFIERS North Carolina

ABSTRACT

Educators in North Carolina, recognizing a need to develop a more objective, performance-based teacher evaluation system, determined that evaluation of teachers should have both formative and summative outcomes and should focus on improvement of teacher skills. The Teacher Performance Appraisal System (TPAS), a companion to the North Carolina Career Development Program, requires at least three classroom observations of a teacher by either the principal or a peer observer. Both of these programs focus on the growth and development of educators, both teachers and teacher evaluators. A large-scale, state-ordered review of the performance system, as perceived by evaluators and evaluatees, led to the present study, which examines the results of the evaluation system over a 3-year period (1985-86 to 1987-88). The study examines the evaluations of teachers (up to 6,257) in 15 different school districts. In general, results of this study confirm and illuminate the improvement in performance expected of both evaluators and teachers for the first 3 years of a 4-year pilot program. Results also confirm the developmental and interactive nature of these improvements. (IAH)

A Longitudinal Study of Teacher Performance Evaluation

by
David Holdzkom
Director, Division of Personnel Relations

Dennis Stacey
Consultant, Division of Personnel Relations

Barbara Kuligowski
Consultant, Division of Personnel Relations

North Carolina Department of Public Instruction
Raleigh, North Carolina 27603

Prepared for presentation at the 1989 Annual Meeting of the American Educational Research Association, San Francisco, CA., March 1989.

BACKGROUND

Since the early 1980's, educational policy-makers and practitioners in North Carolina, like their counterparts throughout the nation, have wrestled with a complex of issues that can loosely be identified as "the reform movement". The fact that this reform movement has reflected and been influenced by research-based knowledge has paradoxically become both a strength and a weakness. Reliance on the research base has been a strength insofar as it has guided the creation of an agenda for action. However, the research base has been called into question by those who naively felt that it guaranteed improvements in educational outcomes in the same way that aspirin conquers headache. The inability to report both immediate and measurable improvements has led some to question the potential efficacy of the changes implemented.

This disgruntlement has, to some degree, been neutralized in North Carolina by continued support from key legislators and policy makers at the state and local level who understand that the effects of a given change may occur at some remove in time from installation. The General Assembly has been generous in both financial support and moral support of a long-term improvement agenda. Evaluation of teachers' skills offers an excellent example of such support. In this paper, we will present information about the teacher evaluation system used in North Carolina, as well as its effects on the performance of teachers over a three-year period.

Beginning in 1979-80, educators in North Carolina recognized a need to develop an evaluation system that would enable a more objective, performance-based view of the skills level of individuals. Apart from other motives, it was felt that, because the State guarantees a uniform salary schedule for teachers, it was a matter of fairness that evaluation be conducted

uniformly through the State. As a first step toward development of such an evaluation system, a working group that combined educators from the North Carolina Department of Public Instruction (DPI) and from local school districts was charged with establishing consensus on the set of skills that ought to constitute the basis of evaluation. At this time, wide latitude was left to local units with respect to how evaluative data were collected, but uniformity about targets was achieved. (Inman, 1982).

Building An Evaluation System

Following this first effort, an immediate revision was undertaken with twin goals. First, the development of a large amount of so-called "effective teaching" research was becoming generally accepted in the practitioner community. This research was seen as useful in the establishment of performance standards and criteria. Second, many of the problems encountered in early implementation of the consensus-based instrument were related to procedural irregularities stemming from choices made by administrators in local school districts. It was hoped that establishment of a standardized evaluation procedure would result from this second effort.

This emphasis on procedure arose from the realization that between research and practice a large gap existed that required some activity akin to engineering. That is, the research knowledge on which an instrumert could be based had been developed in particular settings, with particular people. In science, the replication of a an experiment and the outcomes that are congruent with the first iteration constitute an important aspect of theory-building. Insofar as the replicator reproduces the original conditions, he is able to predict effects similar to those observed by the initiator. Failure to replicate the results can lead to an attack on the original experiment. Failure to replicate the test conditions, however, discredits the replicator.

For most school practitioners, however, innovations are not implemented as a means to increase the power of a given theory. Rather, the practitioner hopes to solve a problem in an environment bearing a semblance to that in which the original study was undertaken. By tailoring an innovation, which preserves the essence of the research knowledge and which is adapted to a specific context, the practitioner develops or engineers a research-based solution. The task, then, is to establish generalizability of the core of the knowledge such that it can influence practice in settings similar to, but not identical with, the original settings. The problem is not one of replication, but one of utilization.

In solving the engineering problem, then, there was a need to specify clearly the ends to which evaluation was to be put. Theoretically, an evaluation tool could be designed so that it simply yields summative information. A thermometer, for example, is useful because it measures with accuracy a person's internal temperature. The thermometer is not expected to tell anything about how to change the temperature. For purposes of teacher performance evaluation, however, summative evaluation was only one outcome. Policy-makers in DPI recognized that the evaluation system could be used as an instrument for improvement of teachers' skills only if the system rendered formative information. Because the reform movement is intended to move behavior from one point to a more desirable one, this emphasis on improvement was integral to the performance evaluation system. Thus, it was decided that evaluation of teachers should have both formative and summative outcomes.

At this point, another decision was made: Evaluation data would be used for improvement of skills. Wise and his colleagues describe several evaluation systems that have both summative and formative outcomes. (Wise et al., 1984). However, generally the systems they describe are applied formatively only to

marginal performers. For those persons teaching at an acceptable level, these systems, presumably, are not used for improvement of skills beyond basic competence. One can wonder how motivated a teacher is to participate in an evaluation process whose advertised goal is to separate the "can do's" from the "can't do's". It would be remarkable if such a system could bring about improvement beyond a minimal level, if only because most teachers will view such a system as irrelevant to themselves, if not actually pernicious. If, however, we begin from the premise that teachers' skills could be improved, regardless of the individual's level of functioning, then the evaluation system would need to be part of an ethos that honors critical self-assessment, reflective observation by a trained observer, and district-supported staff development/training activities. Such an evaluation system could be expected to foster change in levels of observed skills.

Yet, as Joyce and Showers point out, behavioral change is likely to be a fairly slow, incremental process of successive approximations, if not trial and error. (Joyce and Showers, 1982). Moreover, the movement from acquisition of skill to utilization of skill is the result of coaching, not evaluation (Joyce and Showers, 1982; Eaker and Huffman, 1980.) Thus, if evaluation can be used as a sort of needs assessment, only through coaching will the desired change in behavior occur. In addition to evaluation of teachers' skills, then, the new system needed to include a component for <u>improvement</u> of skills, regardless of current level of functioning.

This notion of incremental change is confirmed by Hall and Loucks who have identified discrete stages through which people progress while in the process of change. (Hall and Loucks, 1983). Not only does this insight suggest that change could better be thought of as operating on a continuum rather than in a dichotomy, it also suggests that some time may be required before an innovation, in this case the evaluation system, can be expected to render measurable results.

In summary, then, research shows that:

1. Organizations have developed systems for evaluating teachers' performance.

2. In order to change behavior, specific desirable actions and environments need to be in place.

3. Change occurs in predictable stages over time as people grow with an innovation.

How the TPAS Works

The evaluation system developed in North Carolina requires a minimum of three observations in a teacher's classroom. The observer (either principal or peer observer) keeps detailed notes of the teacher's behaviors and activity during the observation. Later, these behaviors will be analyzed in a narrative report that focuses on the teachers' performance in at least five functions:

1. Management of Instructional Time

2. Management of Student Behavior

3. Instructional Presentation

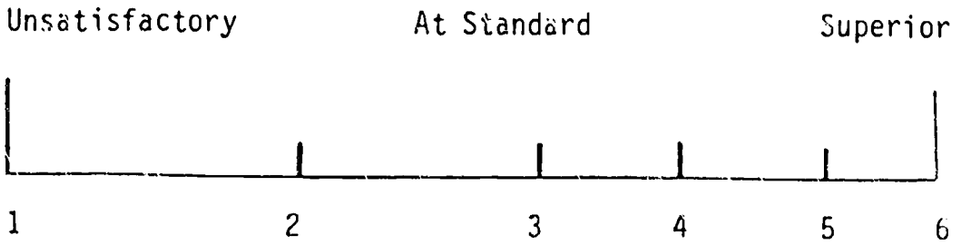
4. Instructional Monitoring

5. Instructional Feedback

The report is presented to and discussed with the teacher during a conference conducted within five days of the observed class period. Based on the report, the observer and teacher examine the skills analysis in terms of the teacher's Professional Development Plan, a systematized plan for skills development.

At the end of the school year, all the observation reports and other data are taken together and a summative evaluation is completed. This summative judgment renders a rating from 1 (Unsatisfactory) to 6 (Superior), with 3 (At Standard) as a mid-point. This results, of course, in a scale of unequal intervals, giving greater precision of discrimination at the upper end: The

distances between 1 and 2, 2 and 3, 3 and 5, and 4 and 6 are equal:



Because ratings are not assigned to individual formative observation reports, averaging of scores is avoided and reliance must be placed on the definitions of the scale points, which evaluate both quality and consistency of the teacher's performance.

In addition to the five functions evaluated primarily through in-class observation, three other functions are included in the evaluation:

6. Facilitating Instruction
7. Communicating Within the Educational Environment
8. Non-Instructional Duties

These functions are not primarily observable in classrooms and administrators are required to collect further evaluative data as a kind of by-product of their normal supervisory responsibilities. As will become clear, this evaluative laissez faire has not been without its difficulties.

Within each of the eight functions the observer is guided in data collection and analysis by the further specification of practices. The five functions observable in classrooms are comprised of 28 specific skills that were identified in the teacher effects literature (Gage and Needels, 1989). In order to be included, a skill had to be identified as associated with increased student achievement and/or increased student time on task in two or more experimental studies. Moreover, each skill has to be observable, generic across grades and

subject matters, and alterable. (The Group, 1983). As a check on the comprehensiveness of the criteria, a third-party evaluation of the instrument and process was conducted. The panel reported that the instrument was "admirably suited to the purposes for which it is intended" (Brandt et al, 1988).

Finally, the Department of Public Instruction conducted a large-scale review of the performance system, as perceived by evaluators and evaluatees. Generally speaking, both groups understood the features of the Teacher Performance Appraisal System (TPAS), accepted the criteria as reasonable and important, and reported satisfaction with initial implementation results (Stacey et al, 1988). A copy of the evaluation criteria is included in Appendix 1.

The present study examines the results of the evaluation system over a three-year period (1985-86 to 1987-88), examining the evaluations of a large number of teachers (up to 6257) in fifteen different school districts. The teachers represent a 100 percent sample in each of the school districts participating in the North Carolina Career Development Plan pilot study. Performance appraisal data were obtained in July and August following each of the pilot years and were reported anonymously to staff of the Division of Personnel Relations by the Career Development coordinators in each of the fifteen districts. While the Career Development Pilot includes sixteen school districts, one of these has been eliminated from consideration in the present study because of large differences in evaluation criteria and teacher participation rates. For our purposes only 15 school districts will be considered. Preliminary analyses (frequency distributions of each rating point by function by district, as well as mean and modal ratings at the district and pilot levels) were shared with school district staff in October of each succeeding year. No effort was made to establish appropriate or desired norms within a school or a district, nor were there organizational consequences (from DPI or the district's central office) aimed at ensuring particular distributions.

Training was provided for teachers (a 30-hour in-service course on effective teaching) and evaluators (a 24-hour in-service course on evaluation of performance, followed by a six-hour course in PDP utilization). Evaluators' reliability was tested at the end of the 24-hour performance appraisal training. In addition, booster workshops (six to ten hours) were provided in February, 1987 and March, 1988 based in part on the analysis of evaluation ratings.

In addition to these state-sponsored activities, local districts were free to develop other training activities for observers, provide individualized technical assistance to observers, and administer inter-rater reliability checks as needed. The State Department of Public Instruction also provided two checks of inter-rater reliability in April 1986 and June 1987.

One additional aspect of CDP should be mentioned. Because the teachers in the sample were participating in this pilot, their evaluation results determined the merit increase in salary, if any, that they would receive. Teachers who were rated at least 3 in all eight job functions were eligible in 1986-87 for a five percent salary rise. Teachers who earned at least seven 4's and one 5 in 1986-87 and who had at least six years' teaching experience were eligible for an additional ten percent rise, beginning in 1987-88. Thus, the evaluations had visible consequences that constituted a motivation for maintaining, if not improving, performance.

Purpose of Study

Given these descriptions, goals, and assumptions of both the North Carolina Career Development Program (CDP) and companion Teacher Performance Appraisal System (TPAS), the purposes and hypotheses of this study will be set forth. Both of these programs are focused on the growth and development of educators, be they teachers or those who evaluate teachers. Consequently, a general expectation of the CDP is that both teacher-evaluators and participating teachers would demonstrate improved performance or skills in

their respective jobs over the course of the pilot, or for purposes of this study, the first three years of the pilot. For evaluators, this improvement might take the form of being able to better recognize demonstrations of quality teaching, as well as being able to properly account for such teaching through reliable and consistent use of the appraisal system. For teachers, this skill improvement would yield progressively higher performance ratings and levels of career status.

Accordingly, the two main expectations of, or predictions for, the CDP are that: (a) Participating teachers, as a whole, will demonstrate progressively improved performance over time, and (b) evaluators will increasingly be able to recognize these improvements in an increasingly consistent way. In more statistical terms, the progressive convergence of overall evaluator ratings will be reflected in decreasing standard deviations of their ratings, whereas improved teacher performance will be reflected in mean performance ratings that increase over time.

In addition, these changes are expected to occur in a particular developmental sequence. The first year of the pilot is likely to be characterized by somewhat random or scattered ratings of teacher performance as evaluators sought to clarify evaluation standards and calibrate their ratings relative to one another. The second year should be characterized by a settling or convergence of ratings, both within and among units, accompanied by a modest overall increase in ratings at the same time that they are converging toward an overall mean. The third year will reflect a continuation of the second-year trend: Continued tightening (or decreased variance) of ratings, as well as continued, modest increases in performance ratings.

It is apparent from the above discussion that these two predicted main effects are covariable or interactive. That is, at the same time that the

variance in ratings of teacher performance is decreasing (or that the reliability of these ratings is increasing), the ratings themselves are progressively increasing. Given the present state of development of the TPAS and related program evaluation techniques, it is unclear how much of the overall increase of performance ratings is due tc improved evaluator skills, and how much is due to improved teacher performance. Thus, this study will attempt to track both effects, without necessarily being able to differentially credit one or the other.

A further consideration of this dual-effects phenomenon has to do with the conceptual and operational differences between the first five, research- and classroom-based functions of the TPAI and the last three functions that relate primarily to the teacher as an employee of a larger organization. As stated earlier, the former functions have received much more attention in the form of training and technical assistance than have the latter functions. Consequently, it is expected that the predicted main effects will look different for Functions 6-8 than for Functions 1-5. This difference would seem to be most evident in terms of more improvement (as measured by higher summative ratings) being attributed to Functions 6-8 than Functions 1-5, especially in CDP units. Likely reasons for this difference include: (a) In the face of relatively less rigorous evaluation criteria, ratings of these functions will tend to be higher; and (b) the likelihood of higher ratings will be enhanced in CDP units where advancement of teachers on the incentive pay scale is dependent on attaining a given level of performance ratings.

Finally, for the third year of the pilot, performance appraisal data were also made available by a sample of non-pilot districts, following the first year of statewide, mandated evaluation of teachers using the TPAS. These districts provide another perspective on or further illumination of the developmental nature of the TPAS by virtue of being in the early stages of

their own use of the system, without nearly the preparation for its use that the pilot districts received. Accordingly, it is expected that TPAS outcomes in these districts would be more consonant with those of the first stage of the pilot districts' experience with the system, namely, relatively less developed levels of evaluator and teacher skill or improvement. In short, these non-pilot districts represent a pseudo-control group for many of the developmental outcomes already proposed.

METHOD

Sample

As indicated earlier, this study addresses the outcomes of the staged development of the North Carolina Teacher Performance Appraisal System (TPAS) over the first three years of a piloted, four-year Career Development Program (CDP). The longitudinal part of this study, which tracks 15 North Carolina public school districts from 1986-1988, is clearly the major thrust of it. Secondary to this major purpose is a comparison of the pilot districts to nine volunteer, non-pilot districts at the end of the third-year of the pilot--the cross sectional component of this study.

For all of these districts year-end performance appraisal data were submitted for virtually all classifications of professional school-based staff (teachers, auxiliary or support staff, and administrators), and, to a lesser degree, central office staff. Because the TPAS is the best developed of all of the North Carolina appraisal systems for professional public school personnel, only teachers are included in this study. The number of full-time employed teachers who constitute the analysis for the three-year period of this study are as follows:

	CDP Districts	Non-CDP, Volunteer Districts
1986	5119	
1987	6131	
1988	6257	2245

The lesser number of non-CDP teachers for 1988 reflects not only fewer participating non-CDP districts, but also a procedural option available to these districts by which school administrators can evaluate a portion (e.g. half) of their total tenured staff in each year of, or stagger the evaluation of their staffs over alternate years. Virtually all non-CDP districts exercised this option.

Participants include teachers of all levels of experience, including those in their first or second year of a mandated Initial Certification Program, third-year provisional teachers, and participants and nonparticipants in the optional career status I and II levels of the CDP units. Because non-CDP districts do not make career status designations of teachers, the only meaningful available variable on which their teachers can be equated to teachers of CDP districts is teaching experience. Such a comparison between the two types of units for 1988, given that years of experience was reported for only 56 percent of non-CDP teachers, reveals that there is no noteworthy difference between the districts in the proportion of teachers with different years of experience. Participants' years of experience range from 1 to 44, and all grade levels and subject areas are represented.

Procedure

Performance appraisal information on the sample was collected via completion of an annual reporting form which was overseen in CDP districts by the Career Development Coordinator, and in the non-CDP districts by personnel administrators. Whereas the requested information varied slightly for each of the three years of the pilot program, standardized information for all three years included each participant's (a) school district and school, (b) unique identification number, (c) career status designation (in the case of CDP districts), and (d) performance appraisal ratings for up to eight functions of the North Carolina Teacher Performance Appraisal Instrument. Additional information acquired in the first and third years of the pilot included (a) years of experience, and (b) grade level(s) and/or subject area(s) taught.

The vast majority of these data were prepared for analysis by completing, correcting, and coding or re-coding the reporting forms, as necessary. The data were then assembled into data files using an IBM-AT microcomputer and PC-File+ data management software. Subsequent data files were verified for

accuracy before being analyzed. Data were analyzed using Base-SAS software, which yields comprehensive descriptive statistics. Given the hypotheses of this study, and the largely exploratory nature of the study, the authors deem descriptive statistics as an appropriate vehicle for examining the hypotheses.

Design of Study

For ease of analysis and communication of this relatively complex, three-year study, the following experimental design considerations are proposed. First, major dependent measures that underlie the study are: (a) mean ratings of job functions, as indicators of quality of performance, and (b) standard deviations of mean function ratings, as indicators of the consistency or convergence of ratings.

Second, as mentioned earlier, the pilot study will be longitudinally analyzed in terms of its three years of staged modification and development. Because development and improvement of teachers is a major expected outcome of the CDP, mean ratings of teacher performance will be examined over time of the pilot. At the same time, it is expected that evaluator skills will also improve. Therefore, the stability and consistency of ratings will also be examined over time of the pilot.

Finally, performance appraisal data of the nine, volunteer non-pilot districts will be cross-sectionally compared to that of the pilot districts. It is hoped that this comparison will serve to further illuminate the evolutionary nature of the TPAS within a staged career development program.

RESULTS

Given the experimental design considerations described above, this study will limit its analyses to the three-year longitudinal study of major trends for the 5 CDP districts, and a comparison of CDP and non-CDP districts for 1988. Accordingly, analyses will be pitched at the level of school districts

or aggregates of school districts. The major dependent variable will consist of frequency distributions or percentile ranks, and means and standard deviations of teachers' performance ratings.

Longitudinal Analyses of CDP Districts

> Hypothesis 1: Mean ratings of TPAI functions in CDP districts will stabilize over time as a function of increased convergence or consistency of aggregated ratings. Thus, over the three-years of the pilot, the standard deviations (SDs) of mean function ratings are expected to modestly decline.

The first and major part of this hypothesis receives overwhelming support (See Table 1). That is, when considering the aggregated (by districts) mean ratings of each function from year to year, the SDs for these means exhibit a progressive decline over the pilot period. As can be seen in Table 1, the decline in SDs from 1986 to 1987 ranges from .024 to .074, whereas the decline from 1987 to 1988 ranges from .011 to .052 across all eight functions.

Moreover, the majority of this settling or convergence of ratings occurs between the first and second years of the pilot: The average reduction of SDs of mean-ratings between 1986 and 1987 is .050, whereas that average reduction decreases to .035 between 1987 and 1988 (See Table 1). The total settling of ratings across all three years of the pilot, therefore, amounts to an average reduction in SD of .085, indicating that evaluators, on the whole, did become more consistent over time.

> Hypothesis 2: TPAI ratings, aggregated for the CDP units, will demonstrate steady, moderate improvement over time. Third-year ratings will be modestly higher than second-year ratings, and both may well be higher than first-year ratings. As a sub-hypothesis, it is predicted that given that the 1986 aggregated district ratings are roughly equal for all functions, mean ratings of Functions 6, 7, and 8 will exceed those for the classroom-based Functions 1 through 5 over the next two years.

As seen in the aggregated ledger of Table 2 for all 15 pilot districts, the ratings for 1988 are higher than those for 1987 for every function (F). These differences range from a 0.11 increase for Functions 3 and 7 to a 0.17

Table 1. Decreasing Standard Deviations of Mean Ratings of TPAI Functions for Aggregated Pilot Districts Over Three Years of the Career Development Program

STANDARD DEVIATIONS OF MEAN FUNCTION RATINGS

Function	1986	1987	Diff.	1987	1988	Diff.	1986-88 Diff.
1	.979	.938	-.041	.938	.892	-.046	-.087
2	1.032	.958	-.074	.958	.925	-.033	-.107
3	1.004	.972	-.032	.972	.920	-.052	-.084
4	.948	.890	-.058	.890	.851	-.039	-.097
5	.943	.882	-.061	.882	.849	-.033	-.094
6	.948	.924	-.024	.924	.902	-.022	-.046
7	.982	.914	-.068	.914	.903	-.011	-.079
8	1.037	.996	-.041	.996	.950	-.046	-.087
			Total = -.399 Mean = -.050			Total = -.282 Mean = -.035	Total = -.681 Mean = -.085

Table 2. Mean Ratings of Teachers Using the N. C. TPAI for the First Three Years of the Career Development Program: By Individual and Aggregated Pilot Districts

DISTRICT	Year	TPAI FUNCTIONS 1	2	3	4	5	6	7	8	Number of Teachers
Alexander	86	4.02	3.53	4.15	3.76	3.66	3.92	3.87	4.15	(207)
	87	4.03	4.00	3.98	4.01	3.89	4.14	4.10	4.48	(234-237)
	88	4.46	4.48	4.24	4.47	4.24	4.38	4.37	4.57	(243)
Buncombe	86	4.68	4.55	4.73	4.73	4.69	4.60	4.84	4.79	(1066-1074)
	87	4.69	4.73	4.76	4.78	4.73	4.81	5.05	5.06	(1115-1042)
	88	4.85	4.91	4.82	4.89	4.84	4.91	5.08	5.16	(1161-1164)
Burke	86	4.09	4.01	4.21	4.08	4.00	4.06	4.12	4.15	(649-651)
	87	4.19	4.17	4.17	4.17	4.10	4.29	4.37	4.40	(579-660)
	88	4.31	4.26	4.30	4.28	4.26	4.45	4.47	4.55	(653-660)
Burlington	86	4.44	4.50	4.57	4.46	4.37	4.36	4.42	4.40	(322-324)
	87	4.44	4.45	4.54	4.40	4.42	4.41	4.52	4.54	(338-341)
	88	4.52	4.54	4.65	4.54	4.57	4.61	4.67	4.69	(336-337)
Chowan	86	4.61	4.47	4.33	4.20	4.36	4.27	4.33	4.47	(99)
	87	4.53	4.60	4.32	4.56	4.65	4.84	5.11	5.11	(133)
	88	4.58	4.71	4.43	4.56	4.80	4.94	5.08	5.18	(136-137)
Greene	86	4.90	4.80	4.73	4.73	4.78	4.89	4.84	5.06	(157)
	87	4.76	4.61	4.53	4.61	4.53	4.87	5.16	5.38	(154-156)
	88	4.48	4.48	4.48	4.53	4.43	4.46	4.63	4.62	(153)
Harnett	86	4.09	4.08	4.10	3.97	4.00	3.93	4.08	4.08	(567-569)
	87	4.34	4.33	4.22	4.18	4.20	4.20	4.42	4.42	(585-586)
	88	4.50	4.45	4.37	4.34	4.34	4.38	4.56	4.59	(607-608)
Haywood	86	4.46	4.35	4.60	4.48	4.45	4.37	4.47	4.46	(410-411)
	87	4.47	4.35	4.57	4.42	4.46	4.39	4.54	4.64	(448)
	88	4.54	4.47	4.59	4.47	4.52	4.52	4.70	4.84	(448)

Table 2. (Cont'd)
TPAI FUNCTIONS

DISTRICT	Year	1	2	3	4	5	6	7	8	Number of Teachers
Montgomery	86	5.31	5.24	5.40	5.51	5.55	5.12	5.31	5.39	(223-224)
	87	4.63	4.83	4.78	4.80	4.79	4.57	4.83	4.95	(215)
	88	4.71	4.73	4.52	4.64	4.59	4.78	5.02	5.06	(221-222)
N. Hanover	86*	4.46	4.36	4.40	4.41	4.24	4.29	4.22	4.13	(1183)
	87	4.24	4.24	4.18	4.24	4.19	4.42	4.57	4.56	(960-961)
	88	4.46	4.53	4.45	4.41	4.37	4.69	4.77	4.82	(990-993)
Orange	86	4.38	4.14	4.29	4.30	4.32	4.21	4.40	4.30	(257-259)
	87	4.40	4.17	4.40	4.45	4.33	4.37	4.67	4.56	(282-285)
	88	4.42	4.24	4.45	4.49	4.45	4.54	4.77	4.71	(294-295)
Perquimans	86	4.21	4.18	4.18	4.20	4.19	4.54	4.68	4.82	(97-98)
	87	4.01	4.12	3.95	4.13	4.11	4.22	4.28	4.40	(101)
	88	4.58	4.54	4.56	4.50	4.49	4.77	4.76	4.80	(92-102)
R. Rapids	86	4.34	4.32	4.18	4.43	4.45	4.94	4.08	4.08	(145-146)
	87	4.70	4.46	4.45	4.70	4.59	4.71	4.76	4.68	(155-156)
	88	4.60	4.48	4.37	4.71	4.74	4.76	4.79	4.80	(161)
Salisbury	86	4.56	4.43	4.43	4.73	4.62	4.30	4.61	4.79	(134-135)
	87	4.73	4.51	4.54	4.63	4.58	4.60	4.61	4.65	(139-140)
	88	4.65	4.62	4.65	4.74	4.71	4.75	4.81	4.71	(139-140)
Tarboro	86	4.05	3.85	3.93	4.09	4.17	4.13	4.25	4.09	(163-164)
	87	4.35	4.27	4.08	4.33	4.34	4.12	4.62	4.47	(171)
	88	4.52	4.61	4.29	4.59	4.60	4.48	4.77	4.79	(173-174)
Aggregated	86	4.43	4.32	4.46	4.41	4.39	4.33	4.46	4.48	(4496-4518)
	87	4.42	4.40	4.40	4.42	4.38	4.47	4.65	4.68	(5610-5732)
	88	4.56	4.56	4.51	4.54	4.52	4.64	4.76	4.82	(5808-5837)

* For 1985-86, New Hanover data include all certified staff of whom 81 percent are teachers.

for Function 6, with the greatest increases occurring for (in order of magnitude) Functions 6, 2 and 8, 5, and 1. In addition, the 1987 ratings are higher than the 1986 ratings for four of eight functions (Fs 8, 7, 6 and 2) and effectively the same (differences of .01) for three others (Fs 4, 1 and 5). Of the four functions exhibiting the greatest improvement from 1986 to 1988 (Fs 2, 6, 7 and 8), Functions 6, 7 and 8 exhibit the most change.

Thus, this important hypothesis receives solid support insofar as performance does exhibit a clear pattern of increase or improvement over the three years of the pilot. In only one case is the change in an average rating a clearly negative one--a -.06 change in the rating of Function 3 between 1986 and 1987. This finding must be viewed against an earlier discovery by the senior author (Division of Personnel Relations, NC Department of Public Instruction (DPR/DPI), 1986) that Function 3 is the least reliably, correctly identified of any of the first five functions, a finding that draws further support from a conceptually different 1987 rater-reliability study in which Function 3 was one of two of the least reliably identified classroom-based functions (DPR/DPI, 1987).

When changes in ratings are considered 1986 to to 1988, the most marked overall increases appear to be associated with Functions (Fs) 6, 7 and 8, as predicted by the secondary hypothesis. This predicted pattern for Fs 6, 7, and 8 is illustrated in a different and more succinct way in Table 3. That is, aggregated mean function ratings for the pilot districts are greater for Fs 6, 7 and 8 as compared to the first five functions for every year of the pilot program. In fact, making the same comparison on a unit-to-unit basis reveals that, out of 45 possible comparisons, there are only six occasions where grand means for Functions 6, 7 and 8 do not exceed those for Functions 1 through 5. When this comparison is repeated for just Functions 7 and 8, the effect is even

Table 3. Grand Means of Teacher Ratings on the N.C. TPAI for the First Three Years of the Career Development Program: By Individual and Aggregated Pilot Districts

		TPAI FUNCTIONS (Fs)			
DISTRICT	Year	Fs 1-5	Fs 6-8	Fs 7-8	Number of Teachers
Alexander	86	3.83	3.98	4.01	(207)
	87	3.98	4.24	4.29	(234-237)
	88	4.38	4.44	4.47	(243)
Buncombe	86	4.68	4.74	4.81	(1065-1073)
	87	4.74	4.97	5.05	(1115-1142)
	88	4.88	5.05	5.12	(1162-1272)
Burke	86	4.08	4.11	4.14	(649)
	87	4.16	4.36	4.39	(579-659)
	88	4.28	4.49	4.51	(653-659)
Burlington	86	4.47	4.38	4.41	(322-324)
	87	4.45	4.49	4.53	(339-341)
	88	4.57	4.65	4.68	(336-337)
Chowan	86	4.40	4.36	4.40	(99)
	87	4.53	5.02	5.11	(133)
	88	4.62	5.07	5.13	(136-137)
Greene	86	4.79	4.93	4.95	(157)
	87	4.61	5.13	5.27	(154-156)
	88	4.49	4.57	4.63	(153-157)
Harnett	86	4.05	4.03	4.08	(567-568)
	87	4.26	4.34	4.42	(585-586)
	88	4.40	4.51	4.57	(608-609)
Haywood	86	4.47	4.44	4.47	(410-411)
	87	4.45	4.52	4.59	(448)
	88	4.52	4.69	4.77	(448)

Table 3. (Cont'd)
TPAI FUNCTIONS (Fs)

DISTRICT	Year	Fs 1-5	Fs 6-8	Fs 7-8	Number of Teachers
Montgomery	86	5.40	5.27	5.35	(222-224)
	87	4.77	4.78	4.89	(215)
	88	4.64	4.95	5.04	(221)
N. Hanover	86*	4.37	4.21	4.18	(1183)
	87	4.22	4.53	4.56	(960-961)
	88	4.44	4.76	4.79	(990-995)
Orange	86	4.28	4.30	4.34	(257-258)
	87	4.35	4.53	4.61	(282-285)
	88	4.41	4.67	4.74	(294-295)
Perquimans	86	4.20	4.66	4.73	(97-98)
	87	4.06	4.30	4.34	(101)
	88	4.51	4.77	4.78	(80-90)
R. Rapids	86	4.34	4.03	4.08	(145-146)
	87	4.59	4.71	4.71	(155)
	88	4.58	4.78	4.80	(161)
Salisbury	86	4.56	4.56	4.69	(134)
	87	4.60	4.62	4.63	(139-140)
	88	4.67	4.76	4.76	(139-140)
Tarboro	86	4.02	4.16	4.17	(163-164)
	87	4.27	4.40	4.54	(171)
	88	4.53	4.68	4.78	(173)
Aggregated	86	4.40	4.42	4.47	(4494-4512)
	87	4.40	4.60	4.66	(5611-5729)
	88	4.55	4.74	4.79	(5798-5934)

* For 1985-86, New Hanover data include all certified staff of whom 81 percent are teachers.

stronger: In only 3 of 45 comparisons do the grand means of Functions 7 and 8 not exceed those for Functions 1 through 5.

Comparison of CDP and Non-CDP Districts: 1988

> Hypothesis 3: Mean function ratings for non-CDP districts will exhibit considerably greater instability than for CDP districts, both in terms of the range of mean function ratings among districts, as well as less consistency among evaluators' mean ratings. Overall, mean function ratings of CDP districts will exceed those of non-CDP districts.

These predicted differences between the two types of districts derive from the fact that non-CDP districts have not experienced the broad scope of training, technical assistance, and professional incentives of the CDP districts. Accordingly, it is expected that neither teachers nor evaluators of non-CDP districts would experience the growth and improvement of their counterparts in CDP districts. This difference will result in (a) lower mean function ratings of teachers, and (b) higher standard deviations (SDs) associated with mean-ratings for non-CDP districts. More specifically, it is expected that mean function ratings for non-CDP districts would clearly not be as high as pilot district ratings in the third year, and probably not as high as second year pilot district ratings. On the other hand, the convergence of non-CDP evaluators' ratings (as measured by the standard deviation) would probably be intermediate between the SDs of the mean function ratings of the first and second years of the pilot--i.e. not as scattered or chaotic as first year ratings, but not as settled as second year ratings.

At the same time, it is expected that considerable variation will exist among the otherwise lower mean function ratings of non-CDP districts--greater variability than would be the case for CDP districts. Again, this is partly due to non-pilot evaluators not having as many opportunities or incentives to refine and standardize their skills to yield more consistent ratings. In

addition, an earlier survey study of the N.C. TPAS (Stacey et al., 1988) suggested that because of the lower stakes that accompany performance evaluation in non-CDP districts, performance ratings could well be inflated under such circumstances, or could result from an increased willingness by evaluators and/or evaluatees to negotiate ratings.

Collectively, the data summarized in Table 4 and 5 offer strong support of hypothesis 3. First, on the question of the stability of non-CDP performance ratings, the findings of Table 4 and the standard deviations displayed in Table 5 depict a much less stable pattern of ratings in these as opposed to the CDP districts. According to Table 4, the difference in mean ratings between the highest and lowest rating districts is greater for the non-CDP districts for every function, usually on an order of three to four times greater. In fact, the non-CDP districts' difference scores of approximately two rating scale points or higher exceeds the most liberal error tolerance (i.e. ± 1 rating scale points) ever considered for the TPAS. Moreover, as further support of the earlier contention that the TPAS has resulted in generally less controlled ratings of Functions 6, 7 and 8 (and especially Functions 7 and 8) than Functions 1-5, the two greatest difference scores for non-pilot districts occur for Functions 7 and 8, whereas two of the three greatest difference scores for the pilot districts also occur for these functions.

The standard deviation (SD) results of Table 5 offer further support of the relative instability of non-pilots' performance ratings. When the SDs of aggregated mean function ratings are compared for CDP and non-CDP districts, the SDs of non-pilots' ratings are greater in every instance, reflecting the lesser convergence or consistency of ratings in these units. Moreover, the convergence or SD of non-pilot districts' ratings of teacher performance tend, as predicted, to be intermediate between the SDs of the 1986

Table 4. Comparative Differences Between the Highest and Lowest Mean (M) Ratings of TPAI Functions for Career Development Versus Non-Career Development Districts for 1988

	Career Development Districts			Non-Career Development Districts		
FUNCTION	High M	Low M	Diff.	High M	Low M	Diff.
1	4.85	4.31	.54	5.20	3.20	2.00
2	4.91	4.24	.67	5.23	3.31	1.92
3	4.82	4.24	.58	5.31	3.25	2.06
4	4.89	4.28	.61	5.31	3.22	2.09
5	4.84	4.26	.58	5.30	3.34	1.96
6	4.94	4.38	.56	5.25	3.22	2.03
7	5.08	4.37	.71	5.35	3.05	2.30
8	5.18	4.55	.63	5.43	2.91	2.52

and the SDs of the 1987 pilot district mean ratings (NOTE: The reader may effect this comparison by matching the non-CDP SDs of Table 5 against the 1986 and 1987 CDP SDs of Table 1). The non-pilots' SDs are intermediate between 1986 and 1987 pilots' SDs for five of eight functions, even higher than the corresponding 1986 SD in two cases (Functions 6 and 8), and less than the corresponding 1987 SD in one case (Function 3). Thus, on the whole, whereas convergence of non-pilot evaluator ratings appears to have largely surpassed that of pilot districts in their first year--at least for the classroom-based functions, they have not evolved to the level of the pilots for their second year. Again, the greatest differences in stability of the ratings between the two types of districts occurs for two of the non-classroom-based functions--6 and 8.

Finally, relative to the second part of hypothesis 3, the "mean ratings" section of Table 5 reveals that teacher performance is rated higher in CDP than in non-CDP districts in the case of every function. That is, even though a particular non-pilot district's mean function ratings surpassed the highest of the pilot districts' mean function ratings, (See Table 4), <u>average</u> performance ratings aggregated over all pilot districts surpassed those of the non-pilot districts. Moreover, that the two types of districts vary the most in their ratings of Fs 6, 7 and 8 testifies, again, to the fact that all evaluators seem to experience the greatest difficulty in being consistent and in control of their ratings of these functions.

Table 5. A Comparison of Mean Function Ratings and Standard Deviations of Mean Ratings for Career Development (CDP) and Non-Career Development (Non-CDP) Districts[a]

	MEAN RATINGS			STANDARD DEVIATION OF RATINGS		
FUNCTION	CDPs	Non-CDPs	Diff.	CDPs	Non-CDPs	Diff.
1	4.56	4.15	.41	.892	.947	-.055
2	4.56	4.12	.44	.925	.971	-.046
3	4.51	4.25	.26	.920	.965	-.045
4	4.54	4.13	.41	.851	.897	-.046
5	4.52	4.07	.45	.849	.886	-.037
6	4.64	4.01	.63	.902	.967	-.065
7	4.76	4.12	.64	.903	.954	-.050
8	4.82	4.18	.64	.950	1.050	-.10

Ns for means and standard deviations range as follows:

	CDP Districts	Non-CDP Districts
Fs 1-5	5834-5837	2118-2123
Fs 6-8	5808-5818	2029-2033

DISCUSSION

Hypothesis 1

As stated in the Results section, this hypothesis is clearly upheld, from at least a couple of standpoints. First, the standard deviations (SDs) of aggregated mean functions ratings for the pilot districts do decline over the course of the pilot, and at a decreasing pace, for all functions. Second, borrowing on some of the results used to test Hypothesis 3 (i.e. Table 5), non-pilot districts did not exhibit SDs as low as pilot districts' SDs for any mean function ratings for 1988. This was an expected outcome, given that more improvement of evaluator performance was expected in the pilot districts.

Of course, the question can be raised as to the usefulness of the SD as a measure of inter-rater consistency, convergence or reliability. The SD was chosen as such a measure for this study because it was the best available indicator on which to compare participating districts. This is not to say that efforts to gauge inter-rater reliability had not occurred for all districts. However, such efforts at both a pilot-wide and district level for the CDP have been quite variable in purpose and methodology, thus yielding outcomes that are incomparable or inconclusive. For the non-pilot districts it not likely that reliability studies have even been attempted; such results were not solicited for this study, in any case.

Some evidence that suggests that the SD is a fairly stable measure of the consistency of ratings is offered by Tables 1 and 5. From the year-to-year SDs itemized in Table 1, it can be seen that the consistently lowest SDs occurred for mean ratings of Function 5, followed by Functions 4 and 1. Table 5 reveals this same rank order of SD-magnitude for non-pilot districts. In fact, the rank orders of SD-magnitudes for all functions is virtually the same for both groups of districts. Even allowing for non-pilots not being as

consistent overall as pilots in their mean function ratings, the fact that the rank order of rating consistency for the different functions is comparable for the two types of districts is at least indirect support for the usefulness of the SD as a measure of rating consistency.

Hypothesis 2

This major hypothesis, as is clear from examination of Tables 2 and 3, is clearly upheld at the level of the pilot-wide aggregation. The apparent contradictions occur at the level of the district, and probably illustrate the covariance of the simultaneous improvement of teachers' skills and evaluators' skills. Two school districts (Greene and Montgomery) represent out-lyers at the upper end of the scale in 1986. Both of these districts had mean ratings exceeding 4.7 on every function. Over the next two years, the mean ratings in these districts systematically declined, except for Fs 1, 6, 7 and 8 in Montgomery. Presumably teachers in these two districts were not demonstrating an erosion of skill. If this interpretation is correct, then evaluators' developing skills, defined as the ability to discriminate quality of performance, must be the source of this unpredicted change.

This speculation receives further support from the observation that in nine school districts, excluding Greene and Montgomery, there was a decline in mean ratings on one or more functions between 1986 and 1987. In only one district, however were such decreases found in more than three functions. This would suggest that some fine-tuning of raters' skill was occurring on a limited number of functions, resulting in a correction in the second year ratings. This conclusion is reinforced by the fact that, whereas 22 of the means were lower in 1987 than in 1986 (excluding Greene and Montgomery), in 1988, only four function means in three districts were lower than they had been in 1987. It is reasonable to infer, then, that the major adjustment in raters' skills

occurred between the first and second years (as already reported). Therefore, the overall improvement of means between the second and third years is indicative of real change in teachers' skills. In effect, then, it is the mean-ratings of the second year that become the baseline from which to measure future improvements.

While we cannot measure with precision how much of the change is accounted for by changes in one group or the other, we can essentially "watch" the early dominance of the change in evaluators' skills and then attribute the net gain in mean ratings between 1987 and 1988 to real growth in teachers' skills. While districts that began with hyper-inflation of ratings required all three years to come into line with other districts, most districts experienced the settling in of raters' skills fairly quickly.

The second part of Hypothesis 2 centers on the differences between Functions (Fs) 1-5 and 6-8. It should be remembered that Fs 1-5 rest on the base of the "process-product" research. Functions 6-8 are a combination of practices drawn from a variety of sources. In training, the emphasis was always placed on the first five functions. Moreover, the peer observers focused on practices in Functions 1-5. Finally, systematic efforts to standardize data collection methods for Functions 6-8 lagged behind those for Fs 1-5. For all these reasons, it is not surprising that the ratings for 6-8 are demonstrably higher than are the ratings for 1-5. Wider discretion to principals, who awarded the ratings, resulted in inflation in these functions relative to the classroom-based functions. While the same pattern of ratings (higher in Year 3 than Year 2) is found in these latter functions, the absolute mean values, especially for Functions 7 and 8 increased with greater rapidity than did the other functions.

This phenomenon suggests not only that limiting the appraisal of Fs 6-8 to principals results in higher ratings of these functions, but also that better training and involvement of all raters in appraising Fs 6-8 would probably have a depressing effect on the ratings. However, such ratings would probably also be more accurate.

Hypothesis 3

This hypothesis is also clearly upheld, in both its parts. First, as predicted by the developmental-readiness model of performance appraisal described earlier, the dependent measures for non-pilots in their first year are clearly less stable than those of pilots. This is true whether it is the range of mean-function ratings of teacher performance or the stability (i.e. SD) of evaluators' ratings that is considered. Second, as predicted by the greater investment of the pilot districts in teacher growth, the measured improvement of pilot teachers is greater than that of non-pilot teachers for the year (1988) of comparison. In fact, all of the 1988 mean-function ratings of non-pilots are lower than those of aggregated pilot districts for any year, and tend to approximate the ratings of the lower-rating pilot districts in the first year of the pilot. This latter finding, too, would fit the developmental-readiness model proposed here in that the non-pilots would be expected to be at an earlier developmental stage of effectiveness, whether it be on the part of evaluators, or teachers, or both.

Conclusion

On the whole, the results of this study tend to confirm and illuminate the improvement in performance expected of both evaluators and teachers for the first three years of a four-year pilot program. The results also bear out the developmental and interactive nature of these improvements. That is, the settling of performance appraisal ratings that appears to be a necessary phenomenon during the first two years of a CDP reflects mostly on the developing skills of evaluators. Once evaluation skills have settled, further

increases can than be regarded as more likely indicators of real teacher growth. Therefore, it is with great anticipation that the investigators look forward to data submitted for the fourth year of the pilot program. The expectation is that the simultaneous growth of evaluators and teachers will continue, and that the trend of the statistical indicators of growth suggested by this study will also be maintained.

Should such trends be maintained, the investigators believe that indicators of essential growth will then be available to serve the programatic expansion of CDPs in the future. Such indicators can minimally be used to assist in determining the readiness of candidate-districts for a CDP, as well as to monitor and track the progress of their implementation of the program.

BIBLIOGRAPHY

1. Brandt, Richard M., Duke, Daniel L.; French, Russell L., and Iwancki, F. Edward (1988.) A Review With Recommendations of the North Carolina Teacher Performance Appraisal Instrument. Raleigh, N. C.: North Carolina General Assembly.

2. Eaker, Robert E. and Huffman, James O. (1980). Helping Teachers Use Research Findings: The Consumer-Validation Process. Washington, D. C.: The National Institute of Education.

3. Gage, Nathaniel and Needels, Margaret C. (1989). "Process-Product Research on Teaching: A Review of the Criticisms" The Elementary School Journal, Vol. 89, No. 3, pp. 253-300.

4. The Group for the Study of Effective Teaching. (1983). Teaching Effectiveness Evaluation Project: Final Report. Chapel Hill, N. C.: School of Education, University of North Carolina at Chapel Hill.

5. Hall, Gene E. and Loucks, Susan F. (1983). The Concept of Innovation Configurations: An Approach To Addressing Program Adaptation. Washington, D. C.: The National Institute of Education.

6. Inman, William. (1982). Reliability of the Teacher and Principal Personnel Performance Appraisal Instruments. Raleigh, N. C.: North Carolina State Board of Education.

7. Joyce, Stephen and Showers, Beverly. (1982). "The Coaching of Teaching". Educational Leadership, Vol. 40, p. 4-10.

8. North Carolina Department of Public Instruction/Division of Personnel Relations. (1986). Outcomes of TPAS Training. Raleigh, N. C.: North Carolina Board of Education.

9. North Carolina Department of Public Instruction/Division of Personnel Relations. (1987). "Inter-rater Reliability Check". Raleigh, N. C.: Department of Public Instruction. Unpublished report.

10. Stacey, Dennis; Kuligowski, Barbara; and Holdzkom, David. (1988). "Effectiveness of the North Carolina Performance Appraisal System". Paper presented at the Annual Meeting of AERA, April 1988.

11. Wise, Arthur E.; Darling-Hammond, Linda; McLaughin, Milbrey W.; and Berstein, Harriet T. (1984). Teacher Evaluation: A Study of Effective Practices. Santa Monica, CA: The Rand Corporation.

END

U.S. Dept. of Education

Office of Educational
Research and Improvement (OERI)

ERIC

Date Filmed
August 10, 1992