APPRAISER ACCURACY UTILIZING THE TEXAS TEACHER APPRAISAL SYSTEM: A DEMOGRAPHIC ANALYSIS

DISSERTATION

Presented to the Graduate Council of the University of North Texas in Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF EDUCATION

By

Bob Evans Griggs, B.S., M.Ed.

Denton, Texas

December, 1988

Griggs, Bob Evans, Appraiser Accuracy Utilizing the Texas Teacher Appraisal System: A Demographic Analysis. Doctor of Education (Administrative Leadership), December, 1988, 146 pages, 22 tables, bibliography, 80 titles.

The purpose of this study was to determine if there are personal and demographic characteristics which can predict the most accurate teacher appraisers. The demographics were limited to the following: campus-level job assignment, employing district size, sex, race, number of years of experience as an administrator, previous level of teaching experience, and curriculum area taught by the appraiser. The 622 subjects were school administrators trained to utilize the Texas Teacher Appraisal System.

The data were analyzed using multiple linear regression. Where an independent variable was significant (.05), a follow-up ANOVA and Tukey's multiple comparison were employed.

Based on the findings of this study the following conclusions were drawn:

1. A summary data set indicated there was little evidence that any of the demographic variables was a significant predictor of accuracy in the evaluation process.

2. Six different data sets indicated that varying instructional settings and methodologies can influence

evaluator accuracy. The campus assignment, years of experience, content area taught, race, and sex of the appraisers were all identified in at least one of the exercise sets as having significance. Except for sex and race, none of the variables was found to be significant when the overall prediction equation with all demographic variables was evaluated.

3. In the prediction equations of this study the percent of variance was so minute that social significance could not be established.

4. The Texas Teacher Appraisal System can be used by appraisers with various backgrounds and experiences without a reduction of accuracy.

5. School boards can appoint appraisers with various backgrounds and experiences without a reduction of accuracy in the process.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In January, 1985, a time line was established by the Texas State Board of Education for the development of a Texas Teacher Appraisal System (TTAS). The development of a system was mandated by the 69th Legislature in House Bill 72. The Texas Teacher Appraisal System's purpose was to appraise teachers for career ladder advancement as well as to improve classroom instruction. The State Board of Education was directed to establish a system of evaluating teachers based upon observable, job-related behaviors. The legislature also instructed the Board to provide for at least two appraisals during each of two appraisal periods within the regular school year, to develop a uniform training program for appraisers of teacher performance (including uniform appraiser certification standards), and to include a teacher self-appraisal in the appraisal process.

The Texas State Board of Education instructed the Texas Education Agency staff to formulate a plan to fulfill the legislative mandate. The Texas Education Agency began by reviewing literature on teaching effectiveness. This review was followed by a survey of what other states were doing with their statewide appraisal system and a survey of what

156 Texas school districts were using in their teacher evaluation system. A job-relatedness survey (Appendix A) was designed by the Texas Education Agency. Information compiled from the returned surveys was to be used in the development of a list of teaching behaviors for inclusion in an appraisal instrument. This sample instrument was mailed to 30,000 Texas teachers who had been proportionately selected according to gender, race/ethnicity, teaching field, teaching assignment, and years of experience. Approximately 17,000 teachers returned the completed survey. A list of teaching behaviors, recommended by these educators, was compiled for a Texas appraisal instrument data base (TEA 1986, 1).

The State Board of Education Committee on Personnel formulated an instrument which was reviewed by nationally recognized experts Dr. Richard Manatt of Iowa State University, Dr. John Goodlad of the University of Washington, Brigadier General Billy Bowles, a military evaluation and staff development expert, and Dr. Lester Solomon of the Georgia State Department of Education (TEA 1986, 1).

Six Texas school districts were chosen in October of 1985 to pilot this appraisal instrument. The six school districts consisted of Grandfalls-Royalty ISD, New Boston ISD, Port Arthur ISD, Santa Rosa ISD, Seguin ISD, and Slaton ISD. These districts were selected to represent each geographical area of the state, all size classifications, and

all ethnicity groupings. These districts consisted of more than 1400 teachers and ninety TTAS appraisers (TEA 1986, 1). The pilot program was designed to develop the foundation for the instrument with regard to reliability and validity. The second purpose of the pilot program was to fine-tune the system before the TTAS was implemented statewide. To reach these objectives, the State Board of Education on September 14, 1985, awarded the pilot program contract to Performance Assessment Systems, Inc., of Athens, Georgia. This corporation was given directions to develop a training program and required materials for implementation of the Texas Teacher Appraisal System (TTAS). Although founded in 1980, the Texas Education Agency felt the corporation had an enviable amount of experience with teacher assessment instruments and could fulfill all requirements of the contract (TEA 1985, 1).

The founders of Performance Assessment Systems, Inc., William Capie (Director of Teacher Assessment Project at the University of Georgia at Athens, Georgia) and Chad Ellett (Coordinator of Research in the College of Education at Louisiana State University in Baton Rouge, Louisiana), had developed the Teacher Assessment and Development System for the Dade County Public Schools, Miami, Florida. This was a system designed to observe and evaluate the effectiveness of the 14,000 teachers in the Dade County School System. This background allowed Capie and Ellett to become a driving

force in the design of the Texas appraisal instrument (TEA 1985, 1).

The pilot program executed its mission and in February of 1986, the State Board of Education held public hearings on the appraisal instrument. These hearings allowed teachers, administrators, and professional organizations to voice their concerns about the instrument. After these hearings were conducted, the instrument went through a revision to incorporate the findings of the pilot project and reflect the comments of the hearings. With these revisions in place, the instrument was prepared for delivery to the 1,063 independent school districts in the state. However, before schools could begin using the instrument, their appraisers were required to be trained in its use.

Approximately 270 individuals were trained by the Texas Education Agency in April and May of 1986 to provide the training to all TTAS appraisers in the state. During the summer of 1986, these 270 individuals, through the twenty educational service centers, provided 43 hours of training to approximately 13,000 persons. This training provided information on statutory requirements and State Board of Education rules concerning the instrument, procedures for scoring the instrument and practice in using the appraisal instrument under simulated conditions. Each appraiser was required to demonstrate a score of 70 percent correct on a test of knowledge over the instrument and 70 percent correct

in scoring segments of instruction which had been videotaped (TEA 1986, 2).

The videotapes used for these tests were scored utilizing the TTAS instrument by a panel of educators, considered by the Texas Education Agency to be experts in teacher appraisal. This panel viewed the videotapes and established scoring standards and a scoring rationale for the taped instructional segments.

Today, after using the instrument for two years in appraising the state's teachers, the state department of education still does not have criteria upon which to select valid teacher appraisers. This study attempted to provide information on which to base these selections.

The data utilized in this study were collected from individuals trained by Region 10 Education Service Center during the summer of 1986. This training encompassed 742 subjects that represented educators who were superintendents, principals, central office administrators and teachers.

Statement of the Problem

The problem of this study is to determine if certain traits or characteristics of appraisers are related to the Texas Teacher Appraisal System accuracy-check test scores.

Purpose of the Study

The purpose of this study was to determine if there are personal and demographic characteristics which can predict teacher appraisers who will be the most accurate.

Hypotheses

The following hypotheses were tested.

1. There is no relationship between campus-level assignment and appraiser's accuracy.

2. There is no relationship between the appraiser's district size and appraiser's accuracy.

3. There is no relationship between the appraiser's sex and appraiser's accuracy.

4. There is no relationship between the appraiser's race and appraiser's accuracy.

5. There is no relationship between the appraiser's number of years experience as an administrator and appraiser's accuracy.

6. There is no relationship between the appraiser's level of teaching experience and appraiser's accuracy.

7. There is no relationship between the curriculum area taught by the appraiser and appraiser's accuracy.

Background and Significance of the Study

Since the implementation of the Texas Teacher Appraisal System, teachers and educational organizations have questioned the validity of appraisers. Many educators seemed to question what made an appraiser a valid or reliable evaluator. Because of the newness of Texas' appraisal system, these questions have not been answered by the Texas Education Agency or the State Board of Education.

Although Stodolsky discovered that nearly all school districts conducted "some type of formal evaluation of teaching performance" and an extensive body of research is available on teacher evaluation and appraisal, there is a limited amount of research on the appraiser (Stodolsky 1984, 12). The question of reliability has received the most attention from researchers looking at teacher appraisers (Brown, et al. 1967). But most researchers are hesitant to even broach this subject when reviewing the subject of teacher appraisers. Brown, states,

> Reliability can be a tricky concept. We know that reliability always refers to consistency throughout a series of measurements, and that it is usually expressed in terms of something called reliability coefficients. Rarely do we make clear what kind of consistency has been figured. Although everybody in educational research reads reliability coefficients, few seem to really understand (or care) what these mean or how they were obtained. All that matters is that they be high. Once the standard for 'highness' has been debated and denoted, then surpassed or fallen short of, what more is there to say about reliability? (Brown 1967, 11)

Regardless of the reliability of teacher appraisers or what the literature indicates about evaluation, the public desires educators to be held accountable for their product. This attitude is presented by Fuller as he states,

> The rational-bureaucratic image of management pervades private firms and many public sector programs, including administration of local schools and categorical interventions. The 'professionalization' of school administration, development of management information systems and rise of 'instructional specialists' represent earnest attempts to improve management ('just like business') and serve to symbolically increase legitimacy of schools trying to survive in an organizational landscape dominated by rationalized economic organizations (Fuller, Wood, Rapoport, and Dornbusch 1982, 16).

This attempt to make education more like business, places an added burden on the administrator responsible for the appraisal of those individuals directly on the "firing" line--teachers.

With the public's increasing belief that educational improvement hinges on upgrading the caliber of the teaching staff in the schools, the appraisal process becomes more important. (Darling-Hammond, Wise, and Pease 1983, 286). Teacher appraisals can assist a district in identifying those teachers who are "master teachers" and in identifying those teachers who need help in improving their teaching skills. Also, a good appraisal instrument and trained appraisers may be used for termination procedures. In fact, prior to the advent of the Texas Teacher Appraisal System, teacher evaluation was "conducted for a limited audience:

administrators holding responsibility for retention decisions" (Peterson 1984, 63).

According to Savage, school administrators accept good evaluation procedures more readily than teachers (Savage 1984, 14). He states that since the time of Socrates teachers have received nothing but grief at the hands of their evaluators. He adds further that administrators are not doing much to improve evaluation procedures. They should become cheerleaders for their staff and not evaluators who stand around to wag fingers at what teachers are doing wrong. Ban and Soudah say administrators should become more as assistants in developing teacher strengths than being fault-finders (Ban and Soudah 1978, 26). Administrators should attempt to instill confidence for the appraisal process in their staff.

This attitude concerning teacher evaluation and the questions raised by teachers concerning the accuracy of their appraisers, lends credence to this study. It is hoped that this research will enable teachers to have more confidence in the ability of their appraisers. This study will attempt to identify those traits and characteristics which typify a proficient appraiser.

Limitations and Assumptions of the Study

The following are limitations and assumptions of this study.

1. It is assumed that the Texas Teacher Appraisal Instrument is a valid appraisal tool in teacher evaluation.

2. The study assumes the scores of individuals in the sample are indicative of their best efforts.

3. The study is limited by the validity with which the panel of experts scored the videotapes used in the training of appraisers.

Definition of Terms

The following terms have been defined for the purposes of this study.

Accuracy.--The State Board of Education arbitrarily established 70 percent correct as the mark for obtaining minimum accuracy on a written test covering the Texas Teacher Appraisal System rules and 70 percent correct on the use of the Texas Teacher Appraisal Instrument. In this paper, accuracy will only be a reference to teacher appraisers and not to teachers.

Appraiser.--An appraiser is an individual trained in a uniform program approved by the central agency and meeting the performance standard set by the State Board of Education of Texas. This individual was "certified" as a TTAS appraiser.

ILT.--Instructional Leadership Training is 36 hours of specialized curriculum required by the State Board of

Education before an individual can take the Texas Teacher Appraisal System training.

Observation/Evaluation Record.--The observation/evaluation record is the recording instrument used in the Texas Teacher Appraisal System to reward or deny credit for each indicator established by the System.

Region 10 Education Service Center.--The Region 10 Education Service Center is one of twenty regional education service centers established by the Texas legislature to assist local education agencies (school districts). Region 10 Education Service Center consists of an eight county area of North Texas with eighty-one independent school districts.

Significant Relationship.--A significant relationship exists if a relationship between or among variables evaluated statistically to be non-zero with probability of error less than .05.

Texas Teacher Appraisal Instrument.--The teacher evaluation instrument adopted by the Texas State Board of Education to be used in evaluating the public school teachers of Texas.

Texas Teacher Appraisal System.--The teacher evaluation system mandated by Texas' 69th Legislature in House Bill 72, to appraise teachers for career ladder purposes and improve classroom instruction.

REFERENCE LIST

Ban, John R., and John R. Soudah. 1978. A new model for professionalizing teacher evaluation. Peabody Journal of Education 56 (October): 24-32.

Brown, Bob Burton, William Mendenhall, and Robert Beaver. 1967. The reliability of observations of teachers' classroom behavior. ERIC, ED 011 520.

Darling-Hammond, Linda, Arthur E. Wise, and Sara R. Pease. 1983. Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research 53 (Fall): 285-327.

Fuller, Bruce, Ken Wood, Tamar Rapoport, and Sanford M. Dornbusch. 1982. The organizational context of individual efficacy. Review of Educational Research 52 (Spring): 7-30.

Peterson, Ken. 1984. Methodological problems in teacher evaluation. Journal of Research and Development in Education 17 (Summer): 62-70.

Savage, John G. 1984. Better ways to evaluate teachers. North Central Association Quarterly 59 (Summer): 14-17.

Stodolsky, Susan S. 1984. Teacher evaluation: The limits of looking. Educational Researcher 13 (November): 11-17.

Texas Education Agency. 1985. Introductory information. Mimeographed memo. Austin, Texas: Texas Education Agency.

Texas Education Agency. 1986. Texas teacher appraisal system appraisal manual. Austin, Texas: Texas Education Agency.

CHAPTER 2

REVIEW OF RELATED LITERATURE

In reviewing the literature on teacher appraisers, it became apparent that there was a lack of research which investigated characteristics of teacher appraisers in relationship to their accuracy in the evaluation process. However, with the public holding educators more accountable for their products, the appraiser comes into the limelight more often. This review of literature will describe five areas which influence teacher appraisers in the evaluation process. These areas are (1) the basis for teacher evaluation, (2) a description of the evaluation process, (3) the types, intents, and designs of appraisal instruments, (4) the training of teacher appraisers, and (5) the reliability and validity of teacher evaluation.

Basis for Teacher Evaluation

Wood and Withall disclosed that teachers on the North American continent have been evaluated since the 17th century when influential community leaders made quick visits to appraise the school's teacher (Withall and Wood 1979, 55). Teacher appraisal evolved into a formal process about 1915 when educators turned to the multifactor teacher rating

scale after it was heralded in the yearbook of the National Society for the Study of Education (Coker 1987, 242). The evolution continued until it has reached its contemporary stage. This level is described by Sapone as follows:

> Today, as never before, the public is demanding educational and fiscal accountability. The message is clear that new dollars for education will not be forthcoming until the taxpayer's confidence is restored in what is currently happening in schools and until they can expect a reasonable return from additional investment.
>
> It is, therefore, incumbent upon educational leaders to develop specific appraisal and evaluation systems for assessing the process and products of education. What citizens are requiring is "proof" of increased effectiveness of teacher and administrative performances as they influence pupil growth and school achievement (Sapone 1980, 44).

With the public demanding this type of accountability from the educational community, it is paramount that effective evaluation systems be utilized in today's schools. But what is evaluation? Tracey describes evaluation as "an aspect of management control. It is a systematic means of determining the extent to which educational plans have been carried out and programmed objectives have been achieved" (Tracey 1978, 240). Another author, Bolton, says "Evaluation is a value judgment made late in a dynamic process . . ." (Bolton 1973, 97). Knapp defines teacher evaluation as ". . . any formalized appraisal . . . with intended consequences for <u>individual teachers</u>, such as improving their teaching or determining their position . . ." (Knapp 1982, 1). To summarize the literature on evaluation, a definition of evaluation for

the purposes of this paper can be stated as: The review of all aspects in the instructional process which enhances or creates change in a classroom teacher's performance.

Society has good reason to make demands on its educational system. Between 1971 and 1978 more money was spent on education in this country than on the national defense. With the largest expenditure in educational budgets consisting of teacher salaries (Gage 1978, 13), the appraisal of the performance for these expenditures gains momentum. Seeley emphasized the case even more firmly when he stated the next few years will either bring ". . . conflict and recrimination between parents and teachers . . . ," or we will recognize that a problem exists and will work ". . . to meet the new demands . . ." (Seeley 1979, 249).

With state legislatures, commissions, advocate groups, teacher unions, and school boards searching for solutions to the public's demands in education, McLaughlin argues that teacher evaluation can be the tool which achieves the sought after school improvements (McLaughlin 1984, 193). White, et al. concur with McLaughlin. Their research indicates that principals, when trained in the use of the Teaching Performance Observation Instrument, can reach reliability approximating .82 on videotaped teaching scenarios (White, Wyne, Stuck, and Coop 1987, 94). Admitting teacher evaluation is a difficult undertaking, Wise concludes that it can be an aid in determining the ability of new teachers, assist with

improvement of all teachers and be a signal when teachers are no longer being productive in their assignment (Wise, Darling-Hammond, McLaughlin, and Bernstein 1985, 62).

The legal system of our country has also been supportive of teacher evaluations, in general. Bridges reports that regardless of the empirical evidence of teacher evaluation's low reliability and validity, the courts have placed great weight on the documented classroom observations by teacher appraisers. He quotes Fowler v. Young, et al., Board of Education, 65 N.W. 2d 399 as proof of his argument:

> Teaching is an art as well as a profession and requires a large amount of preparation in order to qualify one in that profession. The ordinary layman is not well versed in that art, neither is he in a position to measure the necessary qualifications required for the teacher of today. In our judgment this information can be imparted by one who is versed and alert in the profession and aware of the qualifications required. . . . We think the principal with the years of experience possessed by him can be classed properly as an expert in the teaching profession, and is in a similar position as a doctor in the medical profession (Bridges 1985, 62).

However, another issue surrounds the question of teacher evaluation--the accuracy of teacher appraisers and evaluation instruments. Beckham questions if teacher appraisal can be conducted in a "consistent, measurable, objective and meaningful way." He also states that educational researchers have not reached consensus on what elements form effective teaching practices and writes that an instrument or model does not exist which is not open to criticism (Beckham 1981, 1-2).

Peterson asserts that the quality of learning in the classroom and its relationship to teacher evaluation is not clear. He further states that finances and educational resources have not been utilized in improving teacher evaluation (Peterson 1984, 62). Shavelson and Dempsey in their 1976 study of teaching behavior, and the 1978 study of Erlich and Shavelson, revealed that the reliability of specific teacher behaviors which are assessed by teacher observers is low or might be nonexistent (Magoon 1979, 13).

Coker, in his study of teacher effectiveness as it relates to the validity of principal judgments in math and reading classes, discovered a mean correlation between these judgments and achievement gains of average students to be only .20. He says these findings indicate teachers are being evaluated all across the nation using "methods that are not detectably better than chance." He concludes his findings by stating that decisions concerning teachers are being made on judgments by principals which are only a shade better than if "decided by lottery" (Coker 1985, 40). Donald Medley, in an article he co-authored with Coker, noted that studies conducted by Anderson, 1954; Barr, Torgerson, Johnson, Lyon, & Walvoord, 1935; Brookover, 1945; Gotham, 1945; Hellfritsch, 1945; Hill, 1921; Jayne, 1945; Jones, 1946; LaDuke, 1945; Lins, 1946; and Medley and Mitzel, 1959; concluded "that the correlations between the average principal's ratings of teacher performance and

direct measures of teacher effectiveness were near zero." Medley and Coker asserted these early findings did not change teacher evaluation through the present era. In fact, according to their research, contemporary teacher evaluation decisions which are being made on judgments are "only slightly more accurate than they would be if they were based on pure chance" (Medley and Coker 1987, 242-243).

Regardless of which research is correct, one of the most important reasons for developing sound, realistic, and credible evaluations for the teaching vocation is the need to nurture the educational community into a recognized profession. In recent years, several reports written by national task forces on education have emphasized the need to elevate teaching "to a more respected, more responsible, more rewarding and better rewarded occupation" and to create an atmosphere for the professionalization of teaching (Shulman 1987, 3). Even the 1988 Democratic nominee for the President of the United States of America, Michael S. Dukakis, voiced his stance on teachers during his nomination acceptance speech when he said, ". . . make teaching a valued and honored profession--once again" (Dukakis 1988, July 21).

Tommorrow's Teachers, the Holmes Group's first publication, and the Carnegie Task Force's A Nation Prepared: Teachers for the 21st Century suggested one reason for problems in education is the American public's lack of

recognition for teachers as professionals (Hampel 1986, 55). Shulman pointed out that the Holmes Group and the Carnegie Task Force reported a broad knowledge base for teaching. He noted the reports outline a collection of skills, technology, knowledge, understanding, ethics, and responsibility for teaching; and educators possess the methods to communicate them (Shulman 1987, 4).

Shulman, in further defense of the professionalization of education, states there is adequate empirical research of teaching effectiveness to justify the profession. In recent years, the traditional psychological research conducted by Brophy and Good (1986), Gage (1986), and Rosenshine and Stevens (1986) has added to the empirical research which already existed (Shulman 1987, 6). McLaughlin argued that this empirical research base, combined with teacher evaluation which uses specific, concrete terms in diagnosing instruction, supports the established norms for a profession (McLaughlin 1984, 199). White and his co-authors concluded that the anchoring of teaching on empirical based research is the most important step in moving teaching to a recognized profession. He contends this step will provide the foundation of knowledge that exists in other professions. Also, he predicts this approach will provide a "defensible, scientific basis for guiding professional practice in teaching and teacher training (White, Wyne, Stuck, and Coop 1987, 95). With the empirical based research as a

foundation, teaching becomes a learned profession (Shulman 1987, 9).

In addition to the empirical research based argument for the professionalization of teaching, another point of contention has been espoused by many writers--the artistry required of educators. Gage compares the teaching profession with the medical profession. He states the twentieth century medical profession has a scientific basis. This scientific basis consists of thousands upon thousands of different variables but a physician must use his artistry as he arranges these variables to the benefit of his patients. Gage further illustrates his point with a comparison to the engineering profession. He points out that engineers have a strong background in physics and chemistry. Yet, when an engineer is solving a problem, he relies on artistry to manipulate his scientific foundation. Gage argues that an analogy exists between medicine, engineering, and teaching. He contends the professions themselves are not a science but they each utilize a scientific basis and artistry to achieve their goals (Gage 1978, 17-18). Moxley supports this contention by writing that a teacher is not only a facilitator of learning but also adapts to the individual child and becomes a designer and engineer (Moxley 1978, 65). Wise has further substantiated the artistry of teaching by stating, "A professional teacher . . . has sufficient knowledge . . . to make decisions about instructional content and delivery

for different students . . . ascertain their clients' needs and determine how to meet them" (Wise and Darling-Hammond 1985, 31).

After a professional educator practices his artistry, should he be judged, and if so, by whom? McLaughlin tells us an attorney can evaluate success by the number of cases he has won or lost, a dentist is successful if he assists his patients in maintaining good teeth, an engineer can observe the safety of a bridge, and a physician can witness the health of his patients. A teacher's success is based upon the long-term change in a student's behavior (McLaughlin 1984, 196).

Soar contends a parallel can be drawn between the teaching profession and the medical profession. He claims that if the only measure of success in the medical profession was mortality rate of a doctor, then physicians would be hesitant to take terminally ill patients. However, if the judgment was based upon his prescribing a treatment which is the most effective, the evaluation becomes equitable. Soar argues the teaching profession is similar to the medical profession in this regard. He points out a teacher should be evaluated on what he is doing in the classroom and not on the outcome of what he does (Soar 1975, 209).

Yet the public demands perfection from the teaching profession and does not accept the argument that some

students cannot learn all things. No profession guarantees its results. In fact, one universally accepted characteristic of any profession is that the professional is not allowed to guarantee his results. A physician can not guarantee all patients will recover, a lawyer can not win every case, and a dentist can not guarantee he will never lose a tooth. Professionals accept all cases, as teachers do, regardless of the likelihood of success. Therefore, the public has a misconception that the professional educator should be held accountable for how much a student learns (Medley 1982, 10). There are so many internalized rules, exemplars, and knowledge bases in a teacher's performance that only another professional educator who has acquired this knowledge base can conduct a good evaluation on a fellow professional (House 1980, 253). Or as Barber wrote over twenty years ago, "An essential attribute of professional role is . . . application of the body of generalized knowledge in which they alone are expert (Barber 1965, 18).

Every recognized profession uses some method to evaluate the expertise of its members. Some professions use written examinations and others use observational evaluations. Since teaching is the largest of the professions, some would argue it is not realistic to have observational evaluations (Lareau 1986, 555). Others make a case for nation-wide examinations for teachers before being "licensed" to instruct children in the classroom

environment. In fact, many leaders of national teacher organizations have called for the development of a professional examination which would test prospective teachers on subject-matter knowledge and pedagogy (Lareau 1986, 554). Albert Shanker, in a speech to the National Press Club, advocated the administration of a professional examination for teachers (Shanker 1985, January 29). According to Shulman, teaching should follow the model of other professions by utilizing national and state certification procedures which are demanding (Shulman 1987, 20).

English summarizes the question of examinations and evaluations by claiming that "compared to other professions, schools lack systematic peer support systems" (English 1985, 34). McLaughlin argued for evaluations in another way when he wrote, "...while self-reflection lies at the heart of professionalism, self-monitoring and assessment are difficult for teachers to carry out" (McLaughlin 1984, 196). Natriello presented the case in a different way by saying:

> Those who oppose the use of teacher evaluation because they perceive it as a control mechanism injurious to the professional autonomy of teachers might consider the positive effects of such practices in enhancing teachers' control over their teaching tasks (Natriello 1984, 593).

Kauchak and his co-authors were even more explicit when they claimed, "Professionals exert control over the way that their performance is evaluated; workers do not" (Kauchak, Peterson, and Driscoll 1985, 37).

It becomes obvious from this review that evaluation in the field of education is based upon two premises: the need to meet the accountability expectations of the public and to encourage the professionalization of teaching.

The Evaluation Process

In the past fifteen years, many laws requiring the appraisal of teaching performance have been passed. Only six states required teacher evaluation before 1971. By 1983, half of the states in our union required a formal evaluation process (Wuhs 1983, 28). And in 1988, only a handful of states do not have some form of legislation which mandates teacher evaluation. Within the last five years, the emphasis on teacher evaluation has been accelerated by the publication of two major reports: "A Nation at Risk: The Imperative for Educational Reform" and "Action for Excellence."

In April 1983, the National Commission on Excellence in Education published its report, "A Nation at Risk: The Imperative for Educational Reform", and several of its key points concerned the evaluation of teachers. The report concluded:

> Salary, promotion, tenure and retention decisions should be tied to an effective evaluation system that includes peer review so that superior teachers can be rewarded, average ones encouraged, and poor ones either improved or terminated (National Commission on Excellence in Education 1983, 30).

Immediately after this report was published, the Task Force on Education for Economic Growth, Education Commission of the States, released its report--"Action for Excellence." This report also emphasized the need for teacher evaluation when it proposed:

> . . . put in place, as soon as possible, systems for fairly and objectively measuring the effectiveness of teachers and rewarding outstanding performance. . . . Ineffective teachers-those who fall short repeatedly in fair and objective evaluations-should, in due course and with due process, be dismissed (Task Force on Education for Economic Growth 1983, 39).

These reports made sense to the general public since most are products of the public education system. Brown rationalized this dilemma by observing that teachers don't always instruct their students in a manner they know they should (Brown 1968, 8). He further stated that even the best teachers find themselves in this predicament. They have trouble implementing into practice what they know will work (Brown 1968, 9). If teachers do not implement what they know will work, what happens to them and their students? Tracey argues that teachers do not reach the "bottom line." He says the real bottom line in education, as a parallel to business, is the "quality of the product--the knowledge, skills, attitudes, competencies, and potential of graduates." However, he emphasizes there are immediate ways to evaluate the effectiveness of teachers--teacher evaluations (Tracey 1978, 240).

Cruickshank makes a case for teacher evaluation when he emphasizes the need of principals, supervisors and superintendents "to know which teachers are good teachers." They need this information before making a recommendation to their school board concerning hiring, rewards, tenure, and dismissals of teachers (Cruickshank 1986, 81).

Darling-Hammond is even more emphatic in her writing. She claims that if a school district values teaching as a profession and has prioritized outstanding instruction as a valued goal, then personnel evaluation plays a critical role (Darling-Hammond 1986, 531). McLaughlin has expressed it in another way when she stated that teacher evaluation can be a most valuable school improvement technique because it addresses a teacher's sense of professionalism (McLaughlin 1984, 204). Pigford suggests that the way teachers regard evaluation is more important than how the institution regards it. She points out that evaluation can improve teacher performance if designed and used correctly (Pigford 1987, 142).

It would seem from the writings of these authors that there are several purposes to teacher evaluation and not just one global reason for the act. Bolton suggests there are multiple reasons for teacher evaluation. He claims the purposes can be as different as an individual teacher's desire to improve or the school board's desire to meet the public's demands for accountability. He presents the

argument that the first consideration in teacher evaluation is defining the purpose for the evaluation program. Should evaluation assess the instructional program, improve teaching, reward outstanding teachers, or provide the information for teacher growth and development? However, instead of pointing to any one of these areas as the most important aspect of evaluation, he concludes by indicating that a general agreement exists among educators that the purpose of teacher evaluation is to improve instruction (Bolton 1973, 98-99).

Buttram and Wilson disagree with this contention. They claim evaluation is not used to improve effective instructional practices but to document contract terminations, tenure decisions, or salary increments (Buttram and Wilson 1987, 5). Barth suggests the purpose for teacher evaluation is the maintenance of authority and control by principals. Some administrators use it as a technique to break down barriers and show friendship and still others utilize teacher evaluation as a means to display their expertise (Barth 1979, 75).

Knapp's research indicated there are two broad classes of purposes for teacher evaluation: formative and summative. While summative evaluation should be an indicator of a teacher's future role in the district, he says formative evaluation is intended to change a teacher's behavior patterns in the future. He points out that some researchers

will include a third purpose, referred to as diagnostic or predictive, which assesses present capabilities or needs for the planning of future decisions for staff development. But Knapp points out that the third classification can easily be subsumed into either of the first two divisions (Knapp 1982, 3). Knapp also writes most practitioners prefer the two purposes be separated. He says this camp argues that summative evaluation generates a defensive behavior and insecurity in teachers. The philosophy of a formative evaluation only is espoused by the "clinical supervision" approach to evaluation which encourages a supportive role between the principal and teacher (Knapp 1982, 7).

Peterson studied the research-based approach to teacher evaluation. His research, which agrees with Knapp's writings, indicates teacher evaluation should be divided into two separate but related dimensions--summative and formative teacher evaluation. He wrote that improving teacher performance is the purpose of formative evaluation while summative teacher evaluation is a judgement on the performance of a teacher. He also points out that summative evaluations normally use a type of rating scale such as the Lickert scale to rate "how well" a teacher performs. According to Peterson, a summative evaluation, "by law, must be based on a representative sample of teacher performance" (Peterson 1983, 6). In a latter publication, Peterson wrote that quite often the summative rating scale is the only

instrument used in teacher evaluation, and then, they are used for quality judgments. He suggests that when teacher improvement is the desired outcome for teacher evaluation, schools most often use the formative evaluation technique. Peterson adds three methods to the formative purpose for teacher evaluation. He classifies these into the naturalistic inquiry, sign systems, and category systems. According to him, the naturalistic inquiry system uses no predetermined vocabulary and the observer records whatever they choose to write down. Sign systems record a number of items which research indicates are effective teaching practices, and the observer simply checks them off if observed. And last, the category systems is designed for sequence and frequency and is limited by the number of items that can be observed because it is used on a timed basis (Peterson and Peterson 1984, 42).

Darling-Hammond, Wise and Pease refined the purposes of teacher evaluation even further. In their findings, they divide evaluation into four basic purposes: individual staff development, individual personnel decisions, school improvement, and school status decisions. They expressed the need for different processes and methods of evaluation when the focus is to be directed toward the individual versus the organization. According to them, a teacher evaluation system which insists on utilizing one process to evaluate

all four purposes will be unsatisfactory (Darling-Hammond, Wise, and Pease 1983, 302).

Wise, Darling-Hammond, McLaughlin, and Bernstein reported the same four purposes of evaluation. They further testified that teacher evaluation does not need to apply only to an individual teacher or to schools but is conducive to both large groups of teachers or small groups of teachers. The evaluation process, according to their writing, can also represent improvement based on the group process rather than the individual teacher (Wise, Darling-Hammond, McLaughlin, and Bernstein 1985, 68).

In 1984, the Rand Corporation published a report which became very influential within the education community. This report stated teacher evaluation served two purposes: accountability and improvement (American School Board Journal 1985, 25). With the Rand Corporations report and with the above review of the evaluation process, it becomes clear that the two most important purposes of teacher evaluation is for self improvement of a district's teachers or for decisions dealing with the continuation or assignment of its teachers.

Bolton says the improvement of instruction and a teacher's self-improvement are closely related in teacher evaluation. In fact, he suggests that nearly all teacher evaluation systems are based upon the premise that all teachers have a desire for individual growth and development (Bolton

1973, 101). Laing compared evaluation in a different way. He indicated that in most evaluations a principal conducts, dismissal or retention is not the question. He suggests these evaluations are an opportunity for principals to assist teachers in becoming better instructors (Laing 1986, 93).

Roy, a high school principal in Pennsylvania, wrote that if teacher evaluation is to make an impact on student achievement, teacher evaluation must be based upon the assumption that teachers have a desire to improve. He claimed successful evaluation techniques should include a philosophy that will assist in the individual growth and self-improvement of teachers. He predicted that significant student achievement gains will occur if teachers are assisted in improving their existing skills. He urges evaluators to focus on helping their good teachers become better and thereby improving achievement (Roy 1979, 276).

The findings of McLaughlin lends support to the concept of self-improvement of teachers through teacher evaluation. She proposes that evaluation systems which provide "specific, detailed, and believable information about classroom performance can engage teacher commitment to growth and enthusiasm for learning new skills" (McLaughlin 1984, 199).

While self-improvement is one of the most written about aspects of teacher evaluation, the use of teacher evaluation in the dismissal or reassignment process would appear to be

the foremost reason for appraisal in past years. Brieschke tells us:

> . . . it is the accumulation of small mistakes, such as occasional ineptness, laziness, unpreparedness, poor judgments, etc., coupled with low commitment and morale among teachers which poses the most pernicious threat to our nation's schoolchildren.
>
> One of the burdens of the principalship is the identification of educational mistakes among teachers and the development of mechanisms for addressing these mistakes (Brieschke 1986, 249).

However, Bridges reports that neither the National Education Association nor the American Federation of Teachers has taken a stance on the definition of incompetency. He also indicates they are very unlikely to develop an organizational definition of incompetency in the future. Therefore, the individuals who actually evaluate teachers, school administrators, must define incompetency. When school administrators define incompetency, they think of failure, and failure takes one of the following forms: technical failure, bureaucratic failure, ethical failure, productive failure, or personal failure. Fortunately for those educators who are being evaluated, most appraisers use discretion and a great deference in their definition of failure and incompetency (Bridges 1985, 58-59). Pellicer and Hendrix write that appraisers use a remediation process between recognizing incompetence and the formal dismissal in which evaluation can play both the summative role and be a diagnostic tool (Pellicer and Hendrix 1980, 61).

Wise, Darling-Hammond, McLaughlin, and Bernstein found in their research that most districts utilize teacher evaluation to make personnel decisions. They indicated most local education agencies use evaluation to dismiss non-tenured staff members. They report that, in states with particularly difficult laws concerning the termination of tenured staff members, few districts used the system to dismiss tenured teachers. They did discover, however, that in those particular states, administrators complete thorough evaluations of beginning teachers. These authors also claimed that most districts use evaluation in "counseling out" teachers who should not be in the classroom (Wise, Darling-Hammond, McLaughlin, and Bernstein 1985, 76). Darling-Hammond, in a later article, suggested that evaluation for minimal competence incorporates within the evaluation instrument those teaching behaviors possessed and exhibited by all teachers except the incompetent (Darling-Hammond 1986, 534).

The modification of a teacher's assignment is one purpose of teacher evaluation, according to the writings of Bolton. The modifications can include dismissal, as well as promotion, and reduction or increase of teaching load. He expressed that when teacher evaluation emphasized the removal of the ineffective and the weak, teacher moral seems to suffer (Bolton 1973, 100).

English stated, "Most traditional approaches to teacher evaluation are oriented toward inspection rather than growth" (English 1985, 34). Bird concurs with English when he reported most teacher evaluation systems, based on observation, "are designed more to correct incompetence than to foster competence" (Bird and Little 1986, 494).

McLaughlin discovered the same aspects which assisted competent teachers to view teacher evaluation as a self-improvement tool also encouraged incompetent teachers to leave the profession. She indicated those teachers who are ill-suited for teaching, and who continue to have problems in the classroom after being afforded remediation, usually seek another vocation when faced with concrete, detailed documentation of their incompetence (McLaughlin 1984, 199).

Although the removal of teachers who are harming school children is critical, Roy discloses that fewer than 5 percent of our nation's 1.9 million teachers would be considered incompetent. And he argues it is ridiculous to design complete evaluation systems to identify only 5 percent of our teachers and then use the same system on the other 95 percent (Roy 1979, 275).

Another concern of teacher evaluators in the dismissal of incompetent teachers is the legal aspect. With proper documentation and a good teacher evaluation system, their concern would seem to be unfounded. Bridges claims that our nation's judges accept an evaluators definition of

incompetency "without question," if the evaluator has provided his teachers with the criteria for success and has given the teacher specific situations in which their performance has not met the criteria (Bridges 1985, 59).

The findings of Bolton lend support to Bridges conclusions. Bolton says, "From a legal standpoint, protection of both individuals and the school organization is an important purpose of evaluation...evaluation is essential for legal reasons-if for no others." Since a school district's board of trustees is held accountable for the type of system it operates, they must have an evaluation system in place to protect themselves. Bolton has presented additional legal aspects of teacher evaluation systems by reporting that a good system also protects teachers against unjust charges; therefore, assisting both the evaluator and the one being evaluated (Bolton 1973, 100).

Peterson takes another approach to the legal issues involved with teacher evaluation. He reports school administrators should strive to develop teacher evaluation systems which can be used in dismissing ineffective teachers or provide evidence that is defensible for a teacher's retention. He points out systems which meet all legal requirements also have evaluators who are knowledgeable in due process, inference, reliability, validity, and discrimination. He claims that only systems which are based on job

related criteria can be legally applied to a teaching staff (Peterson 1983, 7-8).

According to Beckham, the assessment of educational quality has become linked to teacher evaluation legal issues. He reasons that intervention of the courts can be expected in cases of demotion, reassignment, promotion, grants of tenure or continuing contracts, and withholding of salary increments if school administrators do not use "reasoned, ascertainable standards" in making these decisions (Beckham 1981, 3). However, a well developed teacher evaluation process which follows a standard process for all teachers can eliminate the embarrassment of losing a court case.

Types, Designs, and Intents of Appraisal Instruments

According to Emmer and Peck, the proliferation of systematic observation instruments has assisted the study of classroom behavior. They reported the 17-volume anthology written by Simon and Boyer in 1970 contained over 90 systems. With this type of production, however, they are concerned many unnecessary systems could be developed, and the identification and measurement of the major dimensions of classroom instruction could be impeded (Emmer and Peck 1973, 223). As recently as 1985, Wise, Darling-Hammond, McLaughlin, and Bernstein, writing for the Rand Corporation, suggested that very few school districts in our nation have

"highly developed teacher evaluation systems" (Wise, Darling-Hammond, McLaughlin, and Bernstein 1985, 63). This would appear to be a surprising discovery given the availability of instruments and the legislated requirements for teacher evaluation. Yet, we can sympathize with school boards and administrators in this decision, given the number of instruments available to them.

Competency-based teacher evaluation; outcome-based evaluation; Tylerian model; accreditation model; management-system model; goal-free model; the Bedford, Kalamazoo, Toledo, Salt Lake City and New Hampton evaluation systems; McGreath's Exemplary System; Performance Assessment Record for Teachers; and others have all been heralded as "The" model to change education. But perhaps the best known of this group would be Manatt's "Mutual Benefit Evaluation," Redfern's "Management by Objective Evaluation," Hunter's "Clinical Supervision Model," and then the two models which all others seem to include: "Discrepancy Model" and "Emergent Model." Regardless of the number of systems or models available, Peterson reports, ". . . surprisingly little educational talent and few resources have gone into the important problem of improving teacher evaluation" (Peterson 1984, 62). Darling-Hammond found where state-developed instruments have been required by legislation, the states are adopting "objective," low-inference evaluation instruments (Darling-Hammond 1986, 535).

To have a better understanding of the major evaluation systems, a review of what the literature records is required. Peterson tells us that the most prevalent model in teacher evaluation today is the "Discrepancy Evaluation Model." This model combines a description of ideal teaching characteristics with the comparison of the actual teaching scenario. Those teachers which most closely correspond with the description of ideal teaching are identified as "the best." This is the most common form of teacher evaluation and is usually identifiable by the principal's observation of a teacher while using a checklist of desirable activities (Ken Peterson 1984, 63). The "Emergent Model" uses a large amount of documentation in its ratings. The documentation can include student gains, teacher development progress or effective processes. The documentation is subject to value judgments and "what makes one teacher meritorious or deficient may not apply in the case of another." The key to an "Emergent Model" is the understanding that ratings are context dependent, and that good teaching occurs in a number of non-mutually exclusive forms. Individual teachers are judged on their attainments within a specific setting (Ken Peterson 1984, 64).

According to Darling-Hammond, Wise and Pease, Manatt's "Mutual Benefit Evaluation" model and Redfern's "Management by Objectives Evaluation" model only have one major difference: the step of the process in which the teacher is

included. Both models include teachers in the evaluation process, goal setting, and teaching standards and criteria which are centralized. Manatt describes his process as one to improve the teacher's performance rather than identify incompetent teachers. Redfern's model evolved from the business community although "using behavioral objectives to measure teacher effectiveness was proposed as early as mid-1920 by Franklin Bobbitt of the University of Chicago" (Johnston and Yeakey 1979, 20). The evaluator establishes the learning goals and responsibilities of the teacher as they do in the Manatt model. The critical difference in the two models happens before the evaluation occurs. Redfern's model requires the appraiser and teacher to collectively describe individual objectives, measurable progress indicators, and an action plan while Manatt's model has limited teacher involvement at this stage. Also, although both models are "results-oriented," the Redfern model allows the teacher to provide more input while the Manatt model requires decisions to be made by the supervisor (Darling-Hammond, Wise, and Pease 1983, 309-310).

Hunter's "Clinical Supervision" model is very similar in structure to the Manatt and Redfern models. However, it relies more heavily on communication and dual planning between the teacher and evaluator to establish performance goals. The process is also less defined than the other two models. The clinical supervision model is very time

consuming and is questionable for use in the establishment of teacher incompetence (Darling-Hammond, Wise, and Pease 1983, 311).

The design of teacher evaluation has been going through considerable change in the past 10 years. Research for Better Schools discovered that progressive schools are designing their evaluation systems around the effective teacher practices research, training their evaluators more thoroughly, making administrators more accountable for their evaluations, using results from the evaluation process to develop staff development, and allowing teachers to become active partners in evaluation (Buttram and Wilson 1987, 5).

Teacher evaluation designs have been evolving rapidly in the last ten years. By 1981, 28 states and the District of Columbia had legislated designs for teacher evaluation. With the emphasis by legislatures on improving quality, Beckham wrote:

> . . . development of evaluation . . . rests exclusively with local school district governing boards in eight states. Local boards must seek the assistance of a personnel advisory committee in Arkansas...Oregon and Connecticut mandate the participation of teacher representatives and lay citizens in the development. . . . In Pennsylvania, the Department of Public Instruction develops or must approve the rating system . . . while in Arizona, California, Florida, Hawaii, New Mexico and South Carolina, the state board is responsible for evaluation standards (Beckham 1981, 50-51).

Because of these legislated actions, the design dilemma becomes more complicated. However, educators agree the design should address the district's desired problem or

purpose. Bolton said in designing evaluation instruments, a rule of thumb is to "select the instrument that best fits your purpose, i.e., identify the measurement techniques and strategies that provide the data desired" (Bolton 1973, 111). Wise and his co-authors described it as follows, "Clearly, the design of teacher evaluation systems depends critically on educational goals" (Wise, Darling-Hammond, McLaughlin, and Bernstein 1985, 67). Regardless of the type of instrument utilized, the design must center around the intent or purpose of the evaluation.

The intent of an evaluation system sometimes becomes lost in the emotions of the act itself. Some educators are very vocal that teacher evaluation should be only for teacher improvement. Bolton argues that teacher evaluation is part of a larger effort to evaluate a school's total program. He states teacher evaluation's intent should include changes in curriculum, instructional material availability, building design, and student groupings (Bolton 1973, 127). Other writers are as adamant with their arguments. Wise, Darling-Hammond, McLaughlin, and Bernstein in their writings for the Rand Corporation said, "The primary goal of teacher evaluation is the improvement of individual and collective teaching performance in schools" (Wise, Darling-Hammond, McLaughlin, and Bernstein 1985, 69). Brown and Webb summarize the problem of dealing with the intent of a teacher evaluation system by stating, "No matter how highly

pedigreed the system, if it does not measure the salient features of the program, it will provide meaningless information" (Brown and Webb 1975, 11).

For teacher evaluation to become meaningful for education in this country, both financial and human resources must be expended on the types, intents, and designs of the appraisal instruments used in the evaluation of educational problems. Without these expenditures, teacher evaluation is doomed to become another passing fancy of education that is meaningless in the improvement of the instructional process.

Training of Teacher Appraisers

With teacher evaluation becoming a major responsibility of most principals, they need to be given another piece of equipment to place in their tool box--appraiser training. Faast feels that appraiser training is the key to the entire process of teacher evaluation (Faast 1984, 128). Faast receives support from McLaughlin (1984); White, Wyne, Stuck, and Coop (1987); Webb and Brown (1969); Reilkoff (1981); Berliner (1987); Peterson and Peterson (1984); and Wise, Darling-Hammond, McLaughlin, and Bernstein (1985). According to Peterson and Peterson, those individuals who evaluate must be trained. They stated it more explicitly by writing, "Training is the key to successful application of a well-designed evaluation system" (Peterson and Peterson 1984, 43-44).

In an interview with Berliner in 1987, Brandt discovered that Berliner felt very strongly about training for those who judge teachers. Berliner is quoted as saying, "Judging teaching is absolutely no different from judging figure skating, poultry, potatoes, or cows. Each involves making complex decisions with a good deal of subjectivity." He adds that the difference is in the amount of practice. He says it takes ten years to become a diving judge, fifteen years to be a skating judge at the Olympics, and ten years before one can even submit their name to the Kennel Club for consideration as a breeding dog judge; however he accuses State Departments of Education of picking different people each year, giving them little or no training, and sending them out in schools to judge teaching performance (Brandt 1987, 5-6). His denunciation does not speak well of the emphasis placed upon appraiser training by state departments of education.

Some legislatures, however, feel training is so important that it is mandated for all appraisers. House Bill 72 in Texas and Senate Bill 813 in California are indications of this emphasis. Wickert reports, in California, that Senate Bill 813 requires all administrators "be certified" as to their proficiency in teacher evaluation and instructional methodologies. He suggests that the emphasis on the new legislation is toward the development of instructional

skills by teachers rather than a system designed for termination procedures (Wickert 1987, 23).

A part of Texas House Bill 72 directed the Texas State Board of Education to:

> . . . provide for a uniform training program and uniform certification standards for appraisers to be used throughout the state (Texas School Law Bulletin, Texas Education Code 13.302(c)).

The Texas State Board of Education fulfilled the legislative mandate by developing rules for the teacher appraisal process which defined appraisal standards and procedures. The rules include appraiser qualifications and uniform training requirements. Texas Education Code, paragraph (a)(5) of Section 149.43, Teacher Appraisal Procedures states:

> Before conducting appraisals, each appraiser must receive uniform appraiser training and must reach the required standard of proficiency as established by the State Board of Education. Periodic recertification will be required for each appraiser (Texas School Law Bulletin, Texas Education Code 149.43(a)(5)).

With the legislated mandate and State Board of Education rules, Texas seems to emphasize the need for initial and ongoing training of teacher evaluators; however, according to McLaughlin, most districts do not offer principals adequate training in their teacher evaluation responsibilities. She says that a few days, such as the weekend before a teacher evaluation system is first being implemented, is inadequate for the diagnostic, clinical, and staff-development skills necessary for an effective and successful

evaluation system. McLaughlin adds that for a teacher evaluation system to be successful, it will require not only a healthy initial investment but will also require a continual refining, building and refreshing of principal skills. She concludes by stating that training principals in a teacher evaluation system "is not something that is 'finished'; rather it is an ongoing, interactive activity" (McLaughlin 1984, 201).

Not only does it seem that the success of a teacher evaluation program relies on the training of teacher appraisers but the literature indicates both teachers and appraisers are desirous of training programs. Duckett, Strother, and Gephart feel that "teachers have the right to know who is evaluating them and what their qualifications are." They also indicate that when someone is making decisions about a teacher based upon judgement, then the teacher has the right to know what qualifies the appraiser as an evaluator. These authors claim teachers will have a positive experience with evaluation if the teachers are confident their evaluator is well trained (Duckett, Strother, and Gephart 1982, 1). According to the research conducted by Kauchak, Peterson, and Driscoll, teachers are concerned about the competence of their evaluators. In the elementary school, teachers were concerned when the appraiser had not taught at their particular level. At the secondary level, however, teachers were worried that ap-

praisers had little knowledge about their particular subject matter. But when the teacher was comfortable with the expertise of the appraiser, the evaluation process was viewed as being valuable by teachers (Kauchak, Peterson, and Driscoll 1985, 33- 34). These findings seem to be an argument for extensive training of appraisers before they conduct their first appraisal.

An article written by a junior high school English teacher gives another perspective from teachers. Kult argued that principals are not usually qualified in the subjects which they evaluate, either by experience, license, or certification. He also points to the fact that most principals do not have the practical knowledge in the programs and educational systems in which they evaluate. In his most harsh attack on evaluators, he claimed that college hours beyond a master's degree, a doctorate, a specialist degree, or the title of principal does not automatically give a person expertise and thus become a good appraiser (Kult 1978, 17-18).

According to McLaughlin, there is hope of reconciliation between teachers and principals in the evaluation process through training. She rationalizes that a common language is lacking between teachers and principals without further training of the principals. She indicates that training may permit evaluators to overcome the accusations and concerns voiced by teachers in the articles previously

mentioned. Through training, McLaughlin feels principals will gain the skills necessary to communicate to teachers precisely, clearly, and specifically about their observations. McLaughlin predicts training would allow principals to give specific advise and direction rather than make global statements (McLaughlin 1984, 197-198).

Teachers are not the only educators concerned about the expertise and training of their evaluators. Principals have this same concern. Many evaluators feel it is absolutely essential for them to maintain the perception of excellence in the evaluation process. Wise, et al., tell us that although the instruments used in evaluation contribute to discrepancies within the process, the inadequate training of evaluators also creates problems in the process. Their research indicated many districts felt appraisers did not receive adequate training in the process, and the training provided, did not supply adequate guidance (Wise, Darling-Hammond, McLaughlin, and Bernstein 1985, 75). The training process for teacher evaluators has made great strides in the past ten years. In 1976, Martin wrote that, "Training procedures are extremely difficult to develop since our present knowledge in this area is rather minuscule" (Martin 1976, 13). However, only ten years later, Bird and Little expressed that sufficient training programs had been developed and enough information concerning effective teaching is available to provide a beginning

for teacher evaluator training (Wise and Little 1986, 505). Jackson makes the point that although training of evaluators has made a marked improvement in recent years, and teacher appraiser training may occur at the beginning of an instrument's implementation, it is necessary for principals to continue studying the instrument to reach its most effective utilization (Jackson 1986, 5).

There is a debate about the effectiveness of evaluator training among the researchers, as there is about the validity of the evaluation process. But even Medley and Coker, two of the most prominent skeptics, have written:

> One might ask whether anything can be done to make principals' ratings more responsive to teacher effectiveness. Can principals be trained to be better judges of teacher effectiveness...we doubt that any amount of training can overcome it. But the effort is probably worth making (Medley and Coker 1987, 140).

A 1980 report by Coker, Medley, and Soar (1980) concluded there is little relationship between administrator effort, teacher improvement, and increased student learning. However, Wickert claims training models for teacher evaluators need to be rational models rather than based upon research (Wickert 1987, 24). Faast, however, describes a study conducted in the Des Moines, Iowa Independent Community School District during the 1981-82 school year, which would lend credence to the call for evaluator training. This study concluded: the training program was a success and effective; evaluators who completed the training could

analyze lesson plans more effectively; data is more easily captured during classroom observation by trained evaluators; and trained evaluators recognize and use conference skills more effectively (Faast 1984, 130).

The Petersons reported that the amount of teacher evaluation training needed by appraisers was dependent upon the purpose established by the district for the evaluation. And they wrote, "The only accurate way of determining 'adequacy' is to compare purpose with tested results of trained users" (Peterson and Peterson 1984, 43).

In an interview with Manatt and Schurter, McGreal reported that Manatt feels that if evaluators are given from five to ten days of training, an inter-rater reliability level can be reached which most districts would not be ashamed of. His interview with Schurter, superintendent of schools in Park Forest, Illinois, revealed that Schurter thinks most principals can do an adequate to above-average job of evaluation if they are given enough training. Schurter further stated principals are just like anyone else--some are better than others (McGreal 1986, 12).

Bolton summarizes the effectiveness of training teacher evaluators when he states, "Better training of personnel involved in teacher evaluation is likely to increase the validity, reliability, discrimination, and certainty of decisions" (Bolton 1973, 125).

The debate concerning training of teacher evaluators will continue to rage on into the future. The dilemma certainly creates the need for additional research to clarify the questions. Common sense would tell us, however, that training should increase the proficiency of appraisers. The only question to be resolved would be the type of training, the length of training, and the prerequisites required for the training sessions.

Reliability and Validity of Teacher Evaluation

Teacher evaluation is designed to assist classroom teachers in becoming more efficient or to identify and remove those instructors who are not competent to instruct the young pupil of our nation. To accomplish this task, education must have a teacher evaluation system which is reliable and valid. One author said, "Since around 1915 experts have been using rating scales to assist them in arriving at valid, considered opinions about teacher competence..." (Medley 1982, 15). In the July, 1985 issue of the American School Board Journal, a Rand Corporation report is quoted as saying, "Valid, reliable, and helpful evaluation requires evaluators who recognize good teaching . . ." (Rand Corporation 1985, 25). Yet the educational measurement community has contributed very little writing to the subject of validity and reliability in the teacher evaluation process. Rowley says after reviewing the

literature, it becomes obvious that measurement theory has not contributed to solving the problems of classroom observation (Rowley 1978, 165). Webb and Brown report the problem not only lies with measurement theorist but also with those that are responsible for teacher evaluations. They indicate that the methods used in teacher observations are often not reported and computed "in regard to observer reliability and validity." They conclude that confidence in teacher evaluation can not be built unless there is relevance and accurate data used in observations (Webb and Brown 1969, 1). Withall and Wood add that without valid and reliable data for evaluating a teacher's performance, the intent of evaluation will be lost (Withall and Wood 1979, 55). Peterson disclosed that in the process of developing teacher evaluation systems, "first predictive and content validity are established, then reliability" (Peterson 1983, 9).

To understand the full process of teacher evaluation, a cognizance of validity and reliability must be obtained. Thomas and Young suggest there are three essential elements of any good measurement: reliability, validity, and standardization. They explain reliability occurs if the assessment "measures consistently whatever it measures" (Thomas and Young 1986, 130). According to Kachigan, reliability simply means reproducibility. He says if results cannot be replicated, then the information obtained would be very unstable

or a matter of chance. He also indicates that reliability is "basic to every measurement situation" (Kachigan 1986, 217-218). Bailey tells us that reliability is the "consistency of the measurement." He adds that, "By equating reliability with consistency, we allow a scale to be not valid but still consistent (consistently inaccurate) and thus still reliable." (Bailey 1987, 67).

In an article by McGaw, Wardrop, and Bunda the problem of reliability was discussed. They indicated that some of the confusion about reliability was removed with two publications: Technical Recommendations for Psychological Tests and Diagnostic Techniques (American Psychological Association, 1954) and Technical Recommendations for Achievement Tests (American Educational Research Association, 1955). These authors reported that a subsequent paper, Standards for Educational and Psychological Tests and Manuals (American Psychological Association, 1966), defined reliability in the same terms as the previous two publications. They claim that reliability at the classroom level occurs when two assessments of the same subject are equal or do not vary in the observation by two independent evaluators. McGaw, et. al., also pointed out that most writers of classroom observation report reliability in terms of observer agreement; however they contend this type of definition confuses the issue because of the numerous other sources of unreliability (McGaw, Wardrop, and Bunda 1972, 13-15).

Stodolsky further testified that reliability, stability, and generalizability are closely related but have some differences. She tells us reliability, in studies of teacher behavior, is whether two observers will give conflicting ratings to the same teacher. She says this type of reliability investigates the objectivity of assessing behavior. She indicates a second type of reliability research deals with the reliability of stability. Reliability of stability studies consider the type of instrument used and the number of observations needed to produce consistent results (Stodolsky 1984, 12).

Brown, Mendenhall, and Beaver point to the different types of reliability and how these differences make it a "tricky concept." They say that dealing with this concept is amplified when it is removed from simply considering the tests of intelligence and achievement and becomes a measurement of classroom instruction in the teacher evaluation process. When classroom observation enters the concept, the problem of reliability is amplified by the observers themselves and their recording of the observation information (Brown, Mendenhall, and Beaver 1967, 11-12). These two additions, observer reliability and instrument reliability, could be the topic for an entire study. However, Brown and his co-authors, conclude their findings by stating, "People and conditions can be 'improved' in subsequent studies, but once they are 'out,' (sic. published) instruments rarely

are"; therefore, it is essential to establish reliability of the instrument before it is utilized (Brown, Mendenhall, and Beaver 1967, 23).

In the review of research on reliability and validity, most writers referred to a "classic article" on the subject written by Medley and Mitzel in 1963. The article, "Measuring Classroom Behavior by Systematic Observation," defines reliability in observation as,

> . . . the extent that the average difference between two measurements independently obtained (e.g., by two separate observer-recorders) in the same classroom is smaller than the average difference between two measurements obtained in different classrooms (Medley and Mitzel 1963, 291).

Peterson quotes Medley and Mitzel as saying different authors take different routes in approaching observer and instrument reliability. However, he has found there are two questions relating to reliability that are generic. First is the question of objectivity, and second is the question of stability and generalizability. He indicates objectivity is obtained when two or more observers watch a teacher instruct a class, record their observations, and then compare the records by using statistical measures of variance to discover similarities and differences. He says once objectivity is established, additional observations are made in a variety of settings and over a period of time to reach generalizability and stability. When these two questions are answered, reliability is established (Peterson 1983, 9).

Rowley studied the relationship of reliability and the amount of observation undertaken. He disclosed that reliability is not necessarily the instrument itself but the measurement obtained from the instrument. He proposes reliability depends more on the skill of the observers, the subjects observed, and the number and length of the observations rather than the instrument itself (Rowley 1978, 166).

According to an article published by the Association of Teacher Educators and Bureau of Educational Research, most researchers agree that a reliability coefficient of at least .90 should be obtained by an instrument before it is suitable for evaluating individuals (Medley 1982, 11). Although a number of studies indicate a reliability coefficient this high can not be attained with a high degree of confidence, McGreal says, ". . . we have progressed dramatically in the last 10 years in our ability to evaluate teacher performance more reliably" (McGreal 1986, 11).

But before reliability can be ascertained, validity must be established. Thomas and Young added to the literature by writing that validity is the extent to which an instrument "measures what it purports to measure or to the extent to which it does the job for which it is used." They report there are three kinds of validity. Content validity is the degree which an instrument's content "samples the subject matter or situation about which conclusions are to be drawn." They also state there must be adequate sampling

of the material before content validity can be established. (Thomas and Young 1986, 126). According to them, criterion-related validity is "the extent to which scores . . . correlate with some given criterion measure." They write of two types of criterion-related validity: predictive validity and concurrent validity. Predictive validity usually refers to the correlation between two test scores. While concurrent validity "is evidenced by high correlation between the test and some measure of contemporary criterion performance." Concurrent validity usually reports the "status of an individual with respect to some trait or characteristic." An instrument is said to have concurrent validity if it can measure, indirectly, a trait or characteristic of an individual which it is attempting to measure. The third type of validity is construct validity. This type of validity deals with the "theory underlying the test" and has the ability to indicate a psychological quality or trait (Thomas and Young 1986, 130). Selltiz et al. says that validity can be defined as:

> . . . the extent to which differences in scores on it reflects true differences among individuals on the characteristic that we seek to measure, rather than constant or random errors (Selltiz, Wrightsman, and Cook 1976, 168-169).

And Derek Phillips writes that in measurement for the scientific use of a particular phenomenon, an instrument is considered valid if it measures the phenomenon successfully (Phillips 1971, 197).

Bailey says, that without question, validity has two parts: the instrument must measure the concept in question and not another; and the measurement of the concept must be accurate. He continues by stating an assessment can have the first without the second but they can not be reversed. "The concept cannot be measured accurately if some other concept is being measured." Bailey's research indicates several different validation procedures: face validation which is sometimes referred to as content validation; criterion validation which includes pragmatic validation, concurrent validation and predictive validation; and construct validation. He reports that face validity is the easiest to define since it is simply a matter of judgment. Criterion validity requires several measurements of the same concept and is often referred to in the literature as pragmatic validity, predictive validity, or concurrent validity. Bailey further defines concurrent validity as follows:

> The term 'concurrent validity' has been used to describe a measure that is valid for measuring a particular phenomenon at the present time, while 'predictive validity' refers to the measure's ability to predict future events. . . . The process entails use of a second measure of the concept as a criterion by which the validity of the new measure may be checked. . . . If the two scores were similar, the new method could be said to have criterion validity or, to be more specific, concurrent criterion (pragmatic) validity (Bailey 1987, 68).

He argues that construct validity is the strongest kind of validation procedure since all three types build upon one another. Thus construct validity includes all the features

of the other procedures plus additional elements (Bailey 1987, 69).

According to Kachigan, validity is the extent that assessments do what they claim they can do or what is intended for them to do. Also, he expressed that it was important to note an assessment "can be reliable without being valid, but cannot be valid without being reliable . . . it is impossible for a measurement system to be valid without being reliable" (Kachigan 1986, 219).

In the article by Medley and Mitzel referred to earlier, validity is defined as "the extent that differences in scores yielded by it reflect actual differences in behavior--not differences in impressions made on different observers" (Medley and Mitzel 1963, 291). McNally wrote "validity is the degree to which an evaluation procedure or instrument measures what it is supposed to measure." He also advocates that the best way to attain a high degree of validity in evaluation instruments is by the cooperative planning and implementation of the evaluation system (McNally 1977, 105).

Peterson says from a legal standpoint, evaluation systems must have validity. And he describes validity as the measurement of those attributes which it purports to measure. He points to two types of validity: content validity and predictive validity. He claims that predictive validity assesses a connection between how much students learn and what a teacher does behaviorally. He suggests

that content validity hinges on consensus instead of statistical measures. He reaches this claim since content validity:

> . . . refers to the extent to which knowledgeable people agree that the evaluation system contains items or categories that are clearly articulated and representative of the concepts that are to be measured (Peterson 1983, 8).

In an article discussing competency-based teacher education, Coker, Medley and Soar state:

> Validation of a competence requires not only that the competence be operationally defined but that evidence be produced to show that teachers who possess it are (on the average) more effective in helping pupils learn than teachers who do not. Ideally, evidence would be presented that there is a cause-and-effect relationship between mastery of the competence and effectiveness in the classroom. Evidence that the two are correlated is minimal proof of validity . . . (Coker, Medley, and Soar 1980, 131).

Webb and Brown have expressed a concern for the lack of literature concerning the validity of systematic classroom observation and specifically with the validity of between-observer agreement (the agreement between observers of the same teacher behavior). They suggest that if research on observer validity has been conducted, it had not been published at the time of their research. However, they state their studies show that within-observer reliability (the stability of an individual observer's responses to the same behavior over a period of time) can be achieved if validity is established (Webb and Brown 69, 11).

It becomes obvious in reviewing the literature that research concerning the critical aspects of validity and reliability in the teacher evaluation process is lacking. The question of validity and reliability will have an impact on education in future legal tests of teacher evaluation and become a point of contention for teacher organizations in the future; therefore, it is incumbent on educational researchers to further refine the questions of reliability and validity as they relate to the teacher evaluation process.

REFERENCE LIST

Bailey, Kenneth D. 1987. Methods of social research. 3rd Edition. New York, New York: The Free Press.

Barber, B. 1965. The professions in America. Boston: Houghton Mifflin.

Barth, Roland S. 1979. Teacher evaluation and staff development. Principal 58 (January): 74-77.

Beckham, Joseph C. 1981. Legal aspects of teacher evaluation. National Organization on Legal Problems of Education. Topeka, Kansas. 1-53.

Bird, Tom and Judith Warren Little. 1986. How schools organize the teaching occupation. The Elementary School Journal 86 (March): 493-511.

Bolton, Dale L. 1973. Selection and evaluation of teachers. Berkeley, California: McCutchan Publishing Corporation.

Brandt, Ronald S. 1987. On the expert teacher: A conversation with David Berliner. Educational Leadership 44 (October): 4-9.

Bridges, Edwin M. 1985. Managing the incompetent teacher--What can principals do? NASSP Bulletin 69 (February): 57-65.

Brieschke, Patricia A. 1986. The administrative role in teacher competency. The Urban Review 18 (No. 4): 237-251.

Brown, Bob Burton. 1968. The experimental mind in education. New York, New York: Harper & Row, Publishers.

Brown, Bob Burton and Jeaninne N. Webb. 1975. The use of classroom observation techniques in the evaluation of educational programs. ERIC, ED 117 192.

Brown, Bob Burton, William Mendenhall, and Robert Beaver. 1967. The reliability of observations of teachers' classroom behavior. ERIC, ED 011 520.

Buttram, Joan L. and Bruce L. Wilson. 1987. Promising trends in teacher evaluation. Educational Leadership 44 (April): 4-6.

Coker, Homer. 1985. A study of the correlation between principals' ratings of teacher effectiveness and pupil growth. ERIC, ED 259 460.

Coker, Homer. 1987. The accuracy of principal's judgments of teacher performance. Journal of Educational Research 80 (March-April): 242-247.

Coker, Homer, Donald M. Medley, and Robert S. Soar. 1980. How valid are expert opinions about effective teaching? Phi Delta Kappan 62 (October): 131-149.

Cruickshank, Donald R. 1986. Profile of an effective teacher. Educational Horizons 64 (Winter): 80-86.

Darling-Hammond, Linda. 1986. A proposal for evaluation in the teaching profession. The Elementary School Journal 86 (March): 531-551.

Darling-Hammond, Arthur E. Wise, and Sara R. Pease. 1983. Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research 53 (Fall): 285-327.

Duckett, Bruce, William Strother, and Edward Gephart. 1982. Evaluating the evaluators. Practical Applications of Research 5 (March): 1-12.

Dukakis, Michael S. 1988. Acceptance speech for the Democratic Parties' nomination for the President of the United States of America. Atlanta, Georgia: July 21.

Emmer, Edmund T. and Robert F. Peck. 1973. Dimensions of classroom behavior. Journal of Educational Psychology 64 (April): 223-240.

English, Fenwick W. 1985. Still searching for excellence. Educational Leadership 42 (December-January): 34-35.

Faast, Dorothy A. 1984. Appraiser training: Teacher performance evaluation training for school administrators. The Clearing House 58 (November): 128-130.

Gage, N. L. 1978. The scientific basis of the art of teaching. New York: Teachers College Press.

Hampel, Robert L. 1986. The political side of reform: Are conflicts, power struggles likely to occur? NASSP Bulletin 70 (December): 55-64.

House, Ernest R. 1980. Evaluating With validity. Beverly Hills, California: Sage Publications, Inc.

Jackson, Mary E. 1986. Teacher evaluation: The role of the principal. ERIC, ED 275 049.

Johnston, Gladys Styles and Carol Camp Yeakey. 1979. Supervision of teacher evaluation: Brief overview. Journal of Teacher Education 30 (March-April): 17-22.

Kachigan, Sam Kash. 1986. Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods. New York, New York: Radius Press.

Kauchak, Don, Ken Peterson, and Amy Driscoll. 1985. An interview study of teachers' attitudes toward teacher evaluation practices. Journal of Research and Development in Education 19 (Fall): 32-37.

Knapp, Michael S. 1982. Toward the study of teacher evaluation as an organizational process: A review of current research and practice. Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York (March): 1-26.

Kult, Lawrence E. 1978. Improving teacher evaluations by principals. The Clearing House 52 (September): 17-21.

Laing, Steven O. 1986. The principal and evaluation. NASSP Bulletin 70 (November): 91-93.

Lareau, Annette. 1986. A comparison of professional examinations in seven fields: Implications for the teaching profession. The Elementary School Journal 86 (March): 553-569.

McGaw, Barry, James L. Wardrop, and Mary Anne Bunda. 1972. Classroom observation schemes: Where are the errors? American Educational Research Journal 9 (January): 13-27.

McGreal, Tom. 1986. How well can we truly evaluate teachers? The School Administrator 29 (January): 10-12.

McLaughlin, Milbrey Wallin. 1984. Teacher evaluation and school improvement. Teachers College Record 86 (Fall): 193-207.

McNally, Harold J. 1977. Performance-based teacher evaluation. NASSP Bulletin 61 (October): 104-105.

Magoon, A. Jon. 1979. Sensitive field observation of teaching performance. Journal of Teacher Education 30 (March-April): 13-16.

Martin, Jack. 1976. Developing category observation instruments for the analysis of classroom behavior. Journal of Classroom Interaction 12 (December): 1-16.

Medley, Donald M. 1982. Teacher competency testing and the teacher educator. Charlottesville, Virginia: Bureau of Educational Research.

Medley, Donald M. and Homer Coker. 1987. The accuracy of principals' judgments of teacher performance. Journal of Educational Research 80 (March/April): 242-247.

Medley, Donald M. and Homer Coker. 1987. How valid are principals' judgments of teacher effectiveness? Phi Delta Kappan 69 (October): 138-140.

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

Moxley, Roy A. 1978. Teacher evaluation: Images and analysis. Journal of Teacher Education 29 (November-December): 61-66.

National Commission on Excellence in Education. 1983. A nation at risk: The imperative for educational reform. U.S. Department of Education.

Natriello, Gary. 1984. Teachers' perceptions of the frequency of evaluation and assessments of their effort and effectiveness. American Educational Research Journal 21 (Fall): 579-595.

Pellicer, Leonard O. and O. B. Hendrix. A practical approach to remediation and dismissal. NASSP Bulletin 64 (March): 57- 62.

Peterson, Donovan. 1983. Legal and ethical issues of teacher evaluation: A research based approach. Educational Research Quarterly 7 (Winter): 6-16.

Peterson, Donovan and Kathryn Peterson. 1984. A research-based approach to teacher evaluation. NASSP Bulletin 68 (February): 39-46.

Peterson, Ken. 1984. Methodological problems in teacher evaluation. Journal of Research and Development in Education 17 (Summer): 62-70.

Phillips, Derek L. 1971. Knowledge from what? Theories and methods in social research. Chicago: Rand McNally.

Pigford, Aretha Butler. 1987. Teacher evaluation: More than a game that principals play. Phi Delta Kappan 69 (October): 141-142.

Rand Corporation. 1985. These are the elements of a sound teacher evaluation system. American School Board Journal 172 (July): 25.

Rowley, Glenn. 1978. The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. Journal of Educational Measurement 15 (Fall): 165-180.

Roy, Joseph J. 1979. Teacher evaluation in an era of educational change. The Clearing House 52 (February): 275-276.

Sapone, Carmelo V. 1980. An appraisal and evaluation system for teachers and administrators. Educational Technology (May): 44-49.

Seeley, David S. 1979. Reducing the confrontation over teacher accountability. Phi Delta Kappan 61 (December): 248-251.

Selltiz, Claire, Lawrence J. Wrightman, and Stuart W. Cook. 1976. Research methods in social relations. Third ed. New York, New York: Holt, Rinehart & Winston.

Shanker, A. 1985. A call for professionalism. January 29 speech to the National Press Club. Washington, DC: American Federation of Teachers.

Shulman, Lee S. 1987. Knowledge and teaching: Foundations of the new reform. Harvard Educational Review 57 (February): 1-22.

Soar, Robert S. 1975. Accountability: Assessment problems and possibilities. The Journal of Teacher Education 24 (February): 205-212.

Stodolsky, Susan S. 1984. Teacher evaluation: The limits of looking. Educational Researcher 13 (November): 11-18.

Task Force on Education for Economic Growth, Education Commission of the States. 1983. Action for Excellence. (June).

Texas School Law Bulletin. 1988. Austin: West Publishing Company.

Thomas, Fred L. and Jon I. Young. 1986. An introduction to educational statistics: The essential elements. Lexington, Massachusetts: Ginn Press.

Tracey, William R. 1978. Teacher evaluation--another perspective. The Clearing House 51 (January): 240-242.

Webb, Jeaninne Nelson and Bob Burton Brown. 1969. Establishing reliability and validity estimates for systematic classroom observation. ERIC, ED 028 123.

White, Kinnard, Marvin D. Wyne, Gary B. Stuck, and Richard H. Coop. 1987. Assessing teacher performance using an observation instrument based on research findings. NASSP Bulletin 71 (March): 89-95.

Wickert, Donald M. 1987. Using teacher evaluation for improving instruction. The Clearing House 61 (September): 23-24.

Wise, Arthur E. and Linda Darling-Hammond. 1985. Teacher evaluation and teacher professionalism. Educational Leadership 42 (December-January): 28-33.

Wise, Arthur E., Linda Darling-Hammond, Milbrey W. McLaughlin, and Harriet T. Bernstein. 1985. Teacher evaluation: A study of effective practices. The Elementary School Journal 86 (September): 61-86.

Withall, John, and Fred H. Wood. 1979. Taking the threat out of classroom observation and feedback. Journal of Teacher Education 30 (January-February): 55-58.

Wuhs, Susan K. 1983. The pace of mandated teacher evaluation picks up. The American School Board Journal 170 (May): 28

CHAPTER 3

PROCEDURES FOR DATA COLLECTION AND ANALYSIS

Selection of Sample

Between June 8, 1986, and August 15, 1986, approximately 742 persons in an eight county area of North Texas were given training on the utilization of the Texas Teacher Appraisal System (TTAS) by Region 10 Education Service Center trainers. During training, each person was provided the opportunity to view and score the same six video-taped lessons using the TTAS Instrument (Appendix B). These persons were also requested to provide their personal demographics which included: age, sex, race, teaching field(s), level of teaching experience, size of school district, and length of experience as an administrator. This data was collected on Scantron bubble sheets, scored, and manipulated with the use of an IBM-AT computer. Any data that appeared to be flawed were removed from the study. The final sample consisted of 622 persons who completed the training with usable data.

Collection of Data

A demographic file was compiled from the information available on the subjects used in the study. The following

variables were included: identification number, sex, race, years of administrative experience, content area taught when teaching, campus level of teaching experience, campus assignment, and size of school district.

A total of 636 data sheets were evaluated. Upon examination of the identification numbers, fourteen subjects with duplicate identification numbers were identified. All fourteen of these data records were eliminated from the demographic file. As a result, 622 valid individual cases remained in the demographic file.

Six different tests designed by the state to test appraiser proficiency were utilized in the evaluation of appraisers. The six sets of test results were scanned and collected into six separate files. Each file contained the identification number of the respondent, and the responses indicating the presence or absence of the 55 performance indicators observed on the videotapes. The demographic file was merged with the data in the test files to allow an analysis involving comparison of groups based on the demographic variables. The ID number was used as the key to match data from the test files and the demographic file.

The six separate files of test data were collected, scanned, and saved in computer files. Data from the demographic file was combined with these six test data files by matching the identification number of each test data record with the identification number in the demographic file.

Only test information with matching demographic information will be used in the statistical analysis.

Procedures for Analysis of Data

In order to determine what combination of independent variables best predicts accurate evaluation, a multiple linear regression procedure was used. The dependent variable for all analyses is the total number of correct responses to the proficiency test. The independent variables are:

1. Campus-level assignment
2. Size of school district
3. Sex
4. Race
5. Years of administrative experience
6. Campus level of teaching experience
7. Curriculum area taught when teaching

Each of the seven independent variables was tested using multiple linear regression. In this procedure a full model was tested against a restricted model of prediction. The full model consisted of a prediction equation using all seven independent variables as predictors. The restricted model consisted of all of the predictors except the variable being tested. If the full model gave a significantly (.05) better prediction than the restricted model, then the

variable being tested was considered a significant predictor of evaluation accuracy.

To accomplish the regression procedures the variables had to be transformed from nominal (race, content areas) or ordinal variables (campus assignment, size, years of experience, teaching level) into a series of dummy coded variables. Only the variable sex did not need to be transformed since it was already expressed as a dichotomy. Race, with four values (White, Hispanic, Black, and other) was transformed to three dummy variables (Race1, Race2, Race3). Race1 was set to 1 for Blacks, 0 for all others; Race2 was set to 1 for Hispanics, 0 for all others; Race3 was set to 1 for Whites, 0 for all others. Therefore, the meaning of the variable Race1 is effectively, "Black or not Black." Taken together the three race dummy coded variables represent all four possible values of the original variable race.

In the hypothesis testing part of the regression procedure, the effect of the dummy coded variables was tested together to give an effect equivalent to testing the effect of the original variable. In the stepwise part, each dummy coded variable was free to enter the equation independently.

The best prediction model using a combination of the independent variables was determined using stepwise regression. The hypothesis testing procedures were repeated on all six of the training tapes. Thus, each of the hypotheses was tested six separate times. A seventh data analysis

procedure was performed as well. Average accuracy for individuals who evaluated the teaching performance in all six training tapes was computed. This summary data set gave a representation of the relationship of the demographic variables to evaluation accuracy for the composite of the six teaching situations evaluated.

Where an independent variable was significant in predicting the dependent variable, a follow-up ANOVA and Tukey's multiple comparison were employed to further explain the nature of the relationship between the dependent and independent variables.

All statistical analysis were performed using SPSS-PC+ statistical package on an IBM PC computer.

CHAPTER 4

PRESENTATION AND ANALYSIS OF DATA

The purpose of this study was to determine if there are personal and demographic characteristics which can predict teacher appraisers who will be the most accurate. The effectiveness of seven demographic characteristics in predicting evaluation accuracy was examined. Those seven characteristics were campus assignment, size of school district, sex, race, years of administrative experience, content area taught while in the classroom, and school level taught.

The participants in the study were given training in using the Texas Teacher Appraisal System (TTAS) and then tested for appraisal accuracy by evaluating six previously videotaped teaching sessions. The accuracy of the participants in evaluating the videotapes was computed as agreement with the experts who constructed the scoring key. Each participant's score was computed as the number of items, out of 55, they evaluated correctly. Each of the six exercises comprised a separate test of evaluation accuracy. The results of analysis of the six evaluation exercises are presented in order below.

Evaluation Exercise One--Ninth Grade Grammar

Table 1 presents the means and standard deviations of evaluation accuracy (total items correct out of 55). The categories of each variable are shown with the descriptive information on evaluation accuracy for each one. This descriptive information shows how many individuals were in each category of each variable (N), the average accuracy by category (Mean), and the degree of variability in accuracy within each category (S.D.).

Table 1.--Evaluation accuracy on exercise 1 broken down by levels of the demographic variables

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Campus Assignment			
Elementary	156	41.63	6.71
Middle school	128	41.64	7.57
High school	165	41.75	7.28
Central administration	72	41.60	6.25
Size of school district			
Under 1,000	101	42.03	6.48
1,000 to 2,999	67	41.91	7.47
3,000 to 5,999	99	41.44	7.33
6,000 to 9,999	104	42.80	6.53
10,000 or more	150	40.67	7.27
Sex			
Female	169	41.94	6.93
Male	359	41.47	7.11
Race			
Black	27	45.07	7.74
Hispanic	4	44.25	9.54
White	502	41.48	6.95
Other	4	42.00	9.66

Table 1.--Continued.

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Administrative Experience			
1 to 3 years	107	41.58	7.29
4 to 10 years	190	41.12	7.07
11 to 20 years	153	42.14	6.69
21 or more years	42	42.83	8.14
Not an administrator	47	41.53	6.57
Content Area Taught			
English	49	41.41	7.14
Fine arts	18	39.28	6.52
Math	44	41.30	6.87
Physical education	48	42.27	5.25
Science	60	39.27	7.27
Self-contained elementary	128	42.40	6.66
Social studies	88	41.70	7.83
Special education	33	42.18	5.51
Vocational education	26	44.15	7.16
Not listed	44	41.73	8.52
Level of Teaching Experience			
Elementary	172	42.13	6.47
Middle school	107	41.62	7.26
High school	247	41.26	7.37

Table 2 shows the relationship between each predictor variable ("Source" column) and evaluation accuracy. Each row of the table indicates the statistical significance of the relationship between that predictor and evaluation accuracy. The column, "Sig F" (significance of F) indicates whether the predictor is statistically significant or not. If the value of "Sig F" is less than .05, then the predictor is statistically significant. The table shows that only race is a statistically significant predictor of accuracy.

The "Regression" row depicts the statistical significance of the overall regression equation. As the Sig F column shows, the overall prediction equation with all variables entered does not provide a statistically significant prediction of evaluation accuracy. The adjusted R square for the equation is just over .01, indicating that just over 1 percent of the variance in evaluation accuracy can be predicted from all of the variables taken together.

Even though race was a statistically significant (.0499) predictor of evaluation accuracy in scoring the first training tape, the overall regression equation was not statistically significant. Therefore, no follow-up oneway analysis of variance (ANOVA) for race is reported.

Table 2.--Test for effect of demographic variables on prediction of evaluation accuracy (exercise 1)

Source	DF	Sum of Squares	RSq Chg	F	Sig F
Assignment	3	68.69992	.00273	.46385	.7077
Size	4	249.28519	.00991	1.26235	.2839
Sex	1	31.33376	.00125	.63468	.4260
Race	3	388.79633	.01545	2.62508	.0499
Experience	4	232.59543	.00924	1.17783	.3197
Content area	9	542.28478	.02155	1.22047	.2799
Level taught	2	51.98360	.00207	.52648	.5910
Regression	26	1562.28852		1.21711	.2136
Residual	478	23598.57088			
Total	504	25160.85941			

$R = .24918$; $R^2 = .06209$; Adjusted $R^2 = .01108$ ($p = .2136$)

The stepwise regression procedure yielded a prediction equation with three of the dummy coded variables as predictors. As table 3 shows, the overall adjusted R square was just over .03, indicating that 3 percent of the variance in evaluation accuracy was accounted for by the three predictor variables. This percentage of variance accounted for is small but statistically significant (p = .003).

Table 3.--Stepwise regression equation (exercise 1)

Variable	B	SE B	Beta	T	Sig T
Content5	-2.75084	.98648	-.12237	-2.789	.0055
Race3	-3.83411	1.37734	-.12219	-2.784	.0056
Size4	1.52558	.77444	.08645	1.970	.0494
(Constant)	45.18152	1.35192		33.420	.0000

R = .19049; R^2 = .03629; Adjusted R^2 = .03052 (p = .0030)

The number of the dummy coded variables corresponds to the order the categories appear in the tables of descriptive statistics earlier. Thus, Science, Whites, and schools with enrollments of 1,000 to 2,999 were selected. The sign (none, indicating +, or -) on the B and Beta in the prediction equation indicates the direction of the relationship between the dummy coded variable and evaluation accuracy. A negative sign indicates that being in the category specified is associated with a lower accuracy; whereas a positive B and Beta indicates that being in that category is associated with a higher accuracy. With these guidelines in mind, the interpretation of the stepwise regression equation indicates

that higher accuracy in evaluation was associated with 1) being in a school district of size 1,000 to 2,999, 2) not being white, and 3) not having taught science.

In summary, there is no compelling evidence that any of the demographic variables were significant predictors of accuracy in evaluation in this data set. The overall prediction equation with all demographic variables entered was not statistically significant. The stepwise procedure computed a prediction equation with three of the dummy coded variables. However, since the overall prediction equation was not significant, this finding should be held tentatively until more evidence is gathered.

Evaluation Exercise Two--Third Grade Science

Table 4 presents the means and standard deviations of evaluation accuracy (total items correct out of 55). The categories of each variable are shown with the descriptive information on evaluation accuracy for each one.

Table 4.--Evaluation accuracy on exercise 2 broken down by levels of the demographic variables

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Campus Assignment			
Elementary	171	41.26	2.90
Middle school	137	41.85	2.43
High school	185	41.64	2.67
Central administration	83	40.64	4.78

Table 4.--Continued.

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Size of school district			
Under 1,000	108	41.54	2.49
1,000 to 2,999	79	41.54	2.60
3,000 to 5,999	107	41.36	4.25
6,000 to 9,999	109	41.46	2.56
10,000 or more	173	41.35	3.12
Sex			
Female	189	41.43	2.88
Male	388	41.42	3.19
Race			
Black	26	39.81	4.32
Hispanic	3	44.00	2.00
White	553	41.45	3.00
Other	3	42.00	5.29
Administrative Experience			
1 to 3 years	114	41.73	2.90
4 to 10 years	215	41.13	3.63
11 to 20 years	169	41.63	2.47
21 or more years	44	41.18	2.79
Not an administrator	46	41.37	3.11
Content Area Taught			
English	49	40.86	3.96
Fine arts	21	42.48	2.27
Math	51	41.78	2.26
Physical education	53	41.72	2.65
Science	63	40.92	2.82
Self-contained elementary	140	41.37	2.85
Social studies	100	41.89	2.11
Special education	36	41.28	3.04
Vocational education	28	41.07	2.19
Not listed	46	40.89	5.76
Level of Teaching Experience			
Elementary	187	41.45	2.78
Middle school	123	41.40	3.92
High school	264	41.45	2.84

Table 5 shows the relationship between each predictor variable ("Source" column) and evaluation accuracy. The table shows that campus assignment (Sig F = .0069) and race (Sig F = .0185) are significant predictors of accuracy. The overall prediction equation with all variables entered provides a significant (.0307) prediction of evaluation accuracy. The adjusted R square for the equation is just under .02774 indicating that just under 3 percent of the variance in evaluation accuracy can be predicted from all of the variables taken together.

Table 5.--Test for effect of demographic variables on prediction of evaluation accuracy (exercise 2)

Source	DF	Sum of Squares	RSq Chg	F	Sig F
Assignment	3	114.42231	.02166	4.09176	.0069
Size	4	33.91382	.00642	.90957	.4580
Sex	1	4.96042	.00094	.53216	.4660
Race	3	94.09526	.01781	3.36486	.0185
Experience	4	51.73505	.00979	1.38754	.2370
Content area	9	127.60369	.02416	1.52104	.1371
Level taught	2	11.28195	.00214	.60517	.5464
Regression	26	388.86826		1.60454	.0307
Residual	525	4893.71145			
Total	551	5282.57971			

R = .27132; R^2 = .07361; Adjusted R^2 = .02774 (p = .0307)

As a follow-up, two oneway ANOVAs were performed using campus assignment and race as the independent variables. The ANOVA procedure with campus assignment as the independent variable yielded F (3, 572) = 3.13, p = .025 (see

table 6). Tukey's multiple comparison procedure showed that evaluators assigned to middle schools and high schools had significantly higher rating accuracy than did those assigned to central administration.

Table 6.--Follow-up ANOVA with campus assignment as independent variable

Source	Sum of Squares	DF	Mean Square	F	Sig of F
Campus Assignment	88.943	3	29.648	3.134	.025
Residual	5410.550	572	9.459		
Total	5499.493	575	9.564		

The ANOVA procedure with race as the independent variable yielded F (3, 581) = 3.13, p = .025 (see table 7). Reference to table 4 shows that from high to low in evaluation accuracy the racial groups are arranged as follows: Hispanic, Other, White, Black. However, Tukey's multiple comparison procedure showed only that White evaluators had higher rating accuracy than did Black evaluators. The other pairwise comparisons were not significant due to the small number of individuals in the Hispanic and Other categories.

Table 7.--Follow-up ANOVA with race as independent variable

Source	Sum of Squares	DF	Mean Square	F	Sig of F
Race	88.881	3	29.627	3.131	.025
Residual	5497.113	581	9.461		
Total	5585.993	584	9.565		

The stepwise regression procedure yielded a prediction equation with two of the dummy coded variables as predictors. As table 8 shows, the overall adjusted R square was just over .01, indicating that 1 percent of the variance in evaluation accuracy was accounted for by the two predictor variables.

Table 8.--Stepwise regression equation (exercise 2)

Variable	B	SE B	Beta	T	Sig T
Race1	-4.21818	1.52525	-.26674	-2.766	.0059
Race3	-2.92571	1.38334	-.20399	-2.115	.0349
(Constant)	44.40000	1.37680		32.249	.0000

$R = .12245$; $R^2 = .01499$; Adjusted $R^2 = .01141$ ($p = .0158$)

The number of the dummy coded variables corresponds to the order the categories appear in the tables of descriptive statistics earlier. Thus Blacks and Whites were selected. The interpretation of the stepwise regression equation indicates that higher accuracy in evaluation is associated with 1) <u>not</u> being black, and 2) <u>not</u> being white. In other words, the Hispanics and "Others" had the highest evaluation accuracy. This finding is consistent with the result from the oneway ANOVA on race reported above. The difference resulted from the nature of the comparisons being made. Here each racial group is compared to all other groups combined; in Tukey's procedure groups are compared pairwise.

White and Black show up as significant because of the large number in these groups.

In summary, there is evidence that race is a significant predictor of accuracy in evaluation in this data set. The ANOVA procedure showed that Whites evidenced higher accuracy than Blacks. However, the stepwise regression procedure showed that Whites and Blacks both had lower evaluation accuracy than did Hispanics and "others." Therefore, we can assume that with only 3 Hispanics and 3 "others" in this sample, this result might not be repeated in other population groups. The overall regression procedure and follow-up ANOVA showed campus assignment to be significant; however, the stepwise procedure did not show significance.

Evaluation Exercise Three--Eighth Grade Social Studies

Table 9 presents the means and standard deviations of evaluation accuracy (total items correct out of 55). The categories of each predictor variable are shown with the descriptive information on evaluation accuracy for each one.

Table 9.--Evaluation accuracy on exercise 3 broken down by levels of the demographic variables

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Campus Assignment			
Elementary	171	36.39	2.42
Middle school	143	36.74	2.80
High school	176	36.35	2.47
Central administration	92	36.38	3.14
Size of school district			
Under 1,000	102	36.38	2.76
1,000 to 2,999	79	36.13	2.66
3,000 to 5,999	98	36.62	2.75
6,000 to 9,999	99	36.52	2.55
10,000 or more	204	36.53	2.62
Sex			
Female	188	37.02	2.57
Male	395	36.18	2.67
Race			
Black	25	35.16	2.39
Hispanic	4	37.25	2.06
White	559	36.50	2.66
Other	3	37.67	2.08
Administrative Experience			
1 to 3 years	111	36.62	2.68
4 to 10 years	219	36.56	2.80
11 to 20 years	173	36.47	2.53
21 or more years	46	35.33	2.55
Not an administrator	44	36.52	2.25
Content Area Taught			
English	46	37.07	2.52
Fine arts	22	36.95	2.80
Math	49	36.39	2.99
Physical education	56	36.71	2.26
Science	66	36.41	2.60
Self-contained elementary	138	36.36	2.68
Social studies	102	36.28	2.88
Special education	35	37.00	2.51
Vocational education	31	35.55	2.26
Not listed	47	36.30	2.57

Table 9.--Continued.

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Level of Teaching Experience			
Elementary	187	36.43	2.56
Middle school	128	36.64	2.80
High school	263	36.37	2.61

Table 10 shows the relationship between each predictor variable ("Source" column) and evaluation accuracy. Table 10 shows that sex is a significant (.0002) predictor of accuracy. Females (37.02) had higher evaluation accuracy than did males (36.18) (see table 9). The overall prediction equation with all variables entered provides a significant (.0193) prediction of evaluation accuracy. The adjusted R square for the equation is just over .03 indicating that just over 3 percent of the variance in evaluation accuracy can be predicted from all of the variables taken together.

Table 10.--Test for effect of demographic variables on prediction of evaluation accuracy (exercise 3)

Source	DF	Sum of Squares	RSq Chg	F	Sig F
Assignment	3	13.61973	.00351	.67161	.5697
Size	4	11.54355	.00297	.42692	.7892
Sex	1	93.00855	.02394	13.75909	.0002
Race	3	34.20482	.00880	1.68668	.1688
Experience	4	43.79584	.01127	1.61972	.1679
Content area	9	60.66932	.01561	.99723	.4411
Level taught	2	28.02292	.00721	2.07277	.1269
Regression	26	295.96350		1.68396	.0193
Residual	531	3589.44869			
Total	557	3885.41219			

$R = .27599$; $R^2 = .07617$; Adjusted $R^2 = .03094$ ($p = .0193$)

The stepwise regression procedure yielded a prediction equation with four of the dummy coded variables as predictors. As table 11 shows, the overall adjusted R square was just under .04, indicating that 4 percent of the variance in evaluation accuracy was accounted for by the four predictor variables.

Table 11.--Stepwise regression equation (exercise 3)

Variable	B	SE B	Beta	T	Sig T
Sex	-.83254	.24047	-.14832	-3.462	.0006
YearsExp4	-.95374	.42188	-.09536	-2.261	.0242
Race1	-1.20873	.57606	-.08717	-2.098	.0363
LevelExp2	.54854	.27022	.08567	2.030	.0428
(Constant)	37.87633	.40736		92.979	.0000

$R = .21527$; $R^2 = .04634$; Adjusted $R^2 = .03944$ ($p < .0001$)

The number of the dummy coded variables corresponds to the order in which the categories appear in the tables of descriptive statistics earlier. Thus administrators with 21 or more years, Blacks, Middle School, and sex were selected. The interpretation of the stepwise regression equation indicates that higher accuracy in evaluation was associated with 1) not being male, 2) not having 21 or more years experience as an administrator, 3) not being Black, and 3) having had teaching experience at the middle school level.

In summary, the best evidence for prediction of evaluation accuracy for the third data set came from the demographic variable sex; females had higher average evaluation accuracy than males. The stepwise procedure added evidence that certain levels of three other variables may deserve attention. In race lower accuracy was found for Blacks; in level of experience, higher accuracy was found for those with experience in middle school; and in years of experience, lower accuracy was found for those with 21 or more years of experience.

Evaluation Exercise Four--Middle School Mathematics

Table 12 presents the means and standard deviations of evaluation accuracy (total items correct out of 55). The categories of each predictor variable are shown with the descriptive information on evaluation accuracy for each one.

Table 12.--Evaluation accuracy on exercise 4 broken down by levels of the demographic variables

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Campus Assignment			
Elementary	114	46.26	2.56
Middle school	95	45.99	3.04
High school	127	45.76	2.96
Central administration	44	45.95	2.74
Size of school district			
Under 1,000	62	45.35	2.85
1,000 to 2,999	55	46.65	2.58
3,000 to 5,999	82	45.73	3.24
6,000 to 9,999	71	46.42	2.38
10,000 or more	110	45.94	2.85
Sex			
Female	121	46.29	2.71
Male	262	45.87	2.91
Race			
Black	23	46.78	2.88
Hispanic	3	46.33	.58
White	360	45.95	2.84
Other	3	46.33	3.79
Administrative Experience			
1 to 3 years	74	46.07	2.93
4 to 10 years	152	45.89	3.17
11 to 20 years	118	46.19	2.35
21 or more years	22	44.82	2.02
Not an administrator	25	46.40	3.07
Content Area Taught			
English	31	46.35	3.61
Fine arts	16	45.94	2.62
Math	32	45.84	3.17
Physical education	36	45.50	2.13
Science	45	46.29	3.15
Self-contained elementary	93	46.31	2.35
Social studies	68	45.93	3.08
Special education	21	45.67	2.39
Vocational education	19	46.05	1.96
Not listed	29	45.41	3.61

Table 12.--Continued.

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Level of Teaching Experience			
Elementary	124	46.25	2.26
Middle school	77	45.99	2.87
High school	179	45.74	3.20

Table 13 shows the relationship between each predictor variable ("Source" column) and evaluation accuracy. The table shows that none of the demographic variables were significant predictors of accuracy. The overall prediction equation with all variables entered fails to provide a significant prediction of evaluation accuracy.

Table 13.--Test for effect of demographic variables on prediction of evaluation accuracy (exercise 4)

Source	DF	Sum of Squares	RSq Chg	F	Sig F
Assignment	3	5.58109	.00188	.22457	.8793
Size	4	54.23267	.01828	1.63667	.1646
Sex	1	.88304	.00030	.10660	.7443
Race	3	37.44969	.01262	1.50691	.2125
Experience	4	37.56562	.01266	1.13368	.3405
Content area	9	28.70992	.00968	.38508	.9420
Level taught	2	12.31422	.00415	.74326	.4763
Regression	26	191.84515		.89072	.6224
Residual	335	2775.12998			
Total	361	2966.97514			

$R = .25428$; $R^2 = .06466$; Adjusted $R^2 = -.00793$ ($p = .6224$)

The stepwise regression procedure failed to produce a prediction equation. In summary, data set 4 provides no evidence that the demographic variables are significantly related to evaluation accuracy.

Evaluation Exercise Five--Ninth Grade English

Table 14 presents the means and standard deviations of evaluation accuracy (total items correct out of 55). The categories of each predictor variable are shown with the descriptive information on evaluation accuracy for each one.

Table 14.--Evaluation accuracy on exercise 5 broken down by levels of the demographic variables

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Campus Assignment			
Elementary	158	29.95	1.61
Middle school	126	30.05	1.29
High school	162	29.97	1.55
Central administration	74	30.26	1.50
Size of school district			
Under 1,000	100	30.11	1.48
1,000 to 2,999	68	30.21	1.39
3,000 to 5,999	101	29.91	1.42
6,000 to 9,999	104	29.86	1.59
10,000 or more	147	30.07	1.56
Sex			
Female	169	30.06	1.40
Male	358	30.01	1.54
Race			
Black	28	29.71	1.18
Hispanic	4	29.50	1.73
White	500	30.04	1.51
Other	4	30.00	.82

Table 14.--Continued.

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Administrative Experience			
1 to 3 years	106	30.30	1.30
4 to 10 years	189	29.85	1.53
11 to 20 years	153	30.09	1.43
21 or more years	42	29.81	1.70
Not an administrator	48	29.96	1.64
Content Area Taught			
English	47	29.89	1.72
Fine arts	18	30.22	1.40
Math	44	29.84	1.41
Physical education	48	30.04	1.05
Science	58	29.81	1.53
Self-contained elementary	129	30.02	1.44
Social studies	88	30.11	1.68
Special education	33	30.12	1.56
Vocational education	27	30.11	1.34
Not listed	45	30.09	1.53
Level of Teaching Experience			
Elementary	174	30.03	1.41
Middle school	108	29.80	1.67
High school	244	30.08	1.47

Table 15 shows the relationship between each predictor variable ("Source" column) and evaluation accuracy. The table shows that none of the demographic variables are significant predictors of accuracy. The overall prediction equation with all variables entered fails to provide a significant prediction of evaluation accuracy.

Table 15.--Test for effect of demographic variables on prediction of evaluation accuracy (exercise 5)

Source	DF	Sum of Squares	RSq Chg	F	Sig F
Assignment	3	5.38425	.00467	.78127	.5048
Size	4	6.72932	.00584	.73234	.5702
Sex	1	.68547	.00060	.29839	.5851
Race	3	3.66005	.00318	.53109	.6611
Experience	4	21.26594	.01846	2.31432	.0566
Content area	9	9.97874	.00866	.48265	.8865
Level taught	2	7.20338	.00625	1.56785	.2096
Regression	26	53.73497		.89967	.6099
Residual	478	1098.06701			
Total	504	1151.80198			

$R = .21599$; $R^2 = .04665$; Adjusted $R^2 = -.00520$ ($p = .6099$)

The stepwise regression procedure failed to produce a prediction equation. In summary, data set 5 provides no evidence that the demographic variables are significantly related to evaluation accuracy.

Evaluation Exercise Six--Eighth Grade Language Arts

Table 16 presents the means and standard deviations of evaluation accuracy (total items correct out of 55). The categories of each predictor variable are shown with the descriptive information on evaluation accuracy for each one.

Table 16.--Evaluation accuracy on exercise 6 broken down by levels of the demographic variables

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Campus Assignment			
Elementary	70	50.96	2.14
Middle school	46	50.70	2.19
High school	69	50.90	2.66
Central administration	27	50.78	1.83
Size of school district			
Under 1,000	39	50.69	3.20
1,000 to 2,999	29	51.38	1.72
3,000 to 5,999	40	50.50	2.44
6,000 to 9,999	26	50.81	2.19
10,000 or more	78	50.95	1.86
Sex			
Female	80	50.30	2.87
Male	133	51.17	1.81
Race			
Black	12	49.75	4.41
Hispanic	2	47.50	6.36
White	202	50.95	2.03
Other	0	.	.
Administrative Experience			
1 to 3 years	44	50.93	2.22
4 to 10 years	73	50.81	2.71
11 to 20 years	66	50.85	1.82
21 or more years	16	51.00	2.16
Not an administrator	17	50.65	2.45
Content Area Taught			
English	19	50.47	1.71
Fine arts	8	52.25	1.04
Math	18	51.17	2.28
Physical education	19	51.05	1.68
Science	24	50.50	3.32
Self-contained elementary	53	51.21	1.98
Social studies	35	50.71	2.09
Special education	13	50.08	2.87
Vocational education	13	50.69	2.50
Not listed	14	50.29	2.76

Table 16.--Continued.

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Level of Teaching Experience			
Elementary	76	51.17	1.85
Middle school	44	51.11	2.12
High school	91	50.46	2.66

Table 17 shows the relationship between each predictor variable ("Source" column) and evaluation accuracy. The table shows that sex is a significant (.0036) predictor of accuracy. Reference to table 16 indicates that males had a higher average accuracy of evaluation (51.17) than did females (50.30). However, the overall prediction equation with all variables entered did not provide a significant prediction of evaluation accuracy. The adjusted R square for the equation is just under .05 indicating that just under 5 percent of the variance in evaluation accuracy can be predicted from all of the variables taken together. However, this percentage of variance accounted for is not statistically significant as table 17 shows (significance of F for regression = .1040).

Table 17.--Test for effect of demographic variables on prediction of evaluation accuracy (exercise 6)

Source	DF	Sum of Squares	RSq Chg	F	Sig F
Assignment	3	.75453	.00069	.04957	.9854
Size	4	15.87808	.01454	.78227	.5381
Sex	1	44.28544	.04055	8.72732	.0036
Race	2	26.76319	.02450	2.63711	.0743
Experience	4	18.04398	.01652	.88898	.4717
Content area	9	32.94538	.03016	.72139	.6889
Level taught	2	20.01594	.01833	1.97227	.1421
Regression	25	178.81225		1.40954	.1040
Residual	180	913.38193			
Total	205	1092.19417			

$R = .40462$; $R^2 = .16372$; Adjusted $R^2 = .04757$ ($p = .1040$)

The stepwise regression procedure yielded a prediction equation with three of the dummy coded variables as predictors. As table 18 shows, the overall adjusted R square was just under .07, indicating that nearly 7 percent of the variance in evaluation accuracy was accounted for by the three predictor variables.

Table 18.--Stepwise regression equation (exercise 6)

Variable	B	SE B	Beta	T	Sig T
Sex	1.08589	.34659	.22875	3.133	.0020
LevelExp1	.91501	.34861	.19008	2.625	.0093
Race3	1.34633	.67066	.13695	2.007	.0460
(Constant)	47.51132	.84590		56.167	.0000

$R = .28449$; $R^2 = .08094$; Adjusted $R^2 = .06729$ ($p = .0007$)

The number of the dummy coded variables corresponds to the order the categories appear in the tables of descriptive

statistics earlier. Thus Males, Elementary experience, and Whites were selected. The interpretation of the stepwise regression equation indicates that higher accuracy in evaluation in this data set was associated with 1) being male, 2) having teaching experience at the elementary level, and 3) being white.

In summary, there is little evidence that any of the demographic variables was a significant predictor of accuracy in evaluation in this data set. The overall prediction equation with all variables entered was not statistically significant. Even though the stepwise procedure was able to compute a statistically significant equation with fewer variables, the implication of this finding is uncertain since the overall equation was not significant.

Summary of Results of Six Data Sets

The following table summarizes the results of the analysis of the first six data sets. Two types of findings are summarized: (1) variables found to be significantly related to evaluation accuracy when the regression equation with all seven variables was significant, (2) variables which had one category significantly related to evaluation accuracy in the stepwise regression procedure.

From table 19 it is evident that no one variable was significantly related to evaluation accuracy in each data

set. Though race was significant in 4 of 6 analyses, only in exercise 2 was the full regression model significant.

Table 19.--Summary of findings on six data sets.

Variable	Data Set					
	1	2	3	4	5	6
Campus assignment		*				
Size of school district	#					
Sex			*#			#
Race	#	*#	#			#
Administrative experience			#			
Content area taught	#					
Level of teaching			#			#

* Variable significant in significant full regression model.
\# One level significant in stepwise regression.

Summary--Those With Data on All Six Tapes

Table 20 presents the means and standard deviations of evaluation accuracy (total items correct out of 55) for the summary data set. The categories of each predictor variable are shown with the descriptive information on evaluation accuracy for each one.

Table 20.--Evaluation accuracy on summary of six tapes broken down by levels of the demographic variables

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Campus Assignment			
Elementary	35	40.74	1.55
Middle school	30	40.93	2.00
High school	48	40.98	1.31
Central administration	14	41.12	1.09

Table 20.--Continued.

Demographic Variables	Evaluation Accuracy		
	N	Mean	S.D.
Size of school district			
Under 1,000	24	40.56	1.58
1,000 to 2,999	20	41.04	1.53
3,000 to 5,999	29	40.93	1.71
6,000 to 9,999	14	41.17	.85
10,000 or more	40	40.98	1.58
Sex			
Female	38	41.02	1.41
Male	88	40.88	1.59
Race			
Black	10	40.03	2.30
Hispanic	2	42.67	2.12
White	116	40.96	1.41
Other	0	.	.
Administrative Experience			
1 to 3 years	21	40.75	1.93
4 to 10 years	52	40.85	1.45
11 to 20 years	44	41.16	1.39
21 or more years	6	39.92	1.35
Not an administrator	5	41.43	1.55
Content Area Taught			
English	11	40.61	1.79
Fine arts	6	40.31	1.44
Math	12	41.19	1.03
Physical education	14	40.95	.90
Science	19	39.92	1.82
Self-contained elementary	25	41.01	1.42
Social studies	20	41.47	1.75
Special education	7	41.62	.61
Vocational education	7	40.88	1.63
Not listed	7	41.55	1.35
Level of Teaching Experience			
Elementary	38	41.17	1.31
Middle school	25	41.25	1.34
High school	60	40.57	1.69

Table 21 shows the relationship between each predictor variable ("Source" column) and evaluation accuracy. The table shows that there were no significant predictors of accuracy. The overall prediction equation with all variables entered does not provide a significant prediction of evaluation accuracy ($p = .2080$).

Table 21.--Test for effect of demographic variables on prediction of evaluation accuracy (summary data set)

Source	DF	Sum of Squares	RSq Chg	F	Sig F
Assignment	3	6.94744	.02425	1.02336	.3859
Size	4	4.75001	.01658	.52476	.7178
Sex	1	.06741	.00024	.02979	.8633
Race	2	3.98701	.01391	.88093	.4178
Experience	4	4.03282	.01407	.44552	.7754
Content area	9	34.99783	.12215	1.71839	.0953
Level taught	2	10.60127	.03700	2.34234	.1016
Regression	25	71.54521		1.26463	.2080
Residual	95	214.98142			
Total	120	286.52663			

$R = .49970$; $R^2 = .24970$; Adjusted $R^2 = .05225$ ($p = .2080$)

The stepwise regression procedure yielded a prediction equation with one of the dummy coded variables (content level 5) as predictor. As table 22 shows, the overall adjusted R square was just under .07, indicating that nearly 7 percent of the variance in evaluation accuracy was accounted for by the one predictor variable.

Table 22.--Stepwise regression equation (summary data set)

Variable	B	SE B	Beta	T	Sig T
Content5	-1.16065	.37284	-.27441	-3.113	.0023
(Constant)	41.08170	.14774		278.061	.0000

$R = .27441$; $R^2 = .07530$; Adjusted $R^2 = .06753$ ($p = .0023$)

Content5 is Science, indicating that those who had taught Science had lower evaluation accuracy for the combined data set than those who had not taught Science. In summary, there is little evidence that any of the demographic variables was a significant predictor of accuracy in evaluation in the summary data set. The overall prediction equation with all variables entered was not statistically significant. Even though the stepwise procedure was able to compute a statistically significant equation with one variable, the meaning of this finding is uncertain since the overall equation was not significant.

CHAPTER V

SUMMARY, FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

The purpose of this chapter is to provide a summary of the study, to state the findings, and to draw conclusions. Recommendations are then made based on the findings and conclusions.

Summary

This study was designed to determine if there are personal and demographic characteristics which can predict teacher appraiser accuracy. Specifically, the study tested the following hypotheses.

1. There is no significant relationship between campus-level assignment and appraiser's accuracy.

2. There is no significant relationship between the appraiser's district size and appraiser's accuracy.

3. There is no significant relationship between the appraiser's sex and appraiser's accuracy.

4. There is no significant relationship between the appraiser's race and appraiser's accuracy.

5. There is no significant relationship between the appraiser's number of years experience as an administrator and appraiser's accuracy.

6. There is no significant relationship between the appraiser's level of teaching experience and appraiser's accuracy.

7. There is no significant relationship between the curriculum area taught by the appraiser and appraiser's accuracy.

Chapter 2 reviews of the literature associated with teacher appraisers and teacher evaluation instruments. This review describes five areas which influence teacher appraisers in the evaluation process. These areas are: (1) the basis for teacher evaluation, (2) a description of the evaluation process, (3) the types, intents, and designs of appraisal instruments, (4) the training of teacher appraisers, and (5) the reliability and validity of teacher evaluation.

Chapter 3 describes the procedures for data collection and analysis of the data. It provides specific information concerning manipulation of the data through the use of a computer. Chapter 4 contains an explanation of the data. In order to determine what combination of demographic variables best predicted accurate evaluation, a multiple linear regression procedure was used. Each of the seven independent variables was tested as a predictor of accuracy in evaluation. In addition, the best prediction model using a combination of the independent variables was determined using stepwise regression. Where an independent variable

was significant, a follow-up ANOVA and Tukey's multiple comparison were employed to explain further the nature of the relationship. The level of significance utilized was the .05 level.

Chapter 5 summarizes the entire study, reports the findings, draws conclusions from the findings, and states recommendations from the study. The recommendations are based upon the researcher's opinions after analyzing the data.

Findings

Each of the six exercises and the summary data set discussed in Chapter 4 are reported in the following section.

1. There is no compelling evidence that any of the demographic variables were significant predictors of accuracy in evaluation in evaluation exercise one, ninth grade grammar. Whites performed less accurately than Blacks and Hispanics. However, the follow-up ANOVA did not reveal significance; therefore, this finding should be considered carefully. Although one variable was present in testing district size and content area, the overall tests for these variables were not significant.

2. There is evidence that race is a significant predictor of accuracy in evaluation in evaluation exercise two, third grade science. Hispanics showed more accuracy than

Whites or Blacks. However, the small number of Hispanics in the sample means that this result might not be repeated in other population groups. Also, Tukey's multiple comparison procedure showed that evaluators assigned to middle schools and high schools had significantly higher rating accuracy than did those assigned to central office administration. However, the stepwise regression procedure did not select these variables in the prediction equation. This omission lessens the prospect that campus assignment is a meaningful predictor of evaluation accuracy.

3. There is evidence that sex is a significant predictor of accuracy in evaluation in evaluation exercise three, eighth grade social studies. Females had higher average evaluation accuracy than males. The stepwise procedure identified three other data sets which showed significance: blacks showed less accuracy than Whites or Hispanics; appraisers with middle school experience had higher accuracy; and in years of experience, lower accuracy was demonstrated by those with 21 or more years of experience.

4. Evaluation exercise four, middle school mathematics, provided no evidence that the demographic variables are significantly related to teacher evaluation accuracy.

5. Evaluation exercise five, ninth grade English, provided no evidence that the demographic variables are significantly related to teacher evaluation accuracy.

6. In evaluation exercise six, eighth grade language arts, there is some evidence that persons who taught science were less accurate than persons who taught in other areas. However, the overall prediction equation with all variables entered was not statistically significant; thus, the implications of this finding is uncertain.

7. There is evidence that content area taught is a significant predictor of accuracy in evaluation in the summary data set. The stepwise procedure identified those evaluators who had taught science as being less accurate than other content area teachers. However, the overall prediction equation with all variables entered was not statistically significant thus the implications of this finding is uncertain.

These statistical findings comprised the basis for the acceptance or rejection of the stated hypotheses.

1. There was no statistical relationship between campus-level assignment and appraiser's accuracy.

2. There was no statistical relationship between the appraiser's district size and appraiser's accuracy.

3. There was no statistical relationship between the appraiser's sex and appraiser's accuracy.

4. There was no statistical relationship between the appraiser's race and appraiser's accuracy.

and appraiser's accuracy.

6. There was no statistical relationship between the appraiser's level of teaching experience and appraiser's accuracy.

7. There was no statistical relationship between the curriculum area taught by the appraiser and appraiser's accuracy.

All null hypotheses tested were retained.

Conclusions

In the process of conducting this study the following conclusions were reached which are worthy of mention.

1. The summary data set indicated there was little evidence that any of the demographic variables was a significant predictor of accuracy in the evaluation process. However, there was an indication that persons with a teaching background in the content area of science could possibly preclude an appraiser from being as accurate as other appraisers.

2. The six different exercise data sets indicated that varying instructional settings and methodologies can influence evaluator accuracy. The campus assignment, years of experience, content area taught, race, and sex of appraisers were all identified in at least one of the exercise sets as having significance. Except for sex and race none of the variables was found to be significant when the overall

prediction equation with all demographic variables entered was evaluated. Since whites were found to be the least accurate appraisers in Exercise 1 but the most accurate in Exercise 2, the significance of race should be questioned. The same conclusion would hold true for sex. Exercise 3 found females the most accurate appraisers; however Exercise 6 identified males as the most accurate. With the antonymous nature of these findings, the significance of sex and race should be questioned.

3. In the prediction equations of this study the percent of variance was so minute that social significance could not be established. As a result, this research failed to identify any specific, or combination of demographic variables which would predict an appraiser to be less accurate than any other appraiser.

4. The Texas Teacher Appraisal System is an appraisal system which can be used by appraisers with various backgrounds and experiences without a reduction of accuracy in the process.

5. School boards can appoint appraisers with various backgrounds and experiences without a reduction of accuracy in the evaluation process.

Recommendations

The following recommendations are made on the basis of

the findings, conclusions, and personal observations in the study.

1. The lack of legitimate research investigating teacher evaluation and appraiser accuracy is astonishing when one considers the emphasis being placed on the evaluation process by the educational community. State education agencies, legislatures, and educational organizations should contribute additional monetary and human resources in the area of teacher evaluation research.

2. The amount and type of training received by appraisers in this study could account for the lack of significant findings. Future studies should be conducted to determine what types of training are effective in improving appraiser accuracy and the amount of time required for appraiser training to improve appraiser accuracy.

3. Additional research should be conducted on the accuracy of appraiser evaluations as it correlates with student achievement.

4. The individual demographics of race and sex should be studied in future research. Specific studies should be conducted on the bias or lack of bias in appraisers of different races and sex.

5. Future studies should focus on knowledge concerning the teaching act and scores received by persons utilizing the TTAS training program.

APPENDIX A

Texas Education Agency

• STATE BOARD OF EDUCATION
• STATE COMMISSIONER OF EDUCATION
• STATE DEPARTMENT OF EDUCATION

201 East Eleventh Street
Austin, Texas
78701

April 15, 1985

TO THE EDUCATOR ADDRESSED:

The Texas Education Agency is requesting your participation in a Job-Relatedness Survey to be used in developing the teacher appraisal system mandated by House Bill 72. Teachers have been randomly selected to represent teaching assignments, teaching fields, years of experience, and other characteristics.

Development of the teacher appraisal system began in January 1985. To date, TEA and an advisory committee of Texas educators have met and developed a set of criteria based upon a review of research and surveys of evaluation practices in other states and in Texas. You are being asked to participate in the next step in the development process, the review of those criteria to ensure that the criteria are consistent with teaching practices in public schools in Texas.

Please complete the Job-Relatedness Survey and return it promptly. You may choose to mail it directly to TEA or to return it to your school district office. The district office will mail those which it has received by May 20, 1985. If you are unable or choose not to complete the survey, please return the materials and briefly state the reason why you have not completed the survey. In order for us to include your responses in the analysis, your survey should be returned by **May 20, 1985.**

We cannot overemphasize the importance of your contribution to the survey, both to the development of the teacher appraisal system and to the profession of teaching in Texas. Thank you in advance for your participation. If you have any questions about the survey or about the teacher appraisal system, please feel free to contact Susan Barnes, Director of Teacher Appraisal, at (512) 834-4242.

Sincerely,

W.N. Kirby
Interim Commissioner of Education

TEXAS EDUCATION AGENCY
Division of Professional Support

Job-Relatedness Survey for Teacher Appraisal

Authority for Data Collection: Texas Education Code 13.302(b)
Planned Use of the Data: Development of the statewide teacher appraisal system required in House Bill 72
Instructions: Please read the instructions carefully as you complete the questionnaire. If assistance is needed, call the Division of Professional Support, (512) 834-4242.

CHECK THE APPROPRIATE RESPONSE FOR EACH ITEM.

1. This survey is designed to be completed only by teachers holding a college degree. Do you hold a college degree?

   _____ Yes
   _____ No

2. This survey is designed to be completed only by teachers holding a Texas teaching certificate. Do you hold a Texas teaching certificate?

   _____ Yes
   _____ No

IF THE ANSWER TO EITHER OF THESE QUESTIONS IS "NO," DO NOT CONTINUE. RETURN THE QUESTIONNAIRE TO T.E.A. IN THE ENVELOPE PROVIDED EITHER DIRECTLY OR THROUGH YOUR DISTRICT.

IF THE ANSWER TO BOTH OF THESE QUESTIONS IS "YES," PLEASE CONTINUE.

My responses are representative of my understanding of the questions.

Printed Name of Respondent:

Teacher's Signature

By May 20, 1985, return to:
Texas Education Agency
Data Collection, Data Services
201 East 11th Street
Austin, Texas 78701

DIRECTIONS:
To complete this section of the survey, you will rate each behavior on the list in three ways: first, how frequently the behavior occurs in successful teaching; second, how important the behavior is to successful teaching; and third, whether the behavior is observable or not.

Each dimension should be rated independently. That is, do not let your rating of the IMPORTANCE of a behavior influence your rating of its FREQUENCY or its OBSERVABILITY. For example, a behavior may be very important but occur infrequently; the ratings for that behavior should be very different on those two dimensions.

Follow the steps below to complete this section of the survey:

1. For EACH behavior, indicate how frequently a successful teacher uses that behavior. FREQUENCY refers to how often the behavior occurs during the school year. Circle your response in Column 1 using the following scale:

   5 = very often
   4 = often
   3 = sometimes
   2 = seldom
   1 = very seldom

   Circle only one frequency rating for the behavior and move to the next step.

2. For EACH behavior, judge the importance of that behavior to successful performance as a teacher. IMPORTANCE refers to the value or significance of the behavior to successful teaching. Circle your response in Column 2 using the following scale:

   5 = extremely important
   4 = very important
   3 = important
   2 = not very important
   1 = not at all important

   Circle only one importance rating for the behavior and move to the next step.

3. For EACH behavior, judge the observability of that behavior. OBSERVABILITY means that the behavior can be seen in classroom teaching and/or can be evidenced through materials or documents. Circle your response in Column 3 using the following scale:

   yes = observable
   no = not observable

   Circle only one response to this question.

   EXAMPLE: In the example below, the respondent indicated that the behavior occurred "very often" in successful teaching, was "not at all important" to successful teaching, and was "observable."

Column 1 Frequency	Column 2 Importance	Column 3 Observability
1 2 3 4 (5)	(1) 2 3 4 5	(Y) N

4. Finally, on the last page of the booklet, list other behaviors that you sometimes use as a teacher and believe to be important and observable which have not been included in this survey.

Column 1. FREQUENCY	Column 2. IMPORTANCE	Column 3. OBSERVABILITY
5 = very often	5 = extremely important	yes = observable
4 = often	4 = very important	no = not observable
3 = sometimes	3 = important	
2 = seldom	2 = not very important	
1 = very seldom	1 = not at all important	

DOMAIN I. Teaching Skills

	Column 1 Frequency	Column 2 Importance	Column 3 Observability

CRITERION A. Planning and Preparation for Instruction

The successful teacher:

1. uses diagnostic information in planning	1 2 3 4 5	1 2 3 4 5	Y	N
2. selects or designs performance objectives	1 2 3 4 5	1 2 3 4 5	Y	N
3. selects or designs activities which are appropriate for stated objectives within the daily plan	1 2 3 4 5	1 2 3 4 5	Y	N
4. selects or designs materials for different needs	1 2 3 4 5	1 2 3 4 5	Y	N
5. selects or designs assignments for different needs	1 2 3 4 5	1 2 3 4 5	Y	N
6. selects or designs materials and procedures to assess learner progress	1 2 3 4 5	1 2 3 4 5	Y	N

CRITERION B. Delivery of Instruction

1. establishes focus of the lesson	1 2 3 4 5	1 2 3 4 5	Y	N
2. gives clear and easily understood directions	1 2 3 4 5	1 2 3 4 5	Y	N
3. demonstrates skills and processes	1 2 3 4 5	1 2 3 4 5	Y	N
4. provides opportunities for students to actively participate	1 2 3 4 5	1 2 3 4 5	Y	N
5. provides opportunities for review	1 2 3 4 5	1 2 3 4 5	Y	N
6. provides opportunities for practice	1 2 3 4 5	1 2 3 4 5	Y	N
7. monitors student understanding of instruction	1 2 3 4 5	1 2 3 4 5	Y	N
8. gives immediate feedback to students	1 2 3 4 5	1 2 3 4 5	Y	N
9. reteaches as needed	1 2 3 4 5	1 2 3 4 5	Y	N
10. varies cognitive levels of instruction	1 2 3 4 5	1 2 3 4 5	Y	N
11. provides closure	1 2 3 4 5	1 2 3 4 5	Y	N
12. relates lesson to previous or future lessons	1 2 3 4 5	1 2 3 4 5	Y	N

Column 1. FREQUENCY	Column 2. IMPORTANCE	Column 3. OBSERVABILITY
5 = very often	5 = extremely important	yes = observable
4 = often	4 = very important	no = not observable
3 = sometimes	3 = important	
2 = seldom	2 = not very important	
1 = very seldom	1 = not at all important	

DOMAIN I. Teaching Skills *(continued)*	Column 1 Frequency	Column 2 Importance	Column 3 Observability
CRITERION C. Evaluation of Instruction			
1. uses a variety of evaluation techniques	1 2 3 4 5	1 2 3 4 5	Y N
2. frequently evaluates progress of students	1 2 3 4 5	1 2 3 4 5	Y N
3. provides feedback concerning progress and achievement	1 2 3 4 5	1 2 3 4 5	Y N
4. uses evaluation techniques appropriate to stated objectives	1 2 3 4 5	1 2 3 4 5	Y N
CRITERION D. Motivation for Learning			
1. provides options to students in fulfilling assignments	1 2 3 4 5	1 2 3 4 5	Y N
2. varies learning activities	1 2 3 4 5	1 2 3 4 5	Y N
3. emphasizes value of learning activity	1 2 3 4 5	1 2 3 4 5	Y N
4. varies levels of concern	1 2 3 4 5	1 2 3 4 5	Y N
5. uses student contributions	1 2 3 4 5	1 2 3 4 5	Y N
6. reinforces learning efforts of students	1 2 3 4 5	1 2 3 4 5	Y N
7. holds students responsible for assignments	1 2 3 4 5	1 2 3 4 5	Y N
DOMAIN II. Classroom Management and Organization Skills			
CRITERION A. Management of Time Related to Instruction			
1. maintains quick and efficient roll check	1 2 3 4 5	1 2 3 4 5	Y N
2. begins lesson promptly	1 2 3 4 5	1 2 3 4 5	Y N
3. uses total time available	1 2 3 4 5	1 2 3 4 5	Y N
4. monitors students as they work	1 2 3 4 5	1 2 3 4 5	Y N
5. establishes realistic time limits for class activities	1 2 3 4 5	1 2 3 4 5	Y N
6. accomplishes smooth transitions between activities	1 2 3 4 5	1 2 3 4 5	Y N
7. uses procedures and routines which facilitate instruction	1 2 3 4 5	1 2 3 4 5	Y N

Column 1. FREQUENCY	Column 2. IMPORTANCE	Column 3. OBSERVABILITY
5 = very often	5 = extremely important	yes = observable
4 = often	4 = very important	no = not observable
3 = sometimes	3 = important	
2 = seldom	2 = not very important	
1 = very seldom	1 = not at all important	

	Column 1 Frequency	Column 2 Importance	Column 3 Observability
DOMAIN III. Knowledge of Subject Matter *(continued)*			
CRITERION B. Sequence			
1. selects or designs logical sequence of content within a lesson	1 2 3 4 5	1 2 3 4 5	Y N
2. selects or designs logical sequence of lessons within a unit	1 2 3 4 5	1 2 3 4 5	Y N
CRITERION C. Accuracy and Clarity			
1. presents and explains content accurately	1 2 3 4 5	1 2 3 4 5	Y N
2. presents and explains content clearly	1 2 3 4 5	1 2 3 4 5	Y N
3. gives concrete examples	1 2 3 4 5	1 2 3 4 5	Y N
4. responds knowledgeably to student questions	1 2 3 4 5	1 2 3 4 5	Y N
5. uses vocabulary appropriate to students	1 2 3 4 5	1 2 3 4 5	Y N
6. makes comparisons and points out patterns	1 2 3 4 5	1 2 3 4 5	Y N
CRITERION D. Relevance			
1. relates content to student interests	1 2 3 4 5	1 2 3 4 5	Y N
2. emphasizes value of content	1 2 3 4 5	1 2 3 4 5	Y N
DOMAIN IV. Interpersonal Skills			
CRITERION A. Oral and Written Communication			
1. uses correct syntax in oral communication	1 2 3 4 5	1 2 3 4 5	Y N
2. uses correct syntax and spelling in written materials	1 2 3 4 5	1 2 3 4 5	Y N
3. uses correct pronunciation	1 2 3 4 5	1 2 3 4 5	Y N
4. enunciates clearly	1 2 3 4 5	1 2 3 4 5	Y N
5. modulates voice level appropriately	1 2 3 4 5	1 2 3 4 5	Y N

Column 1. FREQUENCY	Column 2. IMPORTANCE	Column 3. OBSERVABILITY
5 = very often 4 = often 3 = sometimes 2 = seldom 1 = very seldom	5 = extremely important 4 = very important 3 = important 2 = not very important 1 = not at all important	yes = observable no = not observable

	Column 1 Frequency	Column 2 Importance	Column 3 Observability

DOMAIN II. Classroom Management and Organization Skills *(continued)*

CRITERION B. Classroom Organization

1. maintains seating arrangement/student grouping appropriate for the activity	1 2 3 4 5	1 2 3 4 5	Y N
2. has materials and/or facilities ready for use	1 2 3 4 5	1 2 3 4 5	Y N
3. posts daily or weekly schedules, or in other ways makes schedules available to students	1 2 3 4 5	1 2 3 4 5	Y N
4. posts daily or weekly assignments, or in other ways makes assignments available to students	1 2 3 4 5	1 2 3 4 5	Y N
5. revises schedules as needed	1 2 3 4 5	1 2 3 4 5	Y N
6. maintains orderly classroom environment	1 2 3 4 5	1 2 3 4 5	Y N

CRITERION C. Management of Student Behaviors and Discipline

1. specifies expectations for classroom behavior	1 2 3 4 5	1 2 3 4 5	Y N
2. monitors student classroom behavior	1 2 3 4 5	1 2 3 4 5	Y N
3. reinforces appropriate classroom behavior	1 2 3 4 5	1 2 3 4 5	Y N
4. is consistent in the application of class rules	1 2 3 4 5	1 2 3 4 5	Y N
5. corrects deviant classroom behavior	1 2 3 4 5	1 2 3 4 5	Y N

DOMAIN III. Knowledge of Subject Matter

CRITERION A. Essential Elements

1. incorporates the essential elements adopted by TEA in design of lesson	1 2 3 4 5	1 2 3 4 5	Y N
2. assesses mastery of essential elements adopted by TEA	1 2 3 4 5	1 2 3 4 5	Y N

Column 1. FREQUENCY	Column 2. IMPORTANCE	Column 3. OBSERVABILITY
5 = very often 4 = often 3 = sometimes 2 = seldom 1 = very seldom	5 = extremely important 4 = very important 3 = important 2 = not very important 1 = not at all important	yes = observable no = not observable

	Column 1 Frequency	Column 2 Importance	Column 3 Observability

DOMAIN IV. Interpersonal Skills *(continued)*

CRITERION B. Relationships With Students

1. models courteous behavior	1	2	3	4	5		1	2	3	4	5							Y		N
2. models appreciation/acceptance of individual differences	1	2	3	4	5		1	2	3	4	5							Y		N

DOMAIN V. Professional Characteristics

CRITERION A. Professional Responsibilities

1. works cooperatively with others	1	2	3	4	5		1	2	3	4	5							Y		N
2. complies with district policies and procedures	1	2	3	4	5		1	2	3	4	5							Y		N

CRITERION B. Professional Self-development

1. stays current in content taught	1	2	3	4	5		1	2	3	4	5							Y		N
2. stays current in instructional skills	1	2	3	4	5		1	2	3	4	5							Y		N

Job-Relatedness Survey Suggestions

Please use this page to make suggestions concerning this survey. For comments about specific items, refer to the item by Domain, Criterion letter, and behavior number.

If you feel that some behaviors which you sometimes use and believe to be important and observable have been excluded, please list them below.

APPENDIX B

TEXAS
TEACHER APPRAISAL SYSTEM

INSTRUMENT

Texas Education Agency

____ OBSERVATION RECORD _____ Date

____ EVALUATION RECORD _____ Date

School District_____ Co.-Dist. No._____

Teacher _____ #_____ School Year 19___-19___

Assignment/Grade _____ School _____

Appraisal 1 or 2 (Circle) Name of Appraiser _____ Teacher Supervisor or Other Appraiser (Circle One)

Observation Date _____ Beginning Time _____ Ending Time _____

Scheduled _____ Unscheduled _____ Subject Area(s) Observed _____

Directions: Circle the number(s) for each indicator. At least one column will be marked for each indicator. Column "A/BE" is used to note "Absent/Below Expectation" behaviors and has a credit value of 0. Column "SE" represents the "Standard Expectation" for each indicator and has a credit value of 1. Column "EQ" represents "Exceptional Quality" for each indicator and has a credit value of 1. To give credit to Column "EQ," credit must be given in Column "SE." Exceptional Quality credit is not applicable to some indicators. For these, a hyphen has been placed in Column "EQ." Provide documentation when marking either Column "A/BE" or "EQ." The teacher supervisor will compute credits at the end of the appraisal period, and the second appraiser will compute credits within 7 days of the formal observation. Corrections should be initialed by person filling out the form.

I. Instructional Strategies

		Columns A/BE	SE	EQ
1. Provides opportunities for students to participate actively and successfully.				
	a. varies activities	0	1	1
	b. interacts with students	0	1	1
	c. solicits participation	0	1	1
	d. extends	0	1	1
	e. provides time	0	1	1
	f. implements at appropriate level of difficulty	0	1	1
2. Evaluates and provides feedback on student progress during instruction.				
	a. communicates expectations	0	1	1
	b. monitors	0	1	1
	c. solicits responses for assessment	0	1	1
	d. reinforces	0	1	1
	e. provides corrective feedback	0	1	1
	f. reteaches	0	1	1

FOR EVALUATION RECORD
DOMAIN CREDIT TOTAL
(SE credits + EQ credits) [ ]

II. Classroom Management and Organization

		Columns A/BE	SE	EQ
3. Organizes materials and students.	a. secures student attention	0	1	1
	b. uses procedures/routines	0	1	1
	c. gives administrative directions	0	1	1
	d. uses seating/grouping	0	1	1
	e. has materials/aids/facilities ready	0	1	1
4. Maximizes amount of time available for instruction.	a. begins/ends	0	1	-
	b. implements sequence of activities	0	1	-
	c. maintains pace	0	1	-
	d. maintains focus	0	1	-
	e. keeps students engaged	0	1	-
5. Manages student behavior.	a. specifies expectations	0	1	-
	b. prevents off-task behavior	0	1	-
	c. redirects off-task behavior	0	1	-
	d. stops inappropriate behavior	0	1	-
	e. stops disruptive behavior	0	1	-
	f. applies rules	0	1	1
	g. reinforces appropriate behavior	0	1	1

FOR EVALUATION RECORD
DOMAIN CREDIT TOTAL
(SE credits + EQ credits)

III. Presentation of Subject Matter

6. Teaches for cognitive, affective, and/or psychomotor learning and transfer.	a. begins with introduction	0	1	1
	b. uses content sequence	0	1	1
	c. relates prior/future learning	0	1	1
	d. defines/describes	0	1	1
	e. elaborates critical attributes	0	1	1
	f. stresses generalization/principle/rule	0	1	1
	g. transfers	0	1	1
	h. closes instruction	0	1	1

III. Presentation of Subject Matter *(continued)*

		Columns A/SE	SE	EQ
7. Presents information accurately and clearly.	a. makes no significant errors	0	1	-
	b. uses appropriate vocabulary	0	1	-
	c. explains clearly	0	1	1
	d. stresses points/dimensions	0	1	1
	e. clarifies misunderstanding	0	1	1
8. Uses acceptable communication skills.	a. uses correct grammar	0	1	-
	b. pronounces correctly/clearly	0	1	-
	c. uses accurate language	0	1	-
	d. demonstrates written skills	0	1	-

FOR EVALUATION RECORD
DOMAIN CREDIT TOTAL
(SE credits + EQ credits)

IV. Learning Environment

9. Uses strategies to motivate students for learning.	a. relates to interests	0	1	1
	b. emphasizes value/importance	0	1	1
	c. reinforces efforts	0	1	1
	d. challenges	0	1	1
10. Maintains supportive environment.	a. avoids sarcasm/negative criticism	0	1	-
	b. maintains courteous climate	0	1	1
	c. encourages	0	1	1
	d. praises	0	1	1
	e. establishes rapport	0	1	1

FOR EVALUATION RECORD
DOMAIN CREDIT TOTAL
(SE credits + EQ credits)

THE TEXAS TEACHER APPRAISAL INSTRUMENT: EXPLANATIONS AND EXPECTATION STATEMENTS

Indicator statements establish expectations for teacher performance. Each performance indicator is scored by considering the strength and comprehensiveness of the preponderance of evidence related to that indicator. The evidence gathered is evaluated for both quality and quantity of certain teacher behaviors. The quality or effectiveness of teaching behavior will be judged by its observed impact upon student behavior and the apparent success of students engaged in learning activities.

Some indicators, such as closing the lesson, are expected to occur a minimum of once, and credit is given when the indicator occurs. For other indicators, the observer should consider all of the occasions when a teacher demonstrated the skill or characteristic of the indicator in an effective way and all of the occasions when the teacher might have demonstrated the skill or characteristic but did not.

Terms enclosed within parentheses cross reference the terminology frequently used in instructional leadership training to the teacher appraisal instrument. The inclusion of these terms is not a mandate by the state for a particular model of instruction. The referenced concepts are not requirements for receiving credit for an indicator.

Domain I. Instructional Strategies

Criterion 1. provides opportunities for students to participate actively and successfully

---

Performance Indicators	Explanation/Examples
a. appropriately varies activities	a. The teacher actively explains/demonstrates and also provides an opportunity for active student participation. Students participate in ways other than passive listening. (modeling/active participation)
b. interacts with students in group formats as appropriate	b. The teacher interacts with students in more than one group format, i.e., large group, small group, individual, if appropriate. If the teacher is responsible for only one learner, credit is automatically given.

c. solicits student participation	c. The teacher pursues student contributions, demonstrations, and questions with frequency appropriate for the lesson and the learners. The teacher may prompt, rephrase, and call on non-volunteers to increase student participation. (active participation)
d. extends students' responses/contributions	d. The teacher asks a student to give additional information based on a student response/ contribution or the teacher provides such information. (questioning/prompts)
e. provides ample time for students to respond to teacher questions/solicitations and to consider content as it is presented	e. The teacher provides ample time for students to consider information presented, to answer questions asked, and to formulate ideas, responses, and/or contributions. (wait time)
f. implements instruction at an appropriate level of difficulty	f. Scoring for this indicator is based upon the observed effects of instruction upon student performance. For example, if all but a few students appear to understand explanations, are successful in group practice activities, and are able to begin individual assignments without clarification, credit should be given for this indicator. If many students have difficulty performing a task, carrying out assignments, answering questions, and the like, then credit should be denied. The activities or explanations may also be too easy for students; in this case credit should also be denied.

Domain I: Instructional Strategies

Criterion 2. evaluates and provides feedback on student progress during instruction

Performance Indicators	Explanation/Examples
a. communicates learning expectations	a. The teacher indicates standards of success. Communicating to students what they are to accomplish as a result of the lesson/learning activity is sufficient to give credit. (objectives)
b. monitors students' performances as they engage in learning activities	b. The teacher does not miss opportunities to verify that students understand or can perform skill/process. Assigning classwork and failing to circulate to examine student work or performance is cause for no credit.
c. solicits responses or demonstrations from specific students for assessment purposes	c. The teacher may ask questions or have students show steps of a process or skill. Emphasis is on assessment of individual progress toward and/or accomplishment of lesson objectives.
d. reinforces correct responses	d. The teacher tells students when and whose performance is adequate and identifies those aspects of the performance which are adequate.

e. provides corrective feedback, or none needed	e. When student misunderstanding occurs, the teacher takes time to correct it or allows other students to correct it. The teacher tells students when performance is inadequate, identifies specific misunderstandings, and provides suggestions for improvement. Simply informing students that they are "right" or "wrong" is not sufficient to receive credit.(feedback)
f. reteaches, or none needed	f. When ongoing progress checks or other monitoring/assessment methods indicate misunderstanding or other student problems, the teacher instructs using different methods or techniques to explain/demonstrate the same content.

Domain II: Classroom Management and Organization

Criterion 3. organizes materials and students

Performance Indicators	Explanation/Examples
a. secures student attention, or students are attending	a. When directions are given for any activity, students are listening. (focus; management; attention)
b. uses administrative procedures and routines which facilitate instruction	b. The distribution and collection of materials, use of classroom areas, and student movement within the classroom are efficiently managed. (time on task)
c. gives clear administrative directions for classroom procedures or routines, or none needed	c. The teacher communicates to the students what activities and/or tasks are to be done; when, where, and how the activities and/or tasks are to be done; and who will be involved in the activities and/or tasks. (expectations)
d. maintains seating arrangement/grouping appropriate for the activity and the environment	d. Students are able to focus on instruction without difficulty or distraction. Each student has adequate space in which to work without distraction. (time on task)
e. has materials, aids, and facilities ready for use	e. The teacher ensures that a sufficient number of handouts are assembled, audiovisual equipment is set up, transparencies are prepared for use, and tables, desks, and chairs are arranged for the first activity. Everything is ready to use. (time on task)

Domain II: Classroom Management and Organization

Criterion 4. maximizes amount of time available for instruction

Performance Indicators	Explanation/Examples
a. begins promptly/avoids wasting time at the end of the instructional period	a. Clerical routines are completed quickly so that time is not wasted before beginning an activity. The teacher should use the full time available. If students and/or the teacher "run out" of things to do and instructional time is wasted, credit should not be given. (time on task)
b. implements appropriate sequence of activities	b. The activities occur in such an order that students have the necessary background and information to follow instructions or complete assignments. For example, diagnostic activities precede rather than follow homework assignments or practice follows rather than precedes the introduction of a skill. (task analysis of activities)
c. maintains appropriate pace	c. The teacher allots adequate time for activities, does not overdwell in presentation, interaction, and questioning. Attention is also given to allowing sufficient time rather than hurrying through the instructional activity.
d. maintains focus	d. The teacher maintains commitment by staying on the topic in teacher-centered activities and does not interrupt student-centered activities unnecessarily. Focus can also be lost through delays, unnecessary digressions, and lengthy transitions. (task commitment)
e. keeps students engaged	e. Most (85%) of the students are engaged in learning activities for the instructional period. (task commitment)

Domain II: Classroom Management and Organization

Criterion 5. manages student behavior

Performance Indicators	Explanation/Examples
a. specifies expectations for class behavior, or none needed	a. The teacher explains expectations for behavior and gives reasons for students to behave in a certain way. Appropriate student behavior may indicate that expectations have been made clear. However, if inappropriate behavior occurs without subsequent statement or clarification of expectations, no credit should be given. Inappropriate behavior is not consistent with accepted norms or with teacher expectations. The definitions of appropriate and inappropriate behavior vary with the context of instruction. Common inappropriate behaviors include noisy activity, out-of-seat behavior, noisy callouts, and misuse of equipment. In indicators a, d, f, and g inappropriate behavior does not include passive off-task behavior. (management; expectations)
b. uses techniques to prevent off-task behavior, or none needed	b. The teacher observes students and acts to maintain student attention and participation before any off-task behavior occurs.
c. uses techniques to redirect persistent off-task behavior, or none needed	c. The teacher accurately identifies student(s) who are doing something other than the assigned task. The teacher acts quickly to redirect student(s) to the assigned task. (proximity)

d. uses techniques to stop inappropriate behavior, or none needed	d. The teacher indicates to specific student(s) that behavior is inappropriate or inconsistent with teacher expectations, e.g., responding without raising hands when hand raising has been requested. (negative reinforcement; extinction)
e. uses techniques to stop disruptive behavior, or none needed	e. The teacher accurately identifies disruptive student(s) and then acts quickly to stop the behavior. Disruptive behavior distracts one or more students from learning tasks and/or interrupts instruction. The teacher clearly identifies the disruptive behavior, and chooses strategies which minimize disruption of the rest of the class. (negative reinforcement)
f. applies rules consistently and fairly	f. The teacher treats students equitably and maintains consistent expectations for behavior. For example, the teacher does not repeatedly act to correct behavior of a particular student while ignoring the same behavior by another student. (rules)
g. reinforces desired behavior when appropriate	g. The teacher offers specific praise to individuals and/or to the class and reinforces those aspects of behavior which are acceptable. Reinforcement may be nonverbal. Credit should be given if no inappropriate behavior occurs, and reinforcement is judged unnecessary. (positive reinforcement)

Domain III: Presentation of Subject Matter

Criterion 6. teaches for cognitive, affective, and/or psychomotor learning and transfer

The focus may be on cognitive, affective, and/or psychomotor learning. The format may be teacher- or student-centered. The teacher may perform or ask students to perform any or all steps. Student behavior and/or success may be another indicator of success of previous instruction which should be considered in scoring this criterion. If the observer sees "e" and/or "f" and there is an indication that "c", "d", and/or "e" have been taught at a previous time, then credit may be awarded for c, d, and/or e.

Performance Indicators	Explanation/Examples
a. begins instruction/activity with an appropriate introduction	a. The teacher begins with an introduction which directs student attention to the content/purpose of instruction/activity.
b. presents information in an appropriate sequence	b. In instruction for cognitive learning, the teacher moves from simple to complex, concrete to abstract, specific to general (induction), or from complex to simple, abstract to concrete, general to specific (deduction). Emphasis is on student understanding of how parts relate to the whole; specific information is "anchored" to abstract ideas (concepts) and principles/generalizations/ rules. For skill development, the teacher provides 1) explanation, 2) demonstration/modeling, and 3) guided and independent student practice. A step-by-step approach is used. For affective learning, the teacher provides activities which allow the student to 1) explore group/societal interests, attitudes, or opinions and 2) to examine personal interests, attitudes, or opinions within the larger group/societal context; or 1) examine personal interests, attitudes, or opinions and 2) to relate these to group/societal interests, attitudes, or opinions.

c. relates content to prior or future learning	c. As new information, ideas, concepts, and skills are developed and/or new attitudes and interests are examined, they are placed into a meaningful framework for students. The teacher may state the relationships or may provide for students to draw relationships. References to or comparisons with prior and/or future learning can facilitate transfer. For example, introducing the concept of 3/4 time in a music lesson is related to the previous learning of 4/4 time. (transfer)
d. provides for definition of concepts and description of skills and/or attitudes and interests	d. As new concepts, skills, and/or attitudes and interests are introduced, they are given sufficient definition by either the teacher or students. Definitions or descriptions may be oral or written depending upon the complexity of the concept, skill, value, or attitude under consideration and on the nature of the students. Age and ability of students are important considerations in the complexity of definitions/descriptions provided. Young learners, for example, need shorter and less complex definitions than older learners. (explanation, definition)
e. provides for elaboration of critical attributes of concepts, skills, and/or attitudes and interests	e. Critical attributes are elaborated as new concepts, and/or attitudes and interests are introduced; steps or components of a skill are demonstrated. One appropriate strategy is contrasting examples and non-examples. Exploration of similarities and differences between previously acquired and new concepts or skills and/or previously examined attitudes may be desirable. The teacher may elaborate or ask students to elaborate. (explanation, critical attributes)

f. stresses generalization, principle, or rule as a relationship between or among concepts, skills, or attitudes/ interests	f. Generalizations, principles, and rules are statements of relationships between or among concepts. The teacher or student may explain and emphasize these relationships when generalizations, principles, or rules are taught. Cause/effect and ends/means relationships are identified when appropriate. (process steps, rules)
g. provides opportunities for transfer	g. Transfer is achieved in the application of newly acquired learning. The teacher may do this in a variety of ways. For example, the teacher may: use hypothetical or real new examples, demonstrate how a rule applies to a new case, or use a skill or concept in a new setting. Simple drill and practice is not considered transfer. (transfer)
h. closes instruction appropriately	h. The teacher may briefly summarize or ask students to summarize main points and explain how learning will be needed in the future. Closure may take place at the conclusion of any segment of instruction or at the end of the class period. If several content topics/activities occur and only a few are appropriately closed, no credit should be given. Closure need not be lengthy but must be observable. No credit is given for only administrative closure. (management, closure)

Domain III: Presentation of Subject Matter

Criterion 7. presents information accurately and clearly

Performance Indicators	Explanation/Examples
a. makes no significant errors	a. No major errors in teacher presentation of content are observed. A significant error is one which interferes with student understanding or one which is a distortion of fact. If an error is made but corrected by the teacher, credit is given for this indicator.
b. uses vocabulary appropriate to students	b. The teacher uses a simple term as a synonym/ explanation for a more complex term as new vocabulary is introduced.
c. explains content and/or learning tasks clearly	c. Teacher explanations of content are understandable. The teacher explains steps to be followed, provides examples of completed work, identifies potential areas of difficulty, and/or clarifies previously given directions about the task. If student performance/ behavior indicates that most students understand, credit should be given. (explanation, checking information)
d. stresses important points and dimensions of content	d. The teacher uses strategies to emphasize to the students the structure of the content. For example, the teacher uses voice inflection, underlines important points, repeats points for emphasis, and/or explains relationships. If instruction proceeds without some points standing out, important dimensions have not been adequately specified.
e. clarifies student misunderstanding, or none needed.	e. The teacher explains or demonstrates some point or procedure again after student questions or responses indicate that the student misunderstands.

Domain III: Presentation of Subject Matter

Criterion 8. uses acceptable communication skills in presentation

Performance Indicators	Explanation/Examples
a. uses correct grammar	a. Typical errors are 1) use of double negatives, 2) lack of subject-verb agreement, 3) incorrect verb tense, and 4) incorrect pronoun reference. Two or more errors are cause for denying credit.
b. pronounces words correctly and clearly	b. The teacher uses correct vowel/consonant/ diphthong sounds and emphasizes correct syllables. Speech is free of slurring or mumbling of words. The volume and rate of speech is at a level at which all students in the classroom can hear and understand.
c. uses accurate language	c. The teacher does not overuse indefinite or vague terms in presentations and verbal interactions, e.g., false starts, interrupters, qualifiers, or distractors.
d. demonstrates skill in written communication	d. The data source for this indicator will be the written information that is examined or viewed during the observation. Credit is given if no samples are available during the observation. Words should be spelled correctly, and sentences should be structured correctly. Two errors--spelling, grammar, sentence construction, and/or typographical --are cause for denial of credit for this indicator.

Domain IV: Learning Environment

Criterion 9. uses strategies to motivate students for learning

Performance Indicators	Explanation/Examples
a. relates content to student interests/ experiences	a. The teacher may use real or hypothetical examples/cases to increase student perception of relevance of content. Relating students' experiences to lesson content is an effective motivational strategy. (relevance/interest)
b. emphasizes the value/importance of the activity or content	b. The teacher stresses the value or importance of an activity or of content to the content field, to society, or to the student personally. (purpose)
c. reinforces learning efforts of students	c. The teacher may, in a variety of ways, communicate awareness and appreciation of student effort and progress. The teacher acknowledges and encourages students' task/ learning related efforts. The effect that is desired is an increased attempt on the part of students to participate actively. (positive reinforcement).
d. challenges students	d. Challenge is accomplished when the teacher communicates that elements or aspects of a learning task/activity may require extra effort. Challenge may be accomplished through the pace of the lesson, the level of difficulty/complexity of the content or task, or through overt verbal challenge. (expectations; level of concern)

Domain IV: Learning Environment

Criterion 10. maintains supportive environment

Performance Indicators	Explanation/Examples
a. avoids sarcasm and negative criticism	a. Comments to or about learners which personally demean or embarrass them should be avoided. One occurrence is sufficient evidence for denying credit. (learning climate)
b. establishes climate of courtesy and respect	b. The teacher listens to and responds to student questions, requires that students listen to each other in class interactions, encourages cooperation, and models courtesy. All interactions with students should model courtesy. (learning climate)
c. encourages slow and reluctant students	c. The teacher recognizes students who have difficulty in performance, is patient in interaction with these students, and positively reinforces their learning efforts.
d. provides praise for specific performance	d. The teacher singles out specific students or groups and cites specific performance(s). (positive reinforcement)
e. establishes and maintains positive rapport with students	e. The teacher relates to students in a pleasant manner and secures cooperation from the students. The teacher may use student names, make eye contact, smile, use a positive tone of voice, or stand near students, for example.

Domain V. Growth and Responsibilities

Criterion 11. plans for and engages in professional development

Performance Indicators

a. shows progress in completing professional growth requirements as agreed upon with appraiser(s), or none needed

b. stays current in content taught

c. stays current in instructional methodology

Domain V. Growth and Responsibilities

Criterion 12. interacts and communicates effectively with parents

Performance Indicators

a. initiates communications with parents about student performance and/or behavior when appropriate

b. conducts parent-teacher conferences in accordance with local district policy

c. reports student progress to parents in accordance with local district policy

d. maintains confidentiality unless disclosure is required by law

Domain V. Growth and Responsibilities

Criterion 13. complies with policies, operating procedures, and requirements

Performance Indicators

a. follows statutory and Texas Education Agency regulations

b. follows district and campus policies and procedures

c. performs assigned professional duties

d. follows district promotion/retention policy and procedures

Domain V. Growth and Responsibilities

Criterion 14. promotes and evaluates student growth

Performance Indicators

a. participates in campus goal setting for student progress

b. plans instruction in accordance with district requirements

c. documents student progress

d. maintains accurate records

e. reports student progress at appropriate intervals

BIBLIOGRAPHY

Books

Bailey, Kenneth D. Methods of Social Research. 3rd Edition. New York, New York: The Free Press, 1987.

Barber, B. The Professions in America. Boston: Houghton Mifflin, 1965.

Bolton, Dale L. Selection and Evaluation of Teachers. Berkeley, California: McCutchan Publishing Corporation, 1973.

Brown, Bob Burton. The Experimental Mind in Education. New York, New York: Harper & Row, Publishers, 1968.

Gage, N. L. The Scientific Basis of the Art of Teaching. New York: Teachers College Press, 1978.

House, Ernest R. Evaluating With Validity. Beverly Hills, California: Sage Publications, Inc., 1980.

Kachigan, Sam Kash. Statistical Analysis: An Interdisciplinary Introduction to Univariate and Multivariate Methods. New York, New York: Radius Press, 1986.

Medley, Donald M. Teacher Competency Testing and the Teacher Educator. Charlottesville, Virginia: Bureau of Educational Research, 1982.

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

Phillips, Derek L. Knowledge from What? Theories and Methods in Social Research. Chicago: Rand McNally, 1971.

Selltiz, Claire, Lawrence J. Wrightman, and Stuart W. Cook. Research Methods in Social Relations. Third ed. New York, New York: Holt, Rinehart & Winston, 1976.

Texas School Law Bulletin. Austin: West Publishing Company, 1988.

Thomas, Fred L. and Jon I. Young. An Introduction to Educational Statistics: The Essential Elements. Lexington, Massachusetts: Ginn Press, 1986.

Articles

Ban, John R., and John R. Soudah. A new model for professionalizing teacher evaluation. Peabody Journal of Education 56, (October, 1978): 24-32.

Barth, Roland S. Teacher evaluation and staff development. Principal 58 (January, 1979): 74-77.

Bird, Tom and Judith Warren Little. How schools organize the teaching occupation. The Elementary School Journal 86 (March, 1986): 493-511.

Brandt, Ronald S. On the expert teacher: A conversation with David Berliner. Educational Leadership 44 (October, 1987): 4-9.

Bridges, Edwin M. Managing the incompetent teacher--What can principals do? NASSP Bulletin 69 (February, 1985): 57-65.

Brieschke, Patricia A. The administrative role in teacher competency. The Urban Review 18 (No. 4, 1986): 237-251.

Buttram, Joan L. and Bruce L. Wilson. Promising trends in teacher evaluation. Educational Leadership 44 (April, 1987): 4-6.

Coker, Homer. The accuracy of principal's judgments of teacher performance. Journal of Educational Research 80 (March-April, 1987): 242-247.

Coker, Homer, Donald M. Medley, and Rober S. Soar. How valid are expert opinions about effective teaching? Phi Delta Kappan 62 (October, 1980): 131-149.

Cruickshank, Donald R. Profile of an effective teacher. Educational Horizons 64 (Winter, 1986): 80-86.

Darling-Hammond, Linda. A proposal for evaluation in the teaching profession. The Elementary School Journal 86 (March, 1986): 531-551.

Darling-Hammond, Arthur E. Wise, and Sara R. Pease. Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research 53 (Fall, 1983): 285-327.

Duckett, Bruce, William Strother, and Edward Gephart. Evaluating the evaluators. Practical Applications of Research 5 (March, 1982): 1-12.

Emmer, Edmund T. and Robert F. Peck. Dimensions of classroom behavior. Journal of Educational Psychology 64 (April, 1973): 223-240.

English, Fenwick W. Still searching for excellence. Educational Leadership 42 (December-January, 1985): 34-35.

Faast, Dorothy A. Appraiser training: Teacher performance evaluation training for school administrators. The Clearing House 58 (November, 1984): 128-130.

Fuller, Bruce, Ken Wood, Tamar Rapoport, and Sanford M. Dornbusch. The organizational context of individual efficacy. Review of Educational Research 52 (Spring, 1982): 7-30.

Hampel, Robert L. The political side of reform: Are conflicts, power struggles likely to occur? NASSP Bulletin 70 (December, 1986): 55-64.

Johnston, Gladys Styles and Carol Camp Yeakey. Supervision of teacher evaluation: Brief overview. Journal of Teacher Education 30 (March-April, 1979): 17-22.

Kauchak, Don, Ken Peterson, and Amy Driscoll. An interview study of teachers' attitudes toward teacher evaluation practices. Journal of Research and Development in Education 19 (Fall, 1985): 32-37.

Kult, Lawrence E. Improving teacher evaluations by principals. The Clearing House 52 (September, 1978): 17-21.

Laing, Steven O. The principal and evaluation. NASSP Bulletin 70 (November, 1986): 91-93.

Lareau, Annette. A comparison of professional examinations in seven fields: Implications for the teaching profession. The Elementary School Journal 86 (March, 1986): 553-569.

McGaw, Barry, James L. Wardrop, and Mary Anne Bunda. Classroom observation schemes: Where are the errors? American Educational Research Journal 9 (January, 1972): 13-27.

McGreal, Tom. How well can we truly evaluate teachers? The School Administrator 29 (January, 1986): 10-12.

McLaughlin, Milbrey Wallin. Teacher evaluation and school improvement. Teachers College Record 86 (Fall, 1984): 193-207.

McNally, Harold J. Performance-based teacher evaluation. NASSP Bulletin 61 (October, 1977): 104-105.

Magoon, A. Jon. Sensitive field observation of teaching performance. Journal of Teacher Education 30 (March-April, 1979): 13-16.

Martin, Jack. Developing category observation instruments for the analysis of classroom behavior. Journal of Classroom Interaction 12 (December, 1976): 1-16.

Medley, Donald M. and Homer Coker. The accuracy of principals' judgments of teacher performance. Journal of Educational Research 80 (March/April, 1987): 242-247.

Medley, Donald M. and Homer Coker. How valid are principals' judgments of teacher effectiveness? Phi Delta Kappan 69 (October, 1987): 138-140.

Moxley, Roy A. Teacher Evaluation: Images and analysis. Journal of Teacher Education 29 (November-December, 1978): 61-66.

Natriello, Gary. Teachers' perceptions of the frequency of evaluation and assessments of their effort and effectiveness. American Educational Research Journal 21 (Fall, 1984): 579-595.

Pellicer, Leonard O. and O. B. Hendrix. A practical approach to remediation and dismissal. NASSP Bulletin 64 (March): 57-62.

Peterson, Donovan. Legal and ethical issues of teacher evaluation: A research based approach. Educational Research Quarterly 7 (Winter, 1983): 6-16.

Peterson, Donovan and Kathryn Peterson. A research-based approach to teacher evaluation. NASSP Bulletin 68 (February, 1984): 39-46.

Peterson, Ken. Methodological problems in teacher evaluation. Journal of Research and Development in Education 17 (Summer, 1984): 62-70.

Pigford, Aretha Butler. Teacher evaluation: More than a game that principals play. Phi Delta Kappan 69 (October, 1987): 141-142.

Rand Corporation. These are the elements of a sound teacher evaluation system. American School Board Journal 172 (July, 1985): 25.

Rowley, Glenn. The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. Journal of Educational Measurement 15 (Fall, 1978): 165-180.

Roy, Joseph J. Teacher evaluation in an era of educational change. The Clearing House 52 (February, 1979): 275-276.

Sapone, Carmelo V. An appraisal and evaluation system for teachers and administrators. Educational Technology (May, 1980): 44-49.

Savage, John G. Better ways to evaluate teachers. North Central Association Quarterly 59 (Summer, 1984): 14-17.

Seeley, David S. Reducing the confrontation over teacher accountability. Phi Delta Kappan 61 (December, 1979): 248-251.

Shulman, Lee S. Knowledge and teaching: Foundations of the new reform. Harvard Educational Review 57 (February, 1987): 1-22.

Soar, Robert S. Accountability: Assessment problems and possibilities. The Journal of Teacher Education 24 (February, 1975): 205-212.

Stodolsky, Susan S. Teacher evaluation: The limits of looking. Educational Researcher 13 (November, 1984): 11-18.

Tracey, William R. Teacher evaluation--another perspective. The Clearing House 51 (January, 1978): 240-242.

White, Kinnard, Marvin D. Wyne, Gary B. Stuck, and Richard H. Coop. Assessing teacher performance using an observation instrument based on research findings. NASSP Bulletin 71 (March, 1987): 89-95.

Wickert, Donald M. Using teacher evaluation for improving instruction. The Clearing House 61 (September, 1987): 23-24.

Wise, Arthur E. and Linda Darling-Hammond. Teacher evaluation and teacher professionalism. Educational Leadership 42 (December-January, 1985): 28-33.

Wise, Arthur E., Linda Darling-Hammond, Milbrey W. McLaughlin, and Harriet T. Bernstein. Teacher evaluation: A study of effective practices. The Elementary School Journal 86 (September, 1985): 61-86.

Withall, John, and Fred H. Wood. Taking the threat out of classroom observation and feedback. Journal of Teacher Education 30 (January-February, 1979): 55-58.

Wuhs, Susan K. The pace of mandated teacher evaluation picks up. The American School Board Journal 170 (May, 1983): 28.

Government Documents

Dukakis, Michael S. Acceptance speech for the Democratic Parties' nomination for the President of the United States of America. Atlanta, Georgia: July 21, 1988.

National Commission on Excellence in Education. A Nation at Risk: The Imperative for Educational Reform. U.S. Department of Education, 1983.

Task Force on Education for Economic Growth, Education Commission of the States. Action for Excellence. (June, 1983).

Texas Education Agency. Introductory information. Mimeographed memo. Austin, Texas: Texas Education Agency, 1985.

Texas Education Agency. Texas Teacher Appraisal System Appraisal Manual. Austin, Texas: Texas Education Agency, 1986.

Unpublished Materials

Beckham, Joseph C. Legal aspects of teacher evaluation. National Organization on Legal Problems of Education. Topeka, Kansas, 1981.

Brown, Bob Burton and Jeaninne N. Webb. The use of classroom observation techniques in the evaluation of educational programs, 1975. ERIC, ED 117 192.

Brown, Bob Burton, William Mendenhall, and Robert Beaver. The reliability of observations of teachers' classroom behavior, 1967. ERIC, ED 011 520.

Coker, Homer. A study of the correlation between principals' ratings of teacher effectiveness and pupil growth, 1985. ERIC, ED 259 460.

Jackson, Mary E. Teacher evaluation: The role of the principal, 1986. ERIC, ED 275 049.

Knapp, Michael S. Toward the study of teacher evaluation as an organizational process: A review of current research and practice. Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York (March, 1982): 1-26.

Shanker, A. A call for professionalism. January 29 speech to the National Press Club. Washington, DC: American Federation of Teachers, 1985.

Webb, Jeaninne Nelson and Bob Burton Brown. Establishing reliability and validity estimates for systematic classroom observation, 1969. ERIC, ED 028 123.