A QUASI-EXPERIMENTAL STUDY OF INTER-RATER RELIABILITY WHEN AWARDING EXCEPTIONAL QUALITY POINTS ON THE TEXAS TEACHER APPRAISAL INSTRUMENT

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

Louann Dobbs, B.S., M.S.
Denton, Texas

August, 1989

Dobbs, Louann, A Quasi-Experimental Study of Inter-rater Reliability When Awarding Exceptional Quality Points on the Texas Teacher Appraisal Instrument, Doctor of Education (Leadership and Supervision), August, 1989, 135 pp., 19 tables, bibliography, 84 titles.

This study investigates the inter-rater reliability of appraisers who award exceptional quality points on the Texas Teacher Appraisal Instrument. Inter-rater reliability was measured when appraisers scored exceptional quality points after viewing a videotaped lesson. Comparisons were made between appraisers when grouped according to elementary or secondary certification, sex, years of administrative experience, and type of training. A total of 707 subjects from 56 school districts participated in the study. Five research hypotheses were formulated with the .05 level of significance for acceptance. All hypotheses were tested by correlation of coefficients, multiple response procedures, frequencies, and percentages.

The data measuring inter-rater reliability of the appraisers in training imply that there is very little reliability in the awarding of exceptional quality points on the Texas Teacher Appraisal Instrument. The findings of this study are that certification, sex, administrative experience, and type of training made no significant

differences when scoring the instrument. Therefore, it is concluded that the scoring of exceptional quality points is a subjective, professional judgment made by each appraiser when observing a teacher. Since no significant reliability was found, the scoring of exceptional quality points cannot be supported as a reliable means of determining the quality of teaching in Texas schools. Generally, elementary certified appraisers awarded fewer exceptional quality points than secondary appraisers, males awarded slightly more points than females, appraisers indicated no noticeable trend because of years of administrative experience, and less experienced appraisers had the tendency to award more points than experienced appraisers.

Therefore, inter-rater reliability in awarding exceptional quality points cannot be expected on a consistent basis. Each appraiser, regardless of certification, sex, years of administrative experience or training, will use his or her own professional judgment when scoring the instrument.

TABLE OF CONTENTS

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION

House Bill 72, enacted by the 69th Texas Legislature, mandated that the State Board of Education (SBOE) develop a teacher appraisal process to evaluate classroom performance for career ladder purposes. The SBOE was charged with developing an instrument that was based on observable, job-related behaviors. It was also directed to provide (1) "at least two appraisers during each of the two appraisal periods within the regular school year, and (b) a uniform training program for appraisers of teacher performance including uniform appraiser certification standards" (Texas Education Agency, 1989, p. 1). The training program for appraisers consists of two levels of training. Any person who has the responsibility of evaluating teachers must complete an initial 40-hour training program. After successful completion of this program, appraisers are certified by the Texas Education Agency (TEA). The second phase of training consists of update training. Each certified appraiser must complete yearly two-day update sessions to maintain eligibility as an appraiser.

The Texas Education Agency conducted an extensive literature review of effective teaching techniques,

surveyed other states with a statewide evaluation system, and gathered data from 156 school districts regarding evaluation systems being used. After reviewing this information, the TEA sent a job-relatedness survey to 30,000 teachers, randomly selected by gender, race, teaching field, current assignment, and years of teaching experience. These teachers were to respond concerning the observability, importance, and frequency of use of the behaviors. From the 17,000 responses, the Texas Education Agency selected a list of teaching behaviors for inclusion in the appraisal instrument. This instrument, the Texas Teacher Appraisal Instrument (TTAI), is the one currently being used in all Texas schools.

In 1985, the SBOE implemented a pilot program across the state. The pilot program and a public hearing conducted by the SBOE resulted in revisions and changes in the appraisal instrument. In the fall of 1986, the Texas Teacher Appraisal System (TTAS) was implemented on a statewide basis. In January, 1987, some changes were made by the SBOE. These included the removal of zeros from the observation record form and a modification of the requirements for the professional growth plan. Also, guidelines were established for the rating of teacher performance for career ladder decisions. In February, 1987, further refinements were made. They included a

reduction in the number of indicators and criteria, development of the concept of whole instrument scoring, simplification of the Teacher Assessment of Instructional Goals and Outcomes, and the movement of exceptional quality scoring from the indicator level to the criterion level.

The Texas Teacher Appraisal Instrument (Appendix A), developed by the SBOE, comprises 49 standard expectation (SE) points and 27 exceptional quality (EQ) points. The Texas Teacher Appraiser's Manual includes very specific explanations for the SE indicators. The SBOE, however, did not develop a set of specific explanations for the EQ points. The SBOE did develop a set of guiding questions to help appraisers determine what to look for when awarding exceptional quality points, but this has been deleted from the 1988 training manual. Therefore, as from the beginning, EQ points are awarded based on the professional, subjective judgment of the appraiser.

Background research was done to ensure that the instructional content of the instrument was valid and reliable for evaluating teaching performance (Bloom, 1956; Brophy & Evertson, 1974; Gagne & Briggs, 1979; Rosenshine & Stevens, 1986; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). At the same time, however, very little research has been conducted on the standard expectation of exceptional quality points to determine whether or not the

Texas Teacher Appraisal Instrument results in an objective, reliable appraisal of classroom teaching. Sellitz, Wrightsman, and Cool (1976) state that the reliability of an instrument refers to

> the extent to which measures give consistent results. In other words, it refers to the degree of measurement error. Reliability is an important concern for two reasons. First, reliability is a precondition of the success of the instrument in measuring what it is supposed to measure, that is, it is a prerequisite of the validity of the instrument. Second, unless an instrument measures a variable relatively consistently, there is little hope of determining by means of that instrument whether changes in that variable are the result of other variables or are merely the reflection of the unreliability of the instrument. (p. 161)

This citation regarding reliability makes it clear why statistical research and evaluation of inter-rater reliability of the exceptional quality points on the TTAI is needed. An investigation of this nature will be useful to the Texas Education Agency, teachers, school boards, and all appraisers who are currently using the TTAI.

Medley, Coker, and Soar (1984) state that "since its inception, teacher evaluation has consisted of subjective

judgments of teachers' skills; the implicit assumptions have always been that the judges know what good teaching is and can recognize it when they see it" (p. 6). A state-mandated appraisal system with uniform appraiser training should assure those being appraised that their evaluators are able to know good teaching and are capable of judging it.

Teachers, appraisers, and the TEA should be able to feel comfortable about Texas' appraisal system. Since appraisers receive uniform training on the scoring of standard expectation, inter-rater reliability should be relatively high. Standard expectation is a "measure of performance which indicates that the teacher has demonstrated a specific teaching behavior at an appropriate level. The quality of the behavior is at least satisfactory" (Texas Education Agency, 1989, p. 117). These points are very structured and objective. Therefore, the instrument is a "low inference" device when appraisers score SE points. Each indicator on the instrument is clearly defined, and there are specific behaviors that the appraiser looks for and documents. Each score can be traced to the behavior from which it was determined. This, however, is not true of "high inference" scoring of EQ points.

Exceptional quality points are considered "high inference" judgments. High inference observations allow personal standards of effective teaching to enter into the rating. Since EQ points are "a measure of performance which indicates that a teacher consistently demonstrates a teaching behavior at appropriate times (quantity) and that the performance contributes to a consistently high level of student success (quality)" (Texas Education Agency, 1989, p. 115), the scoring made by Texas appraisers is a subjective, professional judgment. Therefore, the reliability of such measures needs to be demonstrated. Since there are no specific guidelines for awarding EQ points, it is reasonable for teachers to feel that these evaluations are subjective, unreliable, open to professional opinion, and perhaps based on irrelevancies.

Another reason to determine the reliability of EQ point measurement is the fact that the appraisal instrument is directly tied to career ladder. Career ladder placement is determined by college or advanced academic training hours, years of experience, and specific scores on the TTAI. Teachers must maintain certain levels of performance on the TTAI to stay at each level and to move up to the next one. Since the TEA has designated career ladder as a purpose of the Texas Teacher Appraisal System, the agency

should be sure to assure teachers that the system is indeed reliable.

Statement of the Problem

The problem of this study is to determine inter-rater reliability with respect to the exceptional quality points awarded on the Texas Teacher Appraisal Instrument which is part of the Texas Teacher Appraisal System.

Purposes of the Study

The purposes of this study are to determine

1. The degree of inter-rater reliability of appraisers who use the Texas Teacher Appraisal Instrument to award exceptional quality points.

2. The degree of inter-rater reliability among different demographic groups of appraisers: (a) elementary and secondary certified appraisers, (b) male and female appraisers, (c) appraisers with five years or less administrative experience and those with more than five years administrative experience, and (d) appraisers involved in 40-hour initial training or two-day update training.

Hypotheses

To carry out the purposes of this study, the following hypotheses were tested.

1. There will be no significant inter-rater reliability on each of the nine performance criteria by which exceptional quality points are measured on the Texas Teacher Appraisal Instrument.

2. There will be no significant difference between the inter-rater reliability of elementary certified appraisers who evaluate elementary teachers and secondary certified appraisers who evaluate elementary teachers.[1]

3. There will be no significant difference between the inter-rater reliability of male and female appraisers.

4. There will be no significant difference between the inter-rater reliability of appraisers with five years or less administrative experience and those with more than five years administrative experience.

5. There will be no significant difference between the inter-rater reliability of appraisers involved in initial 40-hour training and those involved in the two-day update training.

Background and Significance of the Study

The Texas Teacher Appraisal System is currently in its third year of existence. The observation/evaluation record

---

[1]Hypothesis 2 indicates that elementary and secondary subjects will observe an elementary teacher to determine reliability. Since there are no state-prepared EQ tapes of secondary teachers, it is impossible to test reliability of elementary and secondary subjects observing a secondary teacher.

known as the Texas Teacher Appraisal Instrument is one portion of the appraisal system. During its first year, the 1986-1987 school year, the instrument had three categories of performance: absent/below expectations, standard expectations, and exceptional quality. If the teacher did not demonstrate an indicator or did not adequately demonstrate a skill, the teacher received a score of zero for that indicator or skill. If the teacher demonstrated the skill at a standard level, he or she received a score of one. On certain indicators, teachers had the opportunity to score an extra point if they demonstrated the skill in an exceptional manner.

In May, 1987, TEA made several changes in the instrument. First, it removed the score of zero from the observation form. This was done to help with the problem of low morale associated with the score of zero. If the teacher did not demonstrate the skill at a standard level, the appraiser left the indicator blank. Second, the exceptional quality points moved from the indicator level to the criterion level. In 1986, teachers had the opportunity to receive exceptional quality points on specific indicators subsumed under a particular criterion. When exceptional quality moved in 1987 to the criterion level, a teacher had to demonstrate exceptional skill on all the indicators subsumed under a particular criterion in

order to receive EQ points. EQ points awarded at the indicator level had resulted in a score of one for each indicator. EQ points awarded at the criterion level resulted in a maximum of three points for the entire criterion. Third, in 1986-1987, it was possible to receive EQ points on certain behaviors in all five domains. Since 1987, EQ points are awarded with respect to the four performance domains only: (a) Instructional Strategies, (b) Classroom Management and Organization, (c) Presentation of Subject Matter, and (d) Learning Environment. The nine criteria by which points are awarded judge whether a teacher (a) provides opportunities for students to participate actively and successfully, (b) evaluates and provides feedback on student progress during instruction, (c) organizes materials and students (d) maximizes amount of time available for instruction, (e) manages student behavior, (f) teaches for cognitive, affective, and/or psychomotor learning, (g) uses effective communication skills, (h) uses strategies to motivate students for learning, and (i) maintains supportive environment. The fifth domain is designated to evaluate professional growth and responsibilities and is not scored as a result of a classroom observation. Therefore, if a teacher wants exceptional quality credit, he or she must demonstrate exceptional skill on all indicators subsumed under a

particular criterion in the four performance domains. According to the Texas Teacher Appraiser's Manual (Texas Education Agency, 1989), an appraiser must make "a 'holistic' judgment based on information about specific behaviors that provide the basis for the criterion level exceptional quality judgment" (p. 44).

Even though the process for scoring EQ points has changed, the decision to award EQ points is still a subjective, professional judgment. Even though the decision is primarily subjective, the appraiser may identify with some degree of objectivity certain observable student and teacher behaviors that may be considered exceptional.

The appraiser awards EQ points after first determining whether the teacher has demonstrated the behavior listed under a particular criterion at a standard level. If so, the appraiser determines if the teacher has gone above and beyond this standard. Since the instrument is the standard, the standard of the appraiser or the individual school district is not the issue. According to the Texas Teacher Appraiser's Manual (Texas Education Agency, 1989), "standard expectation is not average or mediocre, [sic] it is the standard of effectiveness" (p. 44). Therefore, a teacher who is meeting the standard expectation of the instrument is an effective teacher. Exceptional quality,

therefore, is teaching behavior that goes above and beyond the standard of effectiveness--it is exceptional.

One of the six assumptions underlying the TTAS is that "the difference between successful master and beginning teachers will appear in the number of skills exhibited by the teachers at levels of quality which meet or exceed stated expectations" (Texas Education Agency, 1989, p. 4). There are several critical factors for determining EQ points, but three have been identified as creating the potential for exceptional quality: quantity, quality, and preponderance. Quantity means that the teacher demonstrates a teaching behavior at appropriate times in the lesson. If this quantity contributes to a consistently high level of student success, then quality is present. An appraiser determines preponderance by watching to see whether the teacher demonstrates the skill more times than not. Preponderance is used to judge quality and quantity through weight, power, importance, or strength of the data. When the appraiser judges quantity, quality, and preponderance, the effect on students is the main focus. Regardless of the teacher's creative ability, there must be an observable, positive impact on students. In the Texas Teacher Appraiser's Manual (Texas Education Agency, 1989), "a teacher behavior which probably has no positive impact

on students cannot be considered to be of exceptional quality. Quality rests in probable effectiveness" (p. 45).

When the SBOE developed the TTAS, it did not clearly define statements of exceptional quality. Therefore, each appraiser who awarded these points had no set guidelines for determining whether to award credit or not. In May, 1987, when the SBOE changed the EQ scoring to the criterion level, it developed some "guiding questions" for determining whether a teacher was exceptional or not in a particular criterion. These questions were divided into student and teacher behaviors that might lead to the awarding of EQ points. Some of the questions are as follows:

> Did the teacher exhibit this behavior every time I thought it was appropriate? . . . Was the teacher maximally effective in that all or most of the students achieved the desired understanding/outcome/ behavior change? . . . Did most students participate in ways other than passive listening? . . . Did most students understand what they were expected to learn in each phase of the lesson? (Texas Education Agency, 1989, pp. 44-45)

The appraiser must understand that simply responding "yes" or "no" to these guiding questions does not mean that exceptional quality behavior has been exhibited. These

questions have since been eliminated from training so now there are no guidelines for awarding EQ points. Ultimately, the decision is left to the individual appraiser. For this reason, a study to determine inter-rater reliability in awarding EQ points is needed.

Since the TTAS has no guidelines for determining EQ points, a study of this nature should be very significant for teachers in Texas and other states that purport to have a valid, reliable evaluation system. Many concerns about observable behaviors, objectivity, and reliability have surfaced among Texas teachers. The TTAS states that the evaluation must be based on observable behaviors. Herbert and Attridge (1975) state that "observable refers to the degree to which those behaviors included in the instrument are capable of being perceived by any trained observer, either directly through his own senses, or with the aid of equipment designed to assist them" (p. 6). The major complaint of teachers is that appraisers are not trained to recognize the behaviors that are occurring in the classroom.

Appraiser objectivity is another major concern of teachers even though the standard expectation points on the TTAI are definable and objective. Herbert and Attridge (1975) state that "objectivity pertains to the extent to which the instrument lends itself to change by the observer

due to his own preference, expectations, needs, feelings, and biases and similarly to the extent it may cause the observed subjects to change" (p. 6). Teachers do not feel that an evaluation system that lends itself to observer preferences and biases is objective and reliable. According to Rosenshine and Furst (1973), the teachers point of view is well taken.

> Although an observational category system may provide neutral, objective descriptions of classroom transactions, the people who interpret the data usually make judgments about effective teaching. At present the judgments can be only guesses about what is good, true, and beautiful in classrooms--research in this area has barely begun. (p. 161)

Reliability, a third concern of teachers, is defined by Herbert and Attridge (1975).

> Reliability is not a property of an instrument itself, but a property of the measures derived from the instrument. It is clear that the reliability of any measure will depend on many factors other than the instrument used--for example, the skill of the observer(s), the nature and variability of the subjects of observation, and, particularly the number and length of the observation periods. (p. 14)

The significance of this study is obvious to teachers who are evaluated under the TTAS. This system has replaced all other teacher evaluation techniques used for determining career ladder placement in Texas. The TTAS system not only purports to evaluate effective classroom teaching but attaches a career ladder supplement to all teachers who qualify under the stipulations of House Bill 72 and the individual district. Reliability of what observers see when they are in the classroom is crucial to all involved with this system of evaluation. Therefore, this study attempts to determine the degree of inter-rater reliability in awarding EQ points on the TTAI.

The TTAI consists of 13 criteria, but only the first nine are performance-based and qualify for EQ. The 13 criteria are as follows: (a) *provides opportunities for students to participate actively and successfully, (b) *evaluates and provides feedback on student progress during instruction, (c) *organizes materials and students (d) *maximizes amount of time available for instruction, (e) *manages student behavior, (f) *teaches for cognitive, affective, and/or psychomotor learning, (g) *uses effective communication skills, (h) *uses strategies to motivate students for learning, (i) *maintains supportive environment, (j) plans for and engages in professional development, (k) interacts and communicates with parents,

(l) complies with policies, operating procedures, and requirements, and (m) promotes and evaluates student growth. The items marked with an asterisk are the nine performance criteria on which an appraiser may award EQ points.

If this study demonstrates that the TTAI inter-rater reliability is below 90%, the TEA needs to re-evaluate the reliability of the instrument when using it for career ladder purposes. If, on the other hand, this study demonstrates a higher inter-rater reliability, the TEA might reasonably promote the instrument as a reliable measure of teacher performance.

Definition of Terms

For the purposes of this study, the following definitions are provided.

Appraiser. Both of the individuals assigned to evaluate the performance of a teacher are appraisers. One appraiser is the teacher's supervisor, and the other is designated as the teacher's "other" appraiser. The teacher's supervisor must be designated as such and hold administrator or supervisor certification. The other appraiser must be approved by the local board of trustees, have a valid teaching certificate, and have at least two years of pre-kindergarten, kindergarten, elementary, or secondary classroom experience. The teacher's supervisor

and the other appraiser must have received uniform training and must be certified by the TEA as appraisers.

Criterion. A criterion is "one of the 13 subsets of performance indicators in the Texas Teacher Appraisal System" (Texas Education Agency, 1979, p. 115).

Documentation. Documentation is written verification of an occurrence(s) or condition(s) which affects the teacher's score on one of the indicators in denying credit for standard expectation and on the entire criterion in awarding exceptional quality points.

Exceptional Quality. Exceptional quality is a measure of performance which indicates that the teacher has demonstrated a teaching behavior that had an impact on students, occurred most or all of the time, and provided quality and quantity of instruction whenever possible. Three points are given for the criterion that is awarded exceptional quality points.

Instrument. The instrument is "a list of 65 specific teaching behaviors (indicators) categorized into 13 subsets called criteria. These criteria are grouped into five major areas called domains" (Texas Education Agency, 1979, p. 115).

Observation. Observation is "a visit to a classroom made by an appraiser with the intention of collecting data

with which to assess a teacher's performance" (Texas Education Agency, 1979, p. 115).

Performance Indicator. Performance indicator is "one of the 65 specific teaching behaviors which define the criteria on the TTAS instrument" (Texas Education Agency, 1979, p. 115).

Reliability. Reliability is the degree to which measures give consistent results.

Standard Expectation. Standard expectation is "a measure of performance which indicates that the teacher has demonstrated a specific teaching behavior at an appropriate level. The quality of the behavior must be at least satisfactory" (Texas Education Agency, 1979, p. 117).

Written Record. The written record section of the TTAS is the form which provides space for documentation of observations of the teacher by each appraiser.

Limitations of the Study

This study is limited in that all scoring on the TTAI by subjects is completed in the context of an appraiser training environment and is conducted using a videotaped lesson rather than a live classroom lesson. This limitation is natural for a quasi-experimental study and is necessary in order to ensure that all subjects view the same teaching performance under similar conditions.

Basic Assumptions

It is assumed that the subjects responded honestly to the survey used to gather demographic data and in scoring the EQ points from the videotaped lesson.

Summary

In Chapter 1, the writer has presented a description of the problem of reliability. The purposes and hypotheses relative to the problem have been specified. Data relative to the degree of reliability with respect to elementary and secondary certified appraisers evaluating an elementary teacher, male and female appraisers, years of administrative experience, and type of training are lacking. The problem, purposes, and hypotheses of this study are structured to elicit such study data.

References drawn from the available literature verify the need for the study. Terms are defined and limitations are stated as means of establishing the contextual framework for the study.

CHAPTER 2

SYNTHESIS OF RELATED LITERATURE

The evaluation of a teacher's performance has been identified with many titles: "teacher evaluation, teacher observation, administrator and teacher's progress reporting, merit rating, and most recently, performance appraisal" (Lewis, 1973, p. 23). Regardless of title, all fit one meaning: "the judgment by one or more educators, usually the immediate supervisor, of the manner in which another educator has been fulfilling his professional responsibilities to the school district over a specified period of time" (p. 23). The term "judgment," however, brings up the important question of reliability. No matter what definition is given to teaching, what evaluation instrument is used, or what criteria are being observed, reliability continues to be questioned by teachers and administrators. In the review of literature that follows, reliability and its many facets are addressed.

Borich (1977) states that throughout history, teacher education has lacked sufficient research methods and evaluation techniques to link specific teacher behaviors with precise student outcomes. Nevertheless, teachers have been evaluated with nonempirical methods and criteria.

Typically, the evaluator has used his or her own judgment to conclude which teachers are "effective" and which are not. Teacher evaluation methods that incorporate classroom observations of teaching behaviors should replace the judgmental approach.

The community made up of professional groups, legislators, teachers, administrators, and students has demanded better education so new approaches are imperative. These vocal groups have voiced a need in the past, and again recently, for schools to be evaluated. Therefore, thoughtful consideration has been given to why evaluations are necessary. At least four needs that give both purpose and form to the evaluation of teachers can be identified.

1. Today, parents have little contact with their child's teacher and have a need for renewed assurances about the quality of teaching and the welfare of their child (Phi Delta Kappan, 1979).

2. Administrators trying to improve the quality of education in the schools and faced with several autonomous units which in themselves defy coordination and development need detailed information about the teaching performance in the classroom (Madeus, Kellaghan, & Rakow, 1979).

3. The classroom teacher has a tendency to perceive himself or herself with some uncertainty and distortion,

and therefore needs reliable feedback from external sources (Hardebeck, 1973).

4. Many personnel decisions are made regarding teacher selection, contract renewal, promotions, certification, and in-service education. These decisions need to be based on objective information collected during observations of teaching performance (Cronbach, 1963).

The importance of teachers in the schools has rarely been questioned. Harris (1986) states,

> If teachers are important to learning, if resources are allocated primarily for their services, if schools cannot function without their work, then teacher evaluation is essential for understanding and improving the school's operation. Without teacher evaluation, all other efforts at educational evaluations are relatively nonproductive. (p. 3)

According to the Commission on Elementary Schools (1970) and the National Study of School Evaluation (1973), accreditation programs of many states and accrediting associations give attention to many facets of school and college operations but rarely focus with any rigor on faculty performance. It is obvious that evaluation of educational programs should go beyond teacher evaluation. Teacher evaluation, however, should be viewed as the most important project of any meaningful program evaluation

effort (Madeus et al., 1979). The teacher is not the only important concern of the educational process, but the teacher is central to the teaching/learning process (Rossmiller, 1983).

Contemporary urban societies have caused parents to lose touch with teachers as persons. Harris (1986) states that "the urban society provides for the rearing of its children and youth in a variety of ways--school, playground, church, scouts, little leagues, electronic games, television, street gangs, and movie houses all play a part" (p. 4). Family relationships tend to be lost, and parents turn to the school for help with child rearing. Teachers are no longer neighbors, friends, or members of the same church. Parents have given up trying to know their child's teacher. Harris also states that "desegregation and 'crosstown bussing' illustrate with dramatic overtones some of the parental concerns that grow out of the urbanization of schooling" (p. 5). Accountability seems to be a popular theme of society. Whether right or wrong, dissatisfaction, anxiety, or uncertainty about the educational programs is consciously or unconsciously directed at the classroom teacher by parents and other members of society (Gallup, 1979).

Teacher evaluation may be a tool to help restore public confidence in the teacher as well as in the schools.

In the absence of well-established systems that give public assurances, there has been a movement toward state-mandated teacher evaluation (Popham, 1971) and teacher competency testing (Cole, 1979).

Perhaps administrators and supervisors feel the greatest need for better teacher evaluation. Unfortunately, most of the information which administrators have regarding teacher performance is unreliable and irrelevant. According to Harris (1986),

> The lack of information for administrators and supervisors about teacher performance is perpetuated by several school traditions. The school board and community have long since given up trying to evaluate the teacher directly. Instead, there has emerged the teacher held attitude that "I know my business, so if you don't like how I teach, get another teacher," or "I'm a professional and I have the right to be left alone, to teach in my own way." (p. 6)

Even more widely accepted is the tradition that students are responsible for learning rather than the teacher or the school. This has its roots in rural America with the belief that teachers are good and students are privileged to attend school. It is obvious then, that if students fail, students are at fault. This tradition is currently reflected in standardized testing and in minimum

competency testing programs for students. These traditions have caused the public schools to focus very little on effective teacher evaluation.

Much has been written about the use of teacher evaluation to improve teaching, and this usually indicates that feedback needs to be given high priority. The autonomy and self-sufficiency of the classroom teacher present a difficult problem. Very rarely do evaluators offer useful feedback to the classroom teacher (Harris, 1986).

When little or no feedback is given to teachers, predictable consequences occur such as low morale, anxiety, and distorted perceptions of performance. These negative consequences from no feedback situations support the argument for teacher evaluation methods that are heavily focused on feedback to the teacher rather than judgments made by an evaluator and reported to a supervisor without the teacher's knowledge of results. To be useful, feedback to the teacher should be immediate and frequent and should stress objectivity.

The need to make personnel decisions based on evaluation results is usually cited as important. Many decisions can be made by using appropriate data. McIntyre (1979) cites "validating the teacher selection process" (p. 12) as one of the several purposes for teacher evaluation.

Other purposes include promotion, reassignment, and special recognition, as well as dismissal (Harris, 1979).

Dismissal decisions are probably the most worrisome for both teachers and administrators. Teachers fear the use of evaluation data when dismissal decisions are about to occur. The fear is often unrealistic, but much uncertainty is felt when teacher evaluation methods are subjective, unreliable, and perceived as producing unpredictable decisions. There is substantial reason to believe that rarely is there a connection between teacher evaluation and dismissal decisions. Finlayson (1979) discovered that as far as incompetence is concerned, "only eleven . . . teacher dismissal cases due to competence [were] appealed to the secretary of education" (p. 69) in Pennsylvania from 1971-1976. Obviously, there are more "incompetence" dismissals than those reported by Finlayson (1979), who writes, "there are a few informal, out-of-court purges in some districts. These cases, of course, involve young, nontenured teachers" (p. 69). More common than dismissals, however, are numerous instances of incompetence that will prove more disruptive to quality education and are not dealt with in any systematic way (American Association of School Administrators, 1979).

A well designed evaluation system is a major communication link between administrators and teachers. At

one end, it describes concepts of teaching to teachers and explains standards for their work. At the other end, it helps administrators to manage, structure, and reward the work of teachers.

Renewed national interest was given to teacher evaluation in April, 1983, when A Nation at Risk: The Imperative for Educational Reform was published by the National Commission on Excellence in Education. Several of the recommendations suggested by the commission require teacher evaluation.

> Persons preparing to teach should be required to meet high educational standards, to demonstrate an aptitude for teaching, and to demonstrate competence in an academic discipline. . . . Salaries for the teaching profession should be increased and should be professionally competitive, market sensitive, and performance-based. 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. (p. 30)

Action for Excellence, the June, 1983, report of the Task Force on Education for Economic Growth, Education Commission of the States, had many of the same

recommendations as the National Commission on Excellence in Education.

> We recommend that the boards of education and higher education in each state--in cooperation with teachers and school administrators--put in force, as soon as possible, systems for fairly and objectively measuring the effectiveness of teachers and rewarding outstanding performance.
>
> We strongly recommend that the states examine and tighten their procedures for selecting not only those who come into teaching, but also those who ultimately stay. . . . ineffective teachers--those who fall short repeatedly in fair and objective evaluations--should, in due course and with due process, be dismissed. (p. 39)

Better teachers and better teaching have increasingly become viewed as the key to better education. The Commission on Excellence, in seeking ways to improve the quality of education, recommends improving the quality of teachers. Never has the premise that education could be improved without improving the quality of teachers held true.

"The new concerns for the quality of education and of teaching are being translated into merit pay, career ladder, and master teacher policies that presuppose the

existence of effective teacher evaluation systems" (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984, p. 12). It has become increasingly important for school districts to understand the educational systems they are using because evaluation systems can determine the nature of the teacher and overall education in their schools.

Before any state agency or school district adopts an evaluation system to determine tenure, promotion, merit pay, or master teacher status, educators will need to answer such questions as

1. Is there any evaluation system that can reward outstanding teachers, encourage average ones, and improve or terminate unsatisfactory ones?

2. Can teacher aptitude be recognized on a written test or must prospective teachers be evaluated while teaching?

3. What are the problems connected with linking salary, promotion, and retention decisions to teacher evaluation?

4. Can teacher evaluation itself be used to determine master teacher rank? (Wise et al., 1984).

Since teacher evaluation is being used for merit pay, career ladder, and master teacher decisions, the degree of reliability in an evaluation process is important for school districts to examine. Wise et al. (1984) state,

> Reliability in evaluation refers to the consistency of measurement across evaluators and observations. The degree of reliability required of a teacher evaluation system depends on the use to be made of the results. Personnel decisions demand the highest reliability of evaluation results. Evaluation criteria must be standardized and evaluators must apply these criteria with consistency when the results are to be used for personnel decisions regarding tenure, dismissal, pay, and promotion. (p. 44)

Even for these purposes, it is important to remember that reliability cannot be disregarded because it affects teacher morale and the understood legitimacy of the evaluation process. Reliability may be replaced by variability if the ultimate goal is to encourage individual development based on individual need.

Deneen (1971) approaches reliability by asking three questions: (a) "Do the measures of teacher performance require observable behaviors?" (b) "Have these behaviors been weighted for their importance and scaled on some stable reference?" and (c) "Have those using the measures been trained in observing and valuing teacher performance?" (p. 174). If the prospective school district cannot answer "yes" to all three questions, the reliability of the system is questionable.

The process approved by a district for conducting teacher evaluation affects the reliability and, therefore, the legal value of evaluations. Districts can improve the legal merits of its evaluation process by increasing the reliability of the evaluation measures. One way to achieve this is to use multiple ratings of teachers whose incompetence is in question. More than one skilled evaluator should conduct these assessments. Also, the district should provide all evaluators with training in the use of the instrument and procedures. It would also be beneficial if the evaluators could discuss evaluations with other appraisers to gain inter-rater reliability.

McDonald (1976) contends that teacher behavior is another aspect of reliability of the system. "Teachers do not teach the same way from day to day for many sound reasons. The data will differ by the subject being taught, by the day of the week, by the time of day, and by the week or month" (p. 105). To alleviate this problem, appraisers should observe a teacher as many times as possible. Teaching is a very complex phenomenon, and it is impossible to see total effectiveness unless all aspects are observed.

Validity is another important concern when deciding on an evaluation process. Wise et al. (1984) state,

> Validity of a teacher evaluation process depends on its accuracy and comprehensiveness in assessing

> teaching quality as defined by the agreed-on criteria. Although school districts may seek to finesse the issue of validity by striving for measurement reliability in their evaluation process, they cannot ignore the validity of the process when they use its results as a basis for personnel decisions. (p. 49)

The criteria, the procedure for obtaining data, and the ability of the evaluator contribute to the validity of an evaluation process. In short, the process of evaluation must suit the purpose if the results are to be judged valid and reliable.

Different conceptions of teaching and school organization underlie the type of evaluation process needed to achieve the intended purposes of the district. If teacher evaluation is to work, it must satisfy individual and organizational needs. It must balance the standardization needed for personnel decisions and the responsiveness needed for teacher growth and improvement.

A teacher evaluation process must determine the boundaries of the teaching task and provide a detailed system for judging the teacher. Labor, art, profession, and craft are four ways to look at teaching. These four ways of viewing teaching provide a theoretical basis for analyzing the teacher evaluation process.

When defining teaching as "labor," teaching activities are "rationally planned, programmatically organized, and routinized in the form of standard operating procedures" by administrators (Mitchell & Kerchner, 1983, p. 125). With this definition, teachers must adhere to the prescribed manner and specified routines and procedures.

Evaluating teaching as labor involves viewing lesson plans and classroom performance. In this case, the school administrator is the teacher's supervisor. When one evaluates a teacher using this theoretical framework, one assumes that effective practices can be concretely determined by the evaluation and, in doing these practices, the outcome will be the desired result.

Under the heading of teaching as a "craft," teaching requires a repertoire of specific procedures. Knowledge of these procedures includes knowledge of general rules for application. After the teacher receives an assignment, he or she is expected to work with little supervision or instruction. When teaching is considered a craft, the evaluation is indirect, and the evaluator assumes the teacher has the requisite skills. The evaluator merely holds the teacher to general performance standards and assumes that proper use of the desired techniques and rules will produce the desired outcome.

Defining teaching as a "profession" requires that the teacher have not only a repertoire of specified techniques but also the ability to exercise judgment about when and how he or she should apply those techniques (Shavelson & Stern, 1981). Such professional judgment can be sound only if the teacher has a body of theoretical knowledge as well as a pharmacy of techniques. Broudy (1956) distinguishes between craft and profession in this way.

> We ask the professional to diagnose difficulties, to appraise solutions, and to choose among them. We ask him to take total responsibility for both strategy and tactics. . . . From the craftsman, by contrast, we expect a standard diagnosis, correct performance of procedures, and nothing else. (p. 182)

If a school district chooses an evaluation system that adheres to the conception of teaching as a profession, the administrator would merely ensure that teachers have the resources necessary to carry out their jobs. This view assumes that standards of knowledge can be developed and assessed and their enforcement by the administrator and the teacher will ensure competent teaching.

Teaching techniques that are novel, unconventional, or unpredictable are often referred to as part of teaching. One must not conclude that standards of practice are

ignored, but rather their form is personalized and not standardized.

As Gage (1978) explains, teaching as "art" involves "a process that calls for intuition, creativity, improvisation, and expressiveness--a process that leaves room for departures from what is implied by rules, formulas, and algorithms" (p. 15). He argues that "teaching uses science, but is not itself a science because the teaching environment is not predictable" (p. 15). Teachers who perceive teaching as an art must draw on professional knowledge and techniques as well as their own personal pharmacy of ideas that expresses their personality when interacting with students.

Because teaching viewed as art involves personal autonomy in performance, evaluation should include both self-assessment and critical assessment by others. Such evaluation entails "the study of holistic qualities rather than externally objective points of view" (Gage, 1978, p. 15). It relies on judgmental ("high-inference") rather than countable ("low-inference") variables on assessment of patterns of events rather than counts of specific discrete behaviors (Eisner, 1978; Gage, 1978).

It is obvious that these four definitions of teaching do not exist in pure form in individual classrooms. Teaching is not a clear cut definitional act. Nonetheless,

these views of teaching mean different definitions of success in a teacher evaluation system.

Wise et al. (1986) state that the "disparity implicit in views of teacher evaluation cannot be ignored" (p. 8). McNeil and Popham (1973), for example, prefer to see teaching evaluated by its contribution to student performance measured by test scores rather than by teacher performance criteria. Millman (1981) argues that "criteria and techniques for the fair use of student achievement in both formative and summative roles of teacher evaluation can be devised" (p. 159). He also states that

> students learning as measured by their test performance is a direct function of teacher performance and it measures a teacher's worth in terms of the product or output of his work. Thus, test performance of students envisions teaching as labor and the student as raw material. (p. 159)

In a poll of teachers conducted by the National Education Association (1979), 89% did not consider student scores on standardized tests as a valid measure of their effectiveness. The views of this large number of teachers are based on two points: (a) First, standardized test scores are limited measures of student ability, and (b) other determinants of the teaching and learning process are just as important in determining success in teacher

performance. Other factors such as school and home conditions which are not under the teacher's control are inherent elements that give rise to the idea of teaching as profession or art.

Wise et al. (1984) state,

> Although the various conceptions of teaching differ along several dimensions, one can usually view them as incorporating increasing ambiguity or complexity with regard to the performance of teaching tasks as one moves from labor at one extreme, to art at the other. The role of the teaching environment in determining teacher behavior also increases in importance as one moves from labor to art. The more variable or unpredictable one considers the teaching environment, the more one is impelled to conceive teaching as a profession or art. (p. 9)

Gage (1978) describes how the elements of predictability and environmental control differentiate teaching as a science from teaching as an art. Teaching as a science, he observes, "implies that good teaching will someday be attainable by closely following rigorous laws that yield high predictability and control" (p. 17).

Wise et al. (1984) suggest that "what teachers do in the classroom does affect students" (p. 10). However, assertions that a set of behaviors consistently lead to

increased student performance have been countered by consistent and often contradictory findings that undermine faith in the outcomes of simple process-product research (Doyle, 1978; Dunkin & Biddle, 1974; Shavelson & Dempsey-Atwood, 1976).

In a review of literature on school and teacher effects on student learning, Centra and Potter (1980) observed that "student achievement is affected by a considerable number of variables of which teacher behavior is but one" (p. 187). They conclude,

> Teacher effects are likely to be small when compared with the totality of the effects of the other variables affecting student achievement, [and] . . . the effects of any one of the variables . . . are likely to be small when compared with the combined interactive effect of all other variables. (p. 287)

This interactionist view of teaching is neatly capsulized by Brophy and Evertson (1976) who state,

> Effective teaching requires the ability to implement a very large number of diagnostic, instructional, managerial, and therapeutic skills, tailoring behavior in specific contexts and situations to the specific needs of the moment. Effective teachers not only must be able to recognize which of the many things they know how to do applies at a given moment, but be able

to follow through by performing the behavior correctly. (p. 139)

Research on the stability and generalizability of measures of teacher behaviors lends support to a context-specific view of teaching. Stability refers to the extent that a teacher's behavior if measured at one point will be the same when measured at another time. Generalizability refers to the extent that such measures are the same for different teaching situations. Teachers do not exhibit the same kinds of behavior at different points in time or between different content areas. This may be due to the fact that teachers must need to adjust their behavior to the needs of their classroom (Shavelson & Dempsey-Atwood, 1976).

Effective teaching behaviors vary for students of different socioeconomic, mental, and psychological characteristics (Brophy & Evertson, 1974; Cronbach & Snow, 1977; Peterson,1976) and for different grade levels and subject areas (Gage, 1978; McDonald & Elias, 1976). Peterson and Kauchak (1982) and Soar (1972) state that there are certain teaching behaviors that have proved effective when used in moderation but when overused have produced negative results. This kind of research discourages the development of rules for teaching that can be applied to general situations.

Another finding is that the objective of instruction might cause a difference in teaching behaviors. Then there should be different teaching strategies for test-taking, problem solving, and cognitive learning. This is a problem because, in reality, different teaching strategies and techniques can all lead to positive results. Therefore, it would be difficult to state that there is only one way to become an effective teacher. It can be assumed that different strategies might be delineated for specific goals to be achieved. No one set of criteria would be appropriate for judging every teaching situation.

Several recent reviews (Ellett, Capie, & Johnson, Haefele, 1980; Lewis, 1982; Millman, 1891; Peterson & Kauchak, 1982) of teacher evaluation processes have identified six approaches that are used most often. These authors contend that the approaches used to evaluate teachers measure very different aspects of teaching and the teacher. The approaches rely on different definitions of what demonstrates adequacy and how to recognize or measure adequacy. Some of the evaluation processes purport to assess the quality of the teacher, and some processes assess the quality of teaching. Other evaluation processes seek to assess the quality of student performance or teacher effectiveness.

One tool for assessing school personnel is teacher interviews. Haefele (1981) identified two uses for these interviews: (a) for the purpose of hiring decision and (b) for communication in performance appraisals to the teacher. However, there is no empirical research regarding the ability of interviews to predict the effectiveness of a teacher.

Competency testing for initial certification and hiring is another process for evaluating personnel. This process is based partly on the premise that teachers should demonstrate cognitive ability as a prerequisite for a teaching position and partly on the public's doubt about the effectiveness of teacher education and training (Quirk, Witten, & Weinberg, 1973). Standardized teacher tests guarantee a minimum standardized knowledge on the part of the prospective teacher, but tests cannot assess the classroom performance of a teacher (Haefele, 1980; Harris, 1981). Further, past studies (Coleman, Campbell, Hobson, McPartland, Weinfeld, & York, 1966; Guthrie, 1970) indicate that higher knowledge levels are not clearly translated into more effective teaching.

Classroom observation, coupled with teacher interviews and conferences, is the mainstay of most current teacher evaluation systems. It involves direct observation of the teacher in the classroom by a trained evaluator. Classroom

observation reveals "a view of the climate, rapport, interaction, and functioning of the classroom available from no other source" (Evertson & Holley, 1981, p. 90). Classroom observations may vary according to school district or state policies. Even though the principal usually acts as the observer, trained evaluators and other teachers may observe teachers. A pre-observation conference usually precedes these observations (Garawski, 1980; Redfern, 1980). The district or the state determines the frequency and length of the observation.

The advantage of this method is the ability to see teachers in action in the classroom. However, there are some limitations. Observer bias, insufficient sampling of performance, and poor measurement instruments can threaten the reliability of the results (Evertson & Holley, 1981; Haefele, 1980; Lewis, 1982; Peterson & Kauchak, 1982). Performance ratings have also shown limited stability and generalizability, particularly when low inference measures of specific teaching behaviors are used (Shavelson & Dempsey-Atwood, 1976). Another common limitation of accuracy is evidenced when the teacher puts on a "dog and pony show" for the evaluator. This does not accurately reflect what goes on in the classroom on a regular basis.

Student ratings are another form of classroom observation" since they measure observed performance from

the student's point of view. This method is inexpensive with a high degree of reliability (Peterson & Kauchak, 1982), but questions about its validity restrict its use as the primary evaluation tool (Aleamoni, 1981; Haefele, 1980). According to McNeil and Popham (1973),

> Considerable halo effect is found when students rate their teachers on several traits. As expressions of feeling, student ratings unquestionably have validity. They can be useful indicators that learners have or do not have favorable predisposition to the teacher and the course. (p. 233)

In an especially well-designed study, Davidoff (1970) provided strong evidence leading to the conclusion that student opinion of teacher behavior is very stable over time and that there is no consistent relationship between student opinion of teacher behavior and student gain.

The process of peer reviews is another form of teacher evaluation. This process involves peer evaluators who evaluate teaching through the viewing of lesson plans, graded materials, and classroom observations. The assumption is that the best evaluators of teachers are other teachers. Someone who knows the curriculum, grade requirements, and pupil demands can render more specific and practical suggestions for improvement. According to

Haefele (1980), this method is not generally used as a basis for personnel decisions.

A very controversial evaluation process is that of student achievement. In education, the ultimate concern is student achievement, and to some educators, this is the only true indicator of teacher effectiveness. Even though student scores represent legitimate indicators of success, numerous assumptions must be made if these are to be linked to teacher performance in the classroom.

In studies by Brophy (1973), Rosenshine (1970), Shavelson and Russo (1977), Veldman and Brophy (1974), the reliability of student evaluations as a measure of teacher performance indicates that reliability is very low. This means that the teacher performs differently in different teaching situations so caution should be taken when using this information to rate teacher competence. Further, the use of tests as a measure of teacher effectiveness may inhibit creativity and lead teachers to teach to the test (Shine & Goldman, 1980) and may counteract the effects of teacher behaviors on other desirable outcomes (Centra & Potter, 1980; Peterson, 1979).

Teacher self-evaluation is another evaluation process used in many districts. Self-evaluation is a fairly new method of teacher evaluation. When a teacher combines self-evaluation and individual goal setting, he or she is

more likely to be motivated for change and growth. If used properly, "objective" data may help the teacher evaluate strengths and weaknesses for both personal and professional needs. This process is not acceptable for accountability but is suitable for individual or group development and improvement. Both Redfern (1980) and Lewis (1982) agree that self-evaluation should be regarded not as a process for evaluation in itself but as an important source of information and motivation in the complete evaluation program.

The success of an evaluation system depends on its purpose for the district or state and the ability of the process to measure what it purports to measure. Some evaluation processes measure competence, some measure performance seen through direct observation, and some rely on student performance. According to Darling-Hammond, Wise, and Pease (1983),

> The generally low level of reliability, generalizability, and validity attributed to teacher evaluation methods suggest that unidimensional approaches for assessing competence, performance, or effectiveness are unlikely to capture enough information about teaching attributes to completely satisfy any of the purposes for teacher evaluation. (p. 308)

To this point in the chapter, the need for as well as the importance of teacher evaluation systems have been reviewed. Also, the many different processes available have been discussed. Next, it may be beneficial to examine some case studies of reportedly effective evaluation systems.

In 1983, the Rand Corporation conducted a study of teacher evaluation practices. In the study, the researchers examined instruments and procedures as well as the implementation processes and the organizational contexts of these systems. The authors of this study firmly believe that the evaluation system which a district chooses can "either reinforce the idea of teaching as a profession, or it can further de-professionalize teaching, making it less able to attract and retain talented teachers" (Wise et al., 1984, p. v).

The initial Rand study was designed to produce information for school districts to use in helping teachers improve and in making personnel decisions. The study began with 32 districts that were noted for having highly developed teacher evaluation systems. The differences in these systems varied with respect to the instruments, the number of evaluations, the roles of the teacher and administrator, and how the collected data were used. These differences led the researchers to believe that teacher

evaluation was an "underconceptualized and underdeveloped activity." The following discussion will emphasize the major problems and successes in these exemplary evaluation systems. The opinions stated were gathered from responses to a questionnaire sent to teachers in these selected districts. Even though there were many differences in development and choice of evaluation systems, several of the same teacher concerns appeared.

One of the major areas of concern voiced by these teacher respondents was that even though principals support evaluation systems "principals lack sufficient resolve and competence to evaluate accurately" (Wise et al., 1984, p. 22). This problem might stem from the fact that principals have not been able to resolve the conflict between their roles as instructional leader and teacher appraiser. Many principals have had a reputation of being "good guys" and their evaluations have been thought to be upwardly biased. Wise et al. (1984) state that "principals' disinclination to be tough makes the early identification of problem teachers difficult and masks important variations in teacher performance" (p. 22). In addition, many principals view teacher evaluation as a chore and an extra responsibility that has been added to their many other duties. Therefore, their full support and dedication has been absent.

The second most frequently mentioned concern was teacher resistance or apathy toward evaluation systems. The evaluation itself causes anxiety, and full support of these systems is not present with most teachers. A great amount of discomfort among teachers stems from a third problem area, i.e., lack of uniformity and consistency within a school system. Wise et al. (1984) state,

> Even though evaluation instruments have become more standardized, in many districts teachers believe that the present system depends too much on the judgment or predisposition of the principal and leads to different ratings for similar teacher practices in different schools. (p. 22)

Much of the inconsistency in teacher evaluation stems from the instrument being used, but a larger inconsistency is reflected in the inadequate training of evaluators. There are strong feelings that those responsible for the evaluation of teachers do not receive adequate training and that the training that is being given provides insufficient competence in the process of evaluation.

Another area of teacher concern rests with the evaluation of secondary school staff and secondary content specialists. Both of these issues involve the "difficulty of a generalist evaluator," i.e., the principal "assessing the competence of a specialist teacher" such as a secondary

level chemistry, history, or English specialist (Wise et al., 1984, p. 23).

The problems involved in valid evaluation of teaching are many and are largely unresolved. Even though there are many problems, there are also certain positive features. Wise et al. (1984) state that "teacher evaluation is one of the most powerful ways to impact instruction" (p. 23). Since evaluation, according to Wise et al., is so powerful, it would be important to see how it impacts student instruction and success. Two positive results of teacher evaluation were consistently reported: (a) improved communication between teacher and administrator and (b) heightened teacher awareness of instructional objectives and classroom procedures. Teacher-principal relationships were strengthened by the evaluation process, i.e., pre-observation meetings, classroom visitations, and post-observation conferences between teacher and evaluator. These meetings made teachers more aware of the goals and objectives of classroom instruction.

A formal evaluation system may create a communication network that has never been there before. One teacher cited that "teacher evaluation has brought about a sense of team effort at the building level that did not exist before. More teachers and principals are beginning to jointly establish common goals" (Wise et al., 1984, p. 23).

Many respondents reported that a set evaluation system helps them increase their pride and professionalism and is a motivating factor for improved instruction in the classroom. As one superintendent stated, "Our teacher evaluation program has made teachers prouder of their school system. They are proud of their role in ensuring academic standards in our schools" (Wise et al., 1984, p. 23). It must also be stated that teachers need to be rewarded for their competence if the pride and professionalism is to continue. New evaluation systems have done much to keep the classroom teacher from being isolated from administration. The new growth of two-way communication has led to common goals that can be stressed by both administration and the teachers.

From the 32 districts that were surveyed, four effective districts were chosen. They were selected to represent the diversity of evaluation systems in progress. These four districts were Salt Lake City, Utah; Lake Washington, Washington; Greenwich, Connecticut; and Toledo, Ohio. About a week was spent in each district by a 13-member panel financed by the National Institute of Education. These panel members interviewed the superintendent, other top administrators, members of local teachers' organizations, school board members, parents, and community leaders. In each district, six schools were

visited, and principals, special personnel, and teachers were interviewed.

Each district tackled the process of evaluation differently. The differences were in who evaluated the teachers, the purposes of the evaluation, the instrument used, the judgment process, and how the process would be connected with staff development or other personnel decisions.

The factors that made these systems successful might also contribute to the success of others. Specifically,

> These districts provide top-level leadership and institutional resources for the evaluation process, ensure that evaluators have the necessary expertise to perform their task, encourage teachers and administrators to collaborate to develop a common understanding of evaluation goals and processes, and use an evaluation process and support system that is compatible with each other and with the district's overall goals and organizational context. (Wise et al., 1984, p. 26)

Attention to these four factors--organizational "commitment," evaluator "competence," "collaboration," and strategic "compatibility"--has elevated these four districts from a meaningless procedure to a meaningful process.

Commitment to the evaluation system chosen by a district is of utmost importance. It should be recognized that a key component of a successful evaluation process is time; i.e., time allocated for pre-conferencing, observation, and post-conferencing. Successful districts make time for their evaluation process.

Evaluator competence is a very difficult issue. There are usually two components in the ability to make accurate judgments about the quality of instruction and the ability to recommend appropriate suggestions to help improve the teacher's performance. Successful evaluation systems have built-in mechanisms for checking the accuracy of evaluators' judgments. In these four districts, the evaluators are forced to justify their decision in precise, concrete terms.

These four districts found that collaboration is a major concern if an evaluation system is to be successful. Teachers and administrators must communicate. If communication occurs consistently, major implementation problems can be resolved before they become too complex.

An important note to stress about each of the four case districts is that "teacher evaluation supports and is supported by other key operating functions in the schools. Evaluation is not just an ancillary activity; it is part of

a larger strategy for school improvement" (Wise et al., 1984, p. viii).

There are three main reasons why the case study districts succeeded. First, the districts implemented the evaluation systems as planned. Second, the participants involved in the system understood the process. Third, the results of the evaluation were actually put to some use.

In Lake Washington School District No. 414, Greenwich Public Schools, and Salt Lake City Public Schools, each teacher was appraised every year by an administrator. This requirement decreased evaluation reliability by increasing the chances of variability among evaluators and variability across evaluations and observations. However, thorough evaluator training helped to minimize unreliability to a small extent.

In order to evaluate minimum competency, i.e., standardized, generalizable, and uniformly applied criteria, the evaluator must be able to observe the teacher exhibiting specific generic skills. If appropriateness of the teaching is to be evaluated, the appraiser must know something about the subject matter, classroom make-up, and the characteristics of the teacher being evaluated. The appraiser's level of expertise in the subject matter must be on the same level or above that of the teacher being evaluated.

In Salt Lake City, Lake Washington, and Toledo public schools, the absence of minimal teaching skills automatically triggered help. In these schools, principals generally spent little time evaluating teachers who seemed competent. This practice was not appreciated by most of the teachers. They did not feel they were receiving needed constructive criticism from their principals.

All four of the evaluation systems required that evaluators carefully document any behaviors that were considered unsatisfactory. The Salt Lake City, Toledo, and Lake Washington school districts' evaluation systems required multiple observations, and resources were provided for this complete process.

Toledo and Lake Washington public schools have taken aggressive measures to ensure evaluator competence. The Toledo evaluation system selected appraisers who were consulting teachers and were recognized by peers as having been the best in their field. The appraisers were matched with the teachers in their teaching field or area of expertise. Lake Washington trained all evaluators in the same principles that teachers learn in staff development. This type of prerequisite teaching helped to correlate the district's goals with the appraiser's judgment of the teacher.

Along with evaluator competence and minimum competence of teaching skills, the four case districts found utility to be an important issue. Utility depends on the reliability and validity of the evaluation instrument and how consistently and precisely the process measures minimal competence and degree of skill. Utility of the evaluation process depends also on the cost benefit, that is, on whether it produces usable outcomes without creating excessive costs. The outcomes must be worth the time and money used to acquire them if the process is to outlive competing organizational demands. The utility should be balanced between costs and benefits. The benefits include collection of data to make appropriate decisions, to improve communication between teacher and principal, and to improve personnel decision making.

Toledo's evaluation system had high utility. It helped teachers obtain acceptable teaching competence, or remediation was given, and incompetent teachers were removed from the classroom. The system was able to achieve both without disrupting the total operation or lower teacher morale. Wise et al. (1984) state that three critical factors ensured the utility of the Toledo evaluation process.

1. It was carefully managed, and it was conducted by evaluators who have no other competing responsibilities.
2. It was focused and it used limited resources to reach a carefully defined subset of teachers.
3. It was a collaborative effort and it engaged the key political actors in the design, implementation, and ongoing redesign of the process. (p. 58)

Salt Lake City's evaluation process also had high utility for accountability purposes. This process identified, assisted, and, if needed, removed incompetent teachers from the classroom.

Greenwich Public Schools enabled each school to engage the individual teacher in such a way that it related to professional endeavors. The utility of this system allowed it to motivate teachers and reward teachers' efforts by recognizing their importance.

The utility of Lake Washington's system was considered fairly high. The financial and logistic costs of this system were greater than any of the other districts. The district felt that this great expenditure of money was producing visible benefits to the teachers. Lake Washington did feel that the very specified, time-consuming, and detailed procedures for evaluation decreased

the utility in two ways: (1) The detailed procedures discouraged teacher probation in many cases, and (b) it left little time to be spent on competent teachers.

The four districts achieved utility in different ways. These districts also reported different methods regarding evaluation reliability. The definition of reliability in discussions on teacher evaluation refers to the consistency of measurements across evaluators and observations. According to Mazur and Peterson (1978), there are three major questions to be asked in the testing of an evaluation system for reliability.

1. How consistently does the instrument measure whatever it claims to measure?
2. To what degree are observations or scores consistent over time?
3. To what degree are observations or scores consistent across different raters? (p. 121)

They also state that "if considerable variation exists from one time to another or from one rater to another, the reliability of the procedure is low" (p. 121).

In varying degrees, the four case districts attempted to ensure reliability in their evaluation processes. Of the four, Toledo took the most extensive approach to ensure reliability. The school districts of Lake Washington, Greenwich, and Salt Lake City had a more difficult task

because the primary evaluators were building principals. These administrators were responsible for every teacher, and this increased the number of evaluators, teachers, and observations that had to be standardized.

According to Wise et al. (1984), at least three sources of variability may make teacher evaluation unreliable: (1) variability in the interpretation of appraisers; (b) variability of a single appraiser, i.e., whether the appraiser uses the same criteria consistently from teacher to teacher; and (c) variability in observations, i.e., whether the appraiser uses the same criteria when observing the same teacher in different teaching situations.

These potential sources of unreliability are not a problem for the Toledo school system because a small number of evaluators are utilized. This helps in the attempt to increase system-wide reliability. These evaluators employ a standard of teaching that is the same from school to school. Frequent classroom visits also help enhance the reliability of the process. Observations were made at least twice a year instead of a single observation once a year. Intensive consultation was conducted to incorporate goal setting so the teacher and evaluator would have a common understanding of what was being evaluated and observed.

Lake Washington school system required administrators to evaluate every teacher every year. This type of evaluation system decreases reliability because of variability among evaluators and variability across formal observations and classroom visits. According to Medley (1982), there is a distinction between the reliability of a score based on one or more observational records and the amount of agreement between records of the same behavior made by different observers. Lake Washington's evaluation system was not reliable since one evaluator observed each teacher only one time. There was no way to judge observer agreement (reliability) because different observers did not observe the same behaviors. This system might have had "observer validity" if records showed that the frequency of the evaluator's record agreed with the actual occurrences of the items or categories or behavior recorded.

Lake Washington's evaluation system had an "evaluation checklist" of 29 behaviors that were categorized under seven criteria. A checklist helps evaluators focus their attention on specific behaviors in every classroom. However, the requirement that every teacher be seen every year cuts down on quality time an administrator can spend with each teacher individually. The teacher who needs help is given much more attention than the teacher who is considered competent. As one principal put it,

> I have to evaluate too many people. Four or five people are taking all my attention and I am just doing lip service for the rest. There is no way to fit all of this in within the present system and state constraints. So I just go through the motions with half of them. (Wise et al., 1984, p. 46)

Superficial evaluations reduce the reliability of the evaluators' judgments. There are some teachers who need to be placed on probation, but, because of time, they are left in the classroom teaching students. Because of time constraints put on administrators, reliability across evaluations suffers.

When Greenwich decided to use teacher evaluation for reduction in force decisions, they became very concerned with reliability. Originally their evaluation process was used to evaluate a teacher and help him develop according to specified needs. This required reliability across observations for that one teacher but not for other teachers or for the evaluators as a team. Also, the Greenwich evaluation process lacked reliability because there was no checklist of items to be observed. Evaluators wrote comments stating what they observed during the lesson, but there was no comparability as to what was evaluated.

This process guarantees low reliability because the criteria vary from teacher to teacher and evaluator to evaluator. Low reliability is not a major concern if the evaluations are being used for individual staff development. Greenwich realized the need for more reliability in their evaluation process and is taking steps to upgrade the system and make it more accountable for the sake of reliability, i.e., having supervisors read and evaluate all observations for clarity of description, training evaluators to discuss evaluations to improve observation and reporting techniques so that results are more generalizable across evaluations, and classifying teachers as marginal or outstanding and then assessing data to see if it is adequate for these placements (Wise et al., 1984).

Salt Lake City's evaluation process had low reliability because it lacked a formal observation instrument and a checklist of specific behaviors. Basically, the process dealt with goals set by the district or an individual teacher, but these goals or criteria were not standardized uniformly across teachers. Reliability cannot be high because there was not comparability among teachers. Unlike Greenwich, Salt Lake City evaluators did not receive ongoing training to help increase evaluator reliability.

Salt Lake City's system attempted to increase reliability by having a team composed of specialists that put together remediation plans for teacher. This team was utilized to bring consistency to the process because they served on many remediation teams. This was an attempt to increase reliability across evaluations. However, two members of this team were drawn from a pool which meant less consistency was given from these two participants. Therefore, this inconsistency counteracted the attempt at reliability from using the team of specialists. Salt Lake City felt that this team approach reduced arbitrary personnel decisions and offset other sources of unreliability in the system.

In summary, the four case districts attempted to achieve reliability, but in every case problems existed. Reliability is not one independent issue. It depends upon the construction of the instrument, the skills and training of the evaluator, the behaviors being evaluated, and variations due to the training situation. Therefore, it would be fair to state that most evaluation systems do not produce reliable results.

Reliability has been discussed from the viewpoints of four selected districts. It is important to consider reliability as a statistical measure and discover how

reliability affects the broad concept of teacher evaluation.

In order for teachers, administrators, and school board members to have confidence in an evaluation system, reliability of the evaluation instrument and process must be established. According to Brown (1968), there are three major problems involved in establishing reliability:

1. Selecting types of reliability appropriate to the instrument and the purpose for which it is designed.

2. Selecting a meaningful measure of reliability once the type is specified.

3. Selecting a good estimator of a given measure to give an estimate of reliability based on experimental data.

Statistically, reliability refers to consistency by way of a series of measurements. This is usually expressed in terms of reliability coefficients. Thorndike (1950) studied reliability from two very different viewpoints. First, the approach is about the actual or absolute magnitude of errors of measurement. This type of reliability is secured by scores of repeated testing of the same teacher and is based on standard error of measure. The second approach is when individuals maintain the same relative position in the total group on repetition of a measurement procedure. This type of reliability is

expressed in terms of correlation between two scores called coefficients of reliability.

Reliability becomes even more complex when studying measurement of classroom behavior by systematic classroom observation. Reliability of the evaluators and the documenting of their observations must be added to the dilemma. Most studies have limited their view of reliability to correlation between two classroom observations or to computing the percent of agreement between evaluators.

The percent of inter-rater agreement tells almost nothing about the accuracy of the observation. Research has shown that there can be 99% agreement in recording classroom behaviors on an instrument that has very poor item or category consistency. Even though inter-rater agreement is high, reliability can be low.

It is possible for observers to agree that a certain teaching behavior took place during the observation. Yet, if the exact behavior occurs consistently in all classrooms, the reliability of that behavior as a degree of difference among teachers will be zero. Errors that occur from variations in behaviors from one observation to another are far more important than the failure of two observers to agree exactly in their observations of a given teacher.

The reliability of most evaluation instruments that record the behaviors of teachers require a high percent of inter-rater agreement. "Between-observer" agreement has become a requirement when planning an evaluation system. According to Medley and Mitzel (1963),

> A sample of classrooms from the population to be studied should be visited by trained recorders using the observation instrument in the same way it will be used in the subsequent study. In order to study the "objectivity" of the item, i.e., how closely observers agree in recording identical behaviors, at least two recorders should be present on each visit, sitting in different parts of the room making independent records. (p. 255)

Medley and Mitzel (1963) also found in their studies that if the evaluation system has high reliability, the scoring of the teacher should be free of the opinions of the recorders, or the different conditions under which the evaluation was done. If this can be accomplished, the instrument would be considered "good" or reliable.

One way to ensure higher reliability is to train the evaluators in the ability to score identical classroom behaviors. Even though evaluators are consistently and carefully trained in the proper use of the instrument,

there still may be variables that could influence an evaluator's judgment.

Reliability coefficients which stress strong inter-rater agreement imply that one "single, uniform, objective" system for observing teaching behavior should be used. Medley and Mitzel (1963) state that "between-observer agreement may not only encourage a false sense of confidence with respect to the accuracy of measurements, but also gives a false sense of 'objectivity' regarding the observation" (p. 300). These authors came to the conclusion that if several observers scored a filmed teaching situation and if the observers scored the same teaching behaviors in the same way, then it would be safe to say that the inter-rater score was reliable. They also noted that "observer reliability is always subject to variations in the selection and training of people and the control of conditions under which they use the instrument" (p. 320). It is imperative, therefore, to understand the internal consistency of the evaluation system. If the evaluators are properly trained and are capable of judging effective teaching, an evaluation system with internal consistency and item reliability would be a good measure of total reliability.

Even though Medley and Mitzel (1963) have done extensive research in observer agreement, they agree with

McGraw, Wardrop, and Bunda (1972) that observer agreements alone are not the only indices of reliability. Herbert and Attridge (1975) have urged that those who develop evaluation systems provide data dealing with reliability as well as "a discussion of which reliability measures were selected, and why" (p. 14).

It is unfortunate that reliability may mean one thing in one context and something else in another. In the classroom setting, it has been assumed that reliability is a property that is observed. However, it has never been clear what it is that possesses this reliability. The instrument itself is neither reliable nor unreliable; "it is only when the instrument has been used to collect data, and when the data have been manipulated in some way to produce scores, that we can speak of reliability" (Rowley, 1974, p. 16). Any measure can be reliable or unreliable depending on the way the evaluation is used, the people evaluated, the expertise of the evaluator, and the number of observations, and the time given to the observations. The most important note to make would be which measures produced from an evaluation are reliable and which are not.

Even though inter-rater agreement is not the only reliability issue, it is one of the primary concerns. Inter-rater agreement is usually defined as the "consistency among observers when they are simultaneously

coding the same classroom event" (Frick & Semmel, 1978, p. 159). Researchers usually determine agreement by comparing one observation with another. One faulty assumption is that inter-rater agreement means the observational measure is reliable. Medley, Coker, and Soar (1984) point out that this assumption is not always true. They state that "two observers may put a group of teachers in the same rank order with respect to the amount of negative affect they express, but assume that one regularly sees more affect (gives higher scores) than the other" (p. 65).

In order to prevent observer unreliability, observers should have to prove that they have been adequately trained in the use of the instrument. Even though observers demonstrate high standards of success during training does not ensure reliable observations. Johnson and Bolstad (1973) cite that even when there was high observer agreement right after training, the agreement had a tendency to decline with the passing of time. It is sensible then to conclude that continuous training should be provided to evaluators.

Johnson and Bolstad (1973) note that it is necessary to consider how the collected data will be used when one is looking into observer agreement. If the data are to be analyzed by individual categories, it is necessary that the observer agreement measures be scored using the totals for

each category. Therefore, it should be stated that observer agreement should be calculated using the same behaviors that will be used in the data analyses.

Usually, observer agreement is based on two or more observers comparing their scores or using a key to compare results. According to Ebel (1951), Haggard (1958), and Medley and Mitzel (1958, 1963), obtainment of agreement among observers is possible for each category by using intraclass correlation coefficients if one assumes that an analysis based on categorical frequencies is of interest.

In order to reduce the many problems of using inter-rater agreement, the rater's scores should be compared with an expert coder. This type of rater agreement is known as criterion-related agreement. This type of agreement is more useful when one is making decisions about the adequacy of individual skills.

Moreover, "criterion-related observer agreement lends itself well to the design of instructional materials for observer training" (Thiagarajan, 1973, p. 104). Borich and Malitz (1972) state that "one problem in drawing conclusions from observational studies relating teacher behavior to pupil growth is that it is difficult to generalize findings from independent studies in which information about the behavioral variables is adequate" (p. 75).

Perfect observer agreement is desirable, but the conditions needed to achieve this have not been established. Medley and Norton (1971) note that it should only be necessary to document that the observers have been properly trained and to show evidence that there was agreement when they scored teaching behaviors shown on videotape. In the real world, however, perfect observer agreement may not be desirable. Different teaching situations, content, and student behavior would all influence any given situation. Because of the varying classroom situations, observer disagreement is probably a better picture of what occurs with most evaluation systems.

Trying to reach perfect observer agreement may hinder validity. However, if an observer views a videotape of a typical classroom situation which is the same length as a regular observation and which is fairly representative of the actual situation, the observer can be tested for agreement when he or she scores specific behaviors after he or she has viewed the film twice. This way the observer can see how consistent he or she is. Johnson and Bolstad (1973) state that when an observer goes into the field, inter-rater agreement will be less than in a testing situation.

One must remember that when human judgment is the basis for measurement, the judgment must be based on

careful observation or examination of evidence; in addition, those rendering judgments must have appropriate background against which to compare their observations.

Teacher evaluation is an activity that must satisfy competing personal and organizational needs. The imperative of equal treatment for personnel decisions may result in the standardizing of acceptable teaching behaviors. However, research on teacher performance and teacher effectiveness does not lead to a stable, reliable, valid list of observable teaching behaviors that are effective in all teaching situations. Moreover, research on personal and organizational behavior indicates the need for specific strategies for improving teaching rather than district-wide hierarchial mandates. If teacher evaluation is to be useful for teacher improvement, the process must strive for a balance between standardized, district-administered performance expectations and teacher-specific approaches to evaluation and professional growth and development.

CHAPTER 3

PROCEDURES FOR DATA COLLECTION

The purposes of the study include assessing the degree of inter-rater reliability of appraisers who use the Texas Teacher Appraisal Instrument to award exceptional quality points and the degree of inter-rater reliability among appraisers including elementary and secondary certified appraisers, appraisers with five years or less administrative experience and those with more than five years, male and female appraisers, and type of training completed by the appraisers. Included, also, are the interpretation and analysis of these data in order to compare the reliability of appraisers when scoring EQ points and among the different demographic groups cited.

In Chapter 3 is a description of the procedures used for collecting data, the data-gathering procedures and instrument, and the population of subjects. The testing procedure and the procedures for statistical treatment of the data are also explained.

Description of the Instrument

For the purposes of this study, the data were collected from 707 certified (or certified after training) appraisers from Region X Educational Service Center area.

These appraisers were required to view a 45-minute videotaped lesson and independently score the nine EQ performance criteria on the official TTAI. After scoring the instrument, subjects were asked to complete voluntarily a demographic data questionnaire (Appendix B) indicating size of district (ADA), years of administrative experience, previous teaching experience (elementary or secondary), present level as an appraiser (elementary or secondary), age, ethnicity, present position, type of training, social security number, and sex. Over the course of gathering the data, 710 subjects were asked to participate in this study. The final number of subjects was 707.

Permission was obtained from Region X Education Service Center and the Texas Education Agency to collect and evaluate data regarding exceptional quality points. This permission was granted with the understanding that the subjects would remain anonymous and would agree to participate in the study by completing the questionnaire.

Selection and Description of Subjects

The subjects consisted of educators involved in TTAS training. All subjects were admitted into the training sessions after meeting certification requirements established by House Bill 72. The total number of subjects asked to participate was 710. The actual number completing the questionnaire was 707.

The subjects in this study were from 56 different school districts in the Region X Education Service Center area. This area of North Texas includes a major metropolitan area and many surrounding communities. The subjects came from school districts that ranged in Average Daily Attendance (ADA) from less than 500 to over 125,000. The data relative to district ADA are shown in Table 1.

Table 1

The Number of Subjects Categorized by the Average Daily Attendance of the 56 School Districts

Average Daily Attendance	Number of Subjects	%
Less than 500	44	6.2
501 - 1500	76	10.7
1501 - 5000	182	25.7
5001 - 7500	17	2.4
7501 - 10,000	12	1.7
10,001 - 25,000	120	17.0
25,001 and over	256	16.2
Total	707	100.0

The subjects ranged in years of administrative experience from 0 to 38. The data relative to administrative years of experience are presented in Table 2.

Table 2

Years of Administrative Experience of Subjects

Years of Administrative Experience	Number of Subjects	%
0	112	15.8
1	53	7.5
2	52	7.4
3	51	7.2
4	31	4.4
5	45	6.4
6	34	4.8
7	26	3.7
8	37	5.2
9	23	3.3
10	33	4.7
11	10	1.4
12	29	4.1
13	28	4.0
14	23	3.3
15	27	3.8
16	6	.8
17	14	2.0
18	10	1.4
19	7	1.0
20	15	2.1
21	1	.1
22	7	1.0
23	4	.6
24	3	.4
25	4	.6
26	2	.3
27	9	1.3
28	0	0.0
29	1	.1
30	2	.3
31	1	.1
32	0	0.0
33	0	0.0
34	1	.1
35	0	0.0
36	1	.1
37	0	0.0
38	5	.7
Total	707	100.0

The subjects included appraisers with elementary certification, secondary certification, and dual certification. Data relative to level of certification are given in Table 3.

Table 3

Certification Level of Subjects

Certification	Number of Subjects	%
Elementary	248	35.1
Secondary	379	53.6
Dual*	80	11.3
Total	707	100.0

*The significance of the dual certified appraisers is discussed in Chapter 4.

The age span of subjects ranged from 25 years and under to 46 years and over. The data relative to age are presented in Table 4.

Ethnicity designations were specified as Caucasian, Black, Hispanic, and Other. Data relative to ethnicity are given in Table 5.

Table 4

Age of Subjects

Age of Participants	Number of Subjects	%
25 or under	2	.3
26-35	171	24.2
36-45	302	42.7
46+	232	32.9
Total	707	100.0

Table 5

Ethnicity of Subjects

Ethnicity	Number of Subjects	%
Caucasian	602	85.1
Black	76	10.8
Hispanic	18	2.5
Other	11	1.6
Total	707	100.0

Present positions held by subjects included 16 categories spanning from classroom teacher to superintendent. The data relative to present positions held by the subjects are shown in Table 6.

Table 6

Present Position Held by Subjects

Present Position	Number of Subjects	%
Elementary Principal	153	21.6
Secondary Assistant Principal	137	19.4
Secondary Principal	66	9.3
Curriculum Director	66	9.3
Supervisor/Coordinator	59	8.3
Elementary Assistant Principal	53	7.5
Classroom Teacher	29	4.1
Director--Elementary/Secondary	28	4.0
Administrative Intern	25	3.5
Outside Appraiser	20	2.8
Superintendent	18	2.5
Department Head	17	2.4
Assistant Superintendent	13	1.8
Dean of Instruction	10	1.4
Diagnostician/Counselor	9	1.3
Service Center	4	.6
Total	707	100.0

Appraisers in the state are required to take an initial 40-hour training session to become certified. Each year after this initial training, all are required to

attend update sessions to maintain their certification as appraisers. Some subjects in this study were involved in the initial training sessions, and others who had already completed the initial training were involved in the update training. The data relative to the type of appraiser training are presented in Table 7.

Table 7

Number of Subjects Involved in Each Type of Training

Training	Number of Subjects	%
40-hour initial training	237	33.5
Two-day update training	470	66.5
Total	707	100.0

Data relative to sex of the subjects within the sample population are shown in Table 8.

Table 8

Number of Male and Female Subjects

Sex	Number of Subjects	%
Male	394	55.7
Female	313	44.3
Total	707	100.0

Basic Procedures

The procedures used in the study included the following:

1. The subjects participated in the training session appropriate for their needs (40-hour initial training or two-day update training).

2. The training sessions followed the agenda stated by the Texas Education Agency department for Texas teacher appraisal training.

3. The subjects viewed the videotaped lesson and awarded EQ points on the official observation form.

4. The subjects were given the demographic data questionnaire and were asked to participate voluntarily in the study.

5. There were 710 subjects involved in the training sessions, and 707 chose to complete the questionnaire and to participate in the study.

The data were collected beginning in May, 1987, and continued through October, 1988. It should be noted that the two-day update participants were appraisers who had been actively conducting evaluations during the previous year. Forty-hour training subjects were those who were completing the initial training and had not evaluated teachers using this system.

Procedures for the Treatment of Data

The measurements of the inter-rater reliability study were analyzed after the treatment of raw data using differences between the coefficients of correlation, number of subjects, percentages, and a multiple response procedure. Appropriate comparisons were made based on the hypotheses of this study. The data are reported in tables in conjunction with their respective hypotheses in Chapter 4. Raw data are converted and reported at the .05 level of significance to serve the purposes of this study. Conclusions and recommendations based on the findings of this study are reported in Chapter 5.

Although the subjects were grouped according to certification, sex, years of administrative experience, and type of training, the groups consisted of individuals independently scoring an observation form, and the groups cannot be regarded as matched. The coefficients of correlation allowed each group to be treated on the order of an intact group compared to another intact group.

Summary

In this study, 707 subjects from 56 school districts were involved in a quasi-experimental study on inter-rater reliability. The writer attempted to study a population of appraisers similar to the present population of appraisers in the state of Texas. It should be noted that the ethnic

composition of the subjects indicated a higher concentration of Caucasians and Blacks and a lower concentration of Hispanics in North Texas when compared to overall state ethnic composition. The Texas appraiser population consists of a higher occurrence of males than females while the survey population of males to females is fairly equal. With these differences acknowledged, the writer is satisfied that the subject population remains a strong representative sample of the Texas appraiser population. The study led to the comparison of special groups of subjects in terms of how they awarded EQ points on the TTAI. Special statistical groupings were formed to test the inter-rater reliability of these groups of appraisers.

The procedures for gathering the data, selecting the subjects, and the procedures for implementing the study are described. Finally, the analysis of the statistical data is detailed.

CHAPTER 4

PRESENTATION AND ANALYSIS OF DATA

The purpose of this chapter is to report, analyze, and interpret the findings of this study. In reporting the data of the study, each hypothesis is presented in turn along with the data pertinent to it. The data for each hypothesis were statistically treated using coefficients of correlation, number of subjects, percentages, and a multiple response procedure. Following the reporting of data by hypothesis, a discussion of findings provides the researcher interpretation of the experimental results. Finally, ancillary data relating to peripheral aspects of the study are presented and discussed.

Data Related to Hypothesis 1

The statement of Hypothesis 1 is as follows: There will be no significant inter-rater reliability on each of the nine performance criteria on which exceptional quality points are measured on the Texas Teacher Appraisal Instrument. The data relative to Hypothesis 1 are shown in Table 9.

For the purposes of this study, reliability is set at 90% because the only variable is the appraiser. All subjects viewed the same lesson under the same conditions

Table 9

A Comparison of the Number of Subjects that Awarded or Did Not Award Exceptional Quality Points on the Texas Teacher Appraisal Instrument

Criterion	EQ Points Awarded N	EQ Points Awarded %	EQ Points Not Awarded N	EQ Points Not Awarded %
1	550	77.8	157	22.2
2	412	58.3	295	41.7
3	441	62.4	266	37.6
4	434	61.4	273	38.6
5	125	17.7	582	82.3
6	146	20.7	561	79.3
7	115	16.3	592	83.7
8	75	10.6	632	89.4
9	130	18.4	577	81.6

so the reliability should be higher than if the data had been collected in the field. Since 90% is the stated reliability of this study, the research indicates that none of the criteria are considered reliable. There are various reasons why certain criteria are near 90% and others are not.

Criterion 1. Provides opportunities for students to participate actively and successfully. The subjects that

awarded EQ points probably awarded them on the activity level of the students. The concept of student success, however, may have been when the remaining 22% did not award EQ points.

Criterion 2. Evaluates and provides feedback on student progress during instruction. This criterion was fairly equal in the scoring. This criterion deals with a numerical issue of quantity. The subjects that considered quantity of feedback given would award EQ points. Those that looked for quality of feedback would probably choose not to award credit.

Criterion 3. Organizes materials and students. The subjects that awarded EQ points saw a high quality of materials and the effective, efficient use of them during the lesson. Those who chose not to award EQ points probably thought that what the teacher did was not out of the ordinary for an elementary teacher.

Criterion 4. Maximizes amount of time available for instruction. Before EQ points were moved to the criterion level in 1987, none of the indicators in this criterion were eligible for exceptional quality. Appraisers who were trained in 1986-1987 were not accustomed to awarding EQ points to any of these indicators. This study indicated that 61% chose to award EQ points. These subjects probably felt that the teacher used every available minute for

instruction. The other 38% may have felt that the lesson was too involved or it was taught at a normal pace.

Criterion 5. Manages student behavior. Once again, none of the indicators in this criterion were eligible for EQ points before 1987. Many appraisers feel that if there are no behavior problems, the teacher has not done anything exceptional. Others feel that if no behavior problems occur, it is exceptional because obviously the teacher has done something in the classroom to establish proper behavior. This study indicated that 82.3% agreed not to award EQ points. They must have the philosophy that if there are a few or no problems, EQ points are not awarded.

Criterion 6. Teaches for cognitive, affective, and/or psychomotor learning and transfer. On this criterion, EQ points should be awarded if the quality of the presentation allows students to be successful and gives them the opportunity to apply the learning at a higher cognitive level. In this study, 79% decided not to award credit so they must have felt that student success at higher cognitive levels was not evident.

Criterion 7. Uses effective communication skills. This criterion, like criteria four and five, was not eligible for EQ points in 1986-1987. This criterion is difficult to judge for EQ points. The indicators are all yes or no type decisions. The teacher has to demonstrate

the skills to receive standard credit, but the indicators are not EQ issues. In this study, 83.7% decided not to award credit. It is safe to assume that most people do not see this criterion as eligible for EQ points. Inter-rater reliability is very high on this criterion.

Criterion 8. Uses strategies to motivate students for learning. In looking for EQ points on this criterion, the appraiser would observe how engaged and committed the students were to the activity, how long it took the teacher to get them started, and how reliant the teacher is on mechanical devices to get students motivated to participate. In this study, it is very clear that 89.4% of the subjects felt the students were mechanically motivated and internally committed to the lesson. The inter-rater reliability is very high on this criterion.

Criterion 9. Maintains supportive environment. This is the most subjective criterion in respect to awarding EQ points. The appraiser would look for a warm, safe environment for students where they would feel comfortable participating without the risk of failure or ridicule. In this study, 81.6% of the subjects felt the environment was not conducive to warm feelings. In comments made during training, many felt sarcasm was a big problem with this teacher.

Since none of the criteria indicated inter-rater reliability according to the 90% set by this study, Hypothesis 1 was accepted.

Data Related to Hypothesis 2

The statement of Hypothesis 2 is as follows: There will be no significant difference between the inter-rater reliability of elementary certified appraisers who evaluate elementary teachers and secondary certified appraisers who evaluate elementary teachers. The data relative to Hypothesis 2 are shown in Table 10.

In the comparison of evaluations of elementary teachers by elementary certified appraisers to evaluations of elementary teachers by secondary certified appraisers, the coefficient of correlation was ascertained at +.036 (indicating very little correlation in evaluations). Further statistical treatment based on

H_1: Elementary appraisers deny > 60.1% of the EQ points

H_0: Elementary appraisers deny $\leq$ 60.1% of the EQ points

the resultant z of 3.64 was sufficient to verify difference at the .05 level of significance. The null hypothesis was rejected, and the claim that elementary certified appraisers award more EQ points was supported. Therefore, Hypothesis 2 was rejected.

Table 10

Exceptional Quality Scoring Per Criterion by Elementary, Secondary, and Dual Certified Subjects

		Number of EQ Points Subjects Awarded by Criterion								
Certification	N	1	2	3	4	5	6	7	8	9
Elementary	248	183	137	161	159	38	39	29	22	40
Secondary	379	307	224	229	235	80	89	74	46	77
Dual	80	74	79	75	79	17	30	16	8	16
Total	707									

	Total EQ's Awarded		Total EQ's Denied		Total EQ's
Certification	N	%	N	%	Possible
Elementary	808	36	1424	64	2232
Secondary	1261	40	2050	60	3411
Dual*	391	53	329	46	720

*For purposes of this study, dual certification subjects are addressed as ancillary information only and do not relate to a specific hypothesis.

Data Related to Hypothesis 3

The statement of Hypothesis 3 is as follows: There will be no significant difference between the inter-rater reliability of male and female appraisers. The data relative to Hypothesis 3 are shown in Table 11.

Table 11

Exceptional Quality Scoring Per Criterion by Male and Female Subjects

		Number of EQ Points Subjects Awarded by Criterion								
Sex	N	1	2	3	4	5	6	7	8	9
Male	394	358	274	225	274	88	95	79	44	86
Female	313	230	164	190	199	47	63	40	32	45
Total	707									

	Total EQ's Awarded		Total EQ's Denied		Total EQ's
Sex	N	%	N	%	Possible
Male	1523	47	1699	53	3222
Female	960	34	1857	66	2817

In the comparison of EQ points awarded and denied by male appraisers to EQ points awarded and denied by female appraisers the coefficient of correlation is represented at -.00235 (indicating negligible correlation in the evaluation process). Further statistical treatment based on

H_1: Male appraisers give > 34% of the EQ points

H_0: Male appraisers give $\leq$ 34% of the EQ points

the resultant z of 11.31 was sufficient to verify difference at the .05 level of significance. The null

hypothesis was rejected, and the claim that males give more EQ points was supported. Therefore, Hypothesis 3 was rejected.

Data Related to Hypothesis 4

The statement of Hypothesis 4 is as follows: There will be no significant difference between the inter-rater reliability of appraisers with five years or less administrative experience and those with more than five years administrative experience. The data relative to Hypothesis 4 are shown in Table 12.

In the comparison of EQ points awarded and denied by appraisers with five years or less administrative experience to EQ points awarded and denied by appraisers with more than five years administrative experience, the coefficient of correlation was found to be -.0104 (indicating insignificant correlation). Further statistical treatment based on

H_1: Appraisers with more than five years administrative experience will score > 39% of the EQ points

H_0: Appraisers with more than five years administrative experience will score $\leq$ 39% of the EQ points

the resultant z of 1.545 was sufficient to verify difference at the .05 level of significance. The null

Table 12

Exceptional Quality Scoring Per Criterion by Subjects with Five Years or Less Administrative Experience and Subjects with More than Five Years

		Number of EQ Points Subjects Awarded by Criterion								
Experience	N	1	2	3	4	5	6	7	8	9
5 Years or Less	344	288	199	222	216	71	72	51	36	49
More than 5 years	363	300	239	253	254	64	96	68	40	81
Total	707									

	Total EQ's Awarded		Total EQ's Denied		Total EQ's
Experience	N	%	N	%	Possible
5 Years or Less	1204	39	1892	61	3096
More than 5 years	1395	43	1872	57	3267

hypothesis was rejected, and the claim that appraisers with more than five years administrative experience award more EQ points was supported. Therefore, Hypothesis 4 was rejected.

Data Related to Hypothesis 5

The statement of Hypothesis 5 is as follows: There will be no significant difference between the inter-rater

reliability of appraisers involved in initial 40-hour training and those involved in the yearly update training. The data relative to Hypothesis 5 are shown in Table 13.

Table 13

Exceptional Quality Scoring Per Criterion by Subjects Involved in 40-Hour Initial Training and Subjects Involved in Two-Day Update Training

		Number of EQ Points Subjects Awarded by Criterion								
Training	N	1	2	3	4	5	6	7	8	9
40-hour	237	120	80	98	90	21	28	14	12	32
Update	470	408	316	325	324	98	110	98	59	92
Total	707									

	Total EQ's Awarded		Total EQ's Denied		Total EQ's
Training	N	%	N	%	Possible
40-hour	1638	77	495	23	2133
Update	2400	57	1830	43	4230

In the comparison of EQ points awarded and denied by appraisers involved in 40-hour initial training and EQ points awarded and denied by appraisers with two-day update training, the coefficient of correlation was -.0212

(indicating weak association). After testing data for the level of significance based on

H_1: Appraisers involved in 40-hour initial training will score > 57% of the EQ points

H_0: Appraisers involved in 40-hour initial training will score $\leq$ 57% of the EQ points

the resultant z of 3.020 was sufficient to verify difference at the .05 level of significance. The null hypothesis was rejected, and the claim that appraisers involved in 40-hour initial training gave more EQ points was supported. Therefore, Hypothesis 5 was rejected.

Discussion of Findings

In this section, findings are discussed in terms of the groups and variables investigated in the study. Discrepancies may occur as a result in variable of the sample size of 707 subjects as compared to the actual population of Texas appraisers (17,000). Other variables, while not addressed but nonetheless acknowledged, would entail consideration of fund availability for exceptional quality training as well as fund availability to support changes of career ladder placement.

Inter-rater Reliability by Appraisers Using the Texas Teacher Appraisal Instrument

An examination of EQ points awarded by apraisers using the TTAI does not indicate a constant pattern for overall scoring. If each criterion is analyzed, none will reach 90% reliability according to the standards set by this study. However, the results indicate that the inter-rater reliability of the subjects in denying EQ points came very close to meeting the 90% standard on criteria five (manages behavior), seven (communication), eight (motivating students), and nine (supportive environment).

Elementary and Secondary Certified Appraisers

An examination of EQ points awarded by elementary certified appraisers and secondary certified appraisers using the TTAI did not indicate any noticeable trend. The findings support the idea that elementary certified appraisers award slightly fewer EQ points to an elementary teacher than do secondary certified appraisers scoring the same teacher. A new dimension is added by those appraisers with dual certification (elementary and secondary). The dual certified appraisers awarded more EQ points than elementary appraisers and were more consistent with the rankings by the secondary appraisers.

Male and Female Appraisers

An examination of EQ points awarded by male appraisers indicated that more points were given than from the female appraisers. The statistical findings indicated a negative correlation between the two groups of appraisers, which leaves the significance debatable. It should be noted that once again both groups agreed to award EQ points on criteria one (active and successful participation), two (evaluates feedback), three (organizes students), and four (maximizes time), and deny EQ points on five (manages behavior), six (cognitive, affective, psychomotor learning), seven (communication), eight (motivating students), and nine (supportive environment). However, only 34% of the females awarded EQ points as compared to 47% of the males awarding points.

Years of Administrative Experience of Appraisers

An examination of EQ points awarded by appraisers with five years or less administrative experience did not show any noticeable trend when compared to appraisers with more than five years experience. Again, both groups agreed to give EQ points on criteria one (active and successful participation), two (evaluates feedback), three (organizes students), and four (maximizes time), while denying EQ points on five (manages behavior), six (cognitive, affective, psychomotor learning), seven (communication),

eight (motivating students), and nine (supportive environment). The percentages of agreement and denial were extremely consistent between the two groups. The appraisers with more than five years gave slightly more EQ points and denied slightly fewer EQ points.

Type of Appraiser Training: Forty-Hour or Two-Day Update

An examination of EQ points awarded by appraisers with 40-hour initial training indicated that more EQ points were awarded than by appraisers who completed the two-day update sessions. It should be noted that there were twice as many update subjects as 40-hour subjects. The two-day update appraisers were consistent when they awarded EQ points in the first four criteria. The 40-hour subjects had 51% agreement when they awarded EQ points on criterion one (active and successful participation) only. This trend is noticeably different from the other groups of appraisers tested. These results indicate that appraisers who go through the training for the first time are not likely to award EQ points as consistently as appraisers who have been conducting appraisals for at least one year. This finding might indicate to teachers that appraisers who are conducting appraisals for the first time would be beneficial to them if they are aspiring to move up on the career ladder.

Ancillary Data

For the purpose of evaluating the frequency and percentage of EQ points awarded on each of the nine criteria, tables which present number of subjects and percentage which address the five hypotheses have been included in this section. Individual EQ points were not tested for level of significance in each hypothesis since overall conclusions were more desirable for the hypotheses stated. However, examination of the number of subjects who awarded and denied EQ points within each group provides observational data for the individual criterion addressed on the TTAI.

The data indicate a lack of consensus among elementary, secondary, and dual certified appraisers in awarding EQ points on each of the nine criteria (Table 14). The dual certified appraisers, however, met the 90% reliability of this study on criteria one (active and successful participation), two (evaluates feedback), three (organizes students), and four (maximizes time) when they awarded EQ points. Elementary and dual certified appraisers also met inter-rater reliability on criterion eight (motivating students) when they chose not to award EQ points. Responses from the appraisers at each level of

Table 14

Number of Subjects, Frequency, and Percentage Table for Table 10: Exceptional Quality Scoring per Criterion by Elementary, Secondary and Dual Certified Subjects

Criterion	Elementary	EQ's Awarded	%	Secondary	EQ's Awarded	%	Dual	EQ's Awarded	%
1	248	183	74	379	307	81	80	74	93
2	248	137	55	379	224	59	80	79	99
3	248	161	65	379	229	60	80	75	94
4	248	159	64	379	235	62	80	79	99
5	248	38	25	379	80	21	80	17	21
6	248	39	16	379	89	24	80	30	38
7	248	29	12	379	74	20	80	16	20
8	248	22	9	379	46	12	80	8	10
9	248	40	16	379	77	20	80	16	20

certification are too skewed to indicate any level of predictability. Therefore, information shown by frequency of responses and percentage of EQ points awarded by the three groups of appraisers supports the conclusion that there is no significant inter-rater reliability between elementary and secondary certified appraisers who evaluate an elementary teacher.

The data indicate that male appraisers award more EQ points on every criteria except criterion three (organizes students) (Table 15). On criterion three, the percentage spread is a low 4%. The greatest degree of disagreement, 18%, occurs on criteria two (evaluates feedback) and four (maximizes time) while the lowest degree of disagreement, 4%, occurs on criteria five (manages behavior) and six (cognitive, affective, psychomotor learning).

For the purposes of this study, with 90% reliability, the male subjects had 91% inter-rater reliability on criterion one (active and successful participation). For females, criterion eight (motivating students) met the reliability standards. Therefore, information shown by frequency of responses and percentage of EQ points awarded by the two populations supports the conclusion that there is no significant inter-rater reliability between male and female appraisers in the awarding of EQ points.

Table 15

Number of Subjects, Frequency, and Percentage Table for Table 11: Exceptional Quality Scoring per Criterion by Male and Female Subjects

Criterion	Male	EQ's Awarded	%	Female	EQ's Awarded	%
1	394	358	91	313	230	73
2	394	274	70	313	164	52
3	394	225	57	313	190	61
4	394	274	70	313	199	64
5	394	88	22	313	47	15
6	394	95	24	313	63	20
7	394	79	20	313	40	13
8	394	44	11	313	32	10
9	394	86	22	313	45	14

The data indicate that appraisers with more than five years of administrative experience awarded more EQ points on criteria two (evaluates feedback), three (organizes students), four (maximizes time), six (cognitive, affective, psychomotor learning), seven (communication), eight (motivating students), and nine (supportive environment) (Table 16). Overall, the range of percentages was very consistent between the two groups. Appraisers

Table 16

Number of Subjects, Frequency, and Percentage Table for Table 12: Exceptional Quality Scoring per Criterion by Subjects According to Years of Administrative Experience

Criterion	5 years or less	EQ's Awarded	%	More than 5 years	EQ's Awarded	%
1	344	288	84	363	300	83
2	344	199	58	363	239	66
3	344	222	65	363	253	70
4	344	216	63	363	254	70
5	344	71	21	363	64	18
6	344	72	21	363	96	26
7	344	51	15	363	68	19
8	344	36	10	363	40	11
9	344	49	14	363	81	22

with five years or less administrative experience met the reliability standard of this study on criterion eight (motivating students) only. This type of information indicates that years of administrative experience do not make a significant difference on inter-rater reliability of EQ points.

The data indicate that appraisers who received two-day update training awarded more EQ points on all nine criteria

(Table 17). On criterion one (active and successful participation), two (evaluates feedback), three (organizes students), and four (maximizes time), the percentage difference is almost 30%. The lowest percentage of disagreement was on criterion nine (supportive environment). For the purposes of this study, the subjects involved in initial 40-hour training were at least 90% reliable on criteria five (manages behavior), seven

Table 17

Number of Subjects, Frequency, and Percentage Table for Table 13: Exceptional Quality Scoring per Criterion by Subjects in Forty-Hour and Two-Day Update Training

Criterion	40-hour Training	EQ's Awarded	%	Two-day Update	EQ's Awarded	%
1	237	120	51	470	408	86
2	237	80	34	470	316	67
3	237	98	41	470	325	69
4	237	90	38	470	324	69
5	237	21	9	470	98	21
6	237	28	12	470	110	23
7	237	14	6	470	98	21
8	237	12	5	470	59	13
9	237	32	14	470	92	20

(communication), and eight (motivating students). The two-day update subjects did not meet the 90% standard on any criteria. Even though 40-hour subjects did meet reliability on three criteria, there is no significant difference between the two populations.

The data indicate that the size of school district (ADA) does not make a significant difference in the number of EQ points awarded by subjects (Table 18). The percentage ranged from a high of 50% for appraisers in districts with an ADA of 5,001 to 7,500. The lowest

Table 18

A Comparison of Exceptional Quality Points Awarded According to Size of School District Average Daily Attendance

Criteria	Size of District	# of Subjects per ADA	Total EQ's Awarded	%
1-9	Less than 500	44	154	39
1-9	501 - 1500	76	266	39
1-9	1501 - 5000	175	637	40
1-9	5001 - 7500	17	76	50
1-9	7501 - 10,000	12	32	30
1-9	10,001 - 25000	120	352	33
1-9	25,001 and over	256	917	40

percentage was 30% from districts with 7,501 to 10,000 average daily attendance. Therefore, information presented in Table 18 supports the theory that size of ADA does not make a significant difference in the number of EQ points awarded by appraisers.

The data indicate that when size of school district was paired with subjects having five years or less administrative experience there was no significant difference (Table 19). However, when subjects had more than five years experience, there were disparities. The percentages ranged from a high of 50% in school districts of less than 500 to 28% in districts with 7,501 to 10,000 students. According to the reliability standard set by this study, inter-rater reliability was not met. Therefore, one can conclude that this type of comparison did not show any difference when subjects awarded EQ points.

Summary

The introduction presented in Chapter 1 revealed that reliability refers to the consistent results given by an instrument. Objective research findings have not supported this definition of reliability. Findings in the present study indicated that the demographic groupings of subjects did not show any significant inter-rater reliability in the awarding of EQ points. This study supported the concern of

Table 19

A Comparison of Exceptional Quality Points Awarded According to Size of School District Average Daily Attendance and Years of Administrative Experience

Size of District	Years of Experience	Number of Subjects	Total EQ's Awarded	%
Less than 500	a. 5 or less	31	99	35
	b. more than 5	13	58	50
501 - 1500	a. 5 or less	60	158	29
	b. more than 5	29	95	36
1501 - 5000	a. 5 or less	89	292	36
	b. more than 5	106	403	42
5001 - 7500	a. 5 or less	55	30	36
	b. more than 5	9	36	32
7501 - 10,000	a. 5 or less	6	16	30
	b. more than 5	8	20	28
10,001 - 25000	a. 5 or less	54	134	28
	b. more than 5	73	230	35
25,001 and over	a. 5 or less	128	442	38
	b. more than 5	137	528	43

teachers, administrators, and appraisers that the TTAI is not reliable between appraisers.

The question of significant inter-rater reliability when set at 90% has several results: (a) All appraisers awarding EQ points per criterion did not meet 90% reliability for various reasons already stated; (b) elementary certified appraisers awarded slightly more EQ points than secondary appraisers; (c) although not significant, male appraisers awarded more EQ points than female appraisers; (d) administrative experience did not make a significant difference when appraisers awarded EQ points, and (e) appraisers involved in 40-hour training awarded more EQ points, again not at a significant level.

Questions could be raised in regard to the videotape used to measure the variables in this study. It could be that the videotape used was not the best or most appropriate for the specific purpose intended. This, however, is one of the limitations of this study. No other videotape was made for the purposes of testing EQ points. Numerous videotapes need to be made available to test inter-rater reliability in the future. A perplexing aspect of this study is that teachers are being judged by this instrument and placed on career ladder according to the scores awarded by the appraiser. If there is no significant reliability, as reported by this study, the TEA needs to review the importance placed on the Texas Teacher Appraisal System.

CHAPTER 5

SUMMARY, FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

Chapter 5 presents a summary of the nature and procedures of the study, conclusions based on the findings, and recommendations for further study.

Summary

The purposes of this study were to measure the degree of inter-rater reliability by appraisers who use the Texas Teacher Appraisal Instrument to award exceptional quality points. The other purpose was to measure the degree of inter-rater reliability among different demographic groups of appraisers. These groups included (a) elementary and secondary certified appraisers, (b) male and female appraisers, (c) appraisers with five years or less administrative experience and those with more than five years experience, and (d) appraisers involved in initial 40-hour training or two-day update training.

A total of 707 subjects from 56 school districts in North Texas was used for this study. The study was conducted in training sessions required for certification of appraisers and update renewal certification of appraisers. The purpose of the study was presented to

training subjects, and permission was granted by all who completed the demographic questionnaire during training.

Following the state-mandated training session in which the subjects were involved, a videotaped lesson was shown and exceptional quality points were awarded or denied. The data collected in this study were subjected to statistical treatment and analysis. Data were arranged to include frequencies, percentages, coefficients of correlation, and multiple response procedures.

The instrument used in this study was the official Texas Teacher Appraisal Instrument that was mandated by the Texas State Board of Education in 1985. All subjects viewed the same lesson and received the same training before completing the observation record.

Summary of Findings

Based on the quasi-experimental data collected in this study (reported in Chapter 4), the following findings are presented.

1. The data related to the inter-rater reliability of appraisers in the awarding of EQ points on the nine criteria did not indicate a constant pattern for scoring EQ points. For the purposes of this study, reliability was set at 90%, and this was not met on any criterion.

2. Elementary certified appraisers awarded slightly fewer EQ points to elementary teachers than secondary

certified appraisers awarded. Dual certified appraisers awarded a slightly higher percentage of EQ points than the elementary certified appraisers.

3. Male appraisers awarded more EQ points than their female counterparts.

4. Appraisers with more than five years administrative experience awarded slightly more EQ points than those with five years or less administrative experience.

5. Subjects involved in the 40-hour training sessions awarded many more EQ points than those involved in the two-day update training sessions.

6. Ancillary data indicated that the size of school district did not have any significance on the number of EQ points awarded by the appraisers. Along with this, it was noted that there was no significance in the awarding of EQ points by the appraisers when the size of school district was compared with years of administrative experience.

Conclusions

In this study, significance was set at the .05 level, and inter-rater reliability was set at 90%. Based on the findings and limitations of this study, the following conclusions may be drawn.

1. Appraisers who award EQ points on the TTAI are not consistent throughout the instrument. On the nine

performance criteria, the appraisers did not meet 90% reliability on any criterion.

2. Appraiser certification (elementary or secondary) does not affect the appraiser's decision to award or deny EQ points on the TTAI. Even though 90% reliability was not met, the subjects were consistent in deciding to award EQ points on criteria one through four and in deciding not to award them on criteria five through nine. This consistency should be noted.

3. Male appraisers had the tendency to award 13% more EQ points than the female appraisers. Again, both groups were consistent with the decision to award EQ points on criteria one through four and deny on criteria five through nine.

4. Years of administrative experience did not prove to be significant when appraisers awarded EQ points on the TTAI. These two groups were very close in their decisions to award or deny credit. Both groups were consistent in the decision to award EQ credit on criteria one through four and deny EQ credit on five through nine.

5. Forty-hour training subjects had the tendency to award more EQ points than those involved in two-day update training. It needs to be noted that 40-hour training subjects are participating in training for the first time. Two-day update subjects have previously completed the

40-hour training and are now completing the required yearly update session.

6. The size of school district (ADA) was not significant in the awarding of EQ points. Analysis of the results found that the highest percentage of inter-rater reliability was 50%.

Recommendations

Based on the results and the conclusions of this study, the following recommendations for future investigations are projected.

1. More specific criteria should be used in training appraisers to award or deny exceptional quality points on the TTAI. As was stated previously, these have been deleted from the training manual. At the present time, no information except the definition of exceptional quality is given in training sessions. If the concept of inter-rater reliability is important, and it is, some specific information needs to be established in order to gain consistency across the state. Another reason to gain consistency is because of career ladder. Exceptional quality points determine scores above satisfactory on the instrument. Career ladder placement depends on a teacher achieving a level of exceeding expectations and clearly outstanding to maintain placement on the ladder.

2. Further studies should be conducted to emphasize the conclusion that certification (elementary or secondary) and previous teaching experience or years of administrative experience have no significance on the outcome of the appraisal with respect to the awarding of EQ points. This is a major concern of teachers, and it would benefit appraisers and teachers to study this and disprove theories that an appraiser has to have experience in the field for which he is appraising.

3. More extensive training needs to be developed for all appraisers so inter-rater reliability will increase. Since 40-hour initial training subjects had the tendency to award more EQ points than two-day update subjects, further study needs to be conducted to discover why there is a difference. It would also be useful for groups of appraisers (40-hour and two-day update together) to view lessons and discuss where and why exceptional quality points should be awarded.

APPENDIX A

TEXAS TEACHER APPRAISAL INSTRUMENT

School District _____ Campus _____

TEXAS EDUCATION AGENCY

Texas Teacher Appraisal System

Observation/Evaluation Record

School Year 19___-___

Observation Record Date _____
Evaluation Record Date _____

Teacher _____ Assignment/Grade _____ Appraisal Period 1 or 2 (circle)

Appraiser _____ Title: Teacher's Supervisor _____ Other Appraiser _____

Subject Area Observed _____ Observation Date _____

Beginning Time _____ Ending Time _____ Scheduled _____ Unscheduled _____

TEACHER'S SUPERVISOR:

1. After each formal observation an OBSERVATION RECORD must be completed for Domains I-V. Record the date on which the OR is completed in the space provided in the upper right hand corner of this form.
2. For each indicator observed and/or credited, circle the numeral 1. Evidence concerning indicators for which credit is denied must be documented in the space provided.
3. For each criterion in which Exceptional Quality is awarded, circle the numeral 3. Evidence concerning the basis for awarding EQ credit must be documented in the space provided.
4. At the end of each appraisal period and/or prior to the summative conference, an EVALUATION RECORD must be developed. Review the completed OBSERVATION RECORD(S) and any cumulative data collected up to the end of the appraisal period to determine whether changes need to be made regarding SE and EQ credit. Record the date the EVALUATION RECORD is developed in the space provided in the upper right hand corner of this form. If after reviewing the data there are no changes to be made, complete steps 5 and 6 below. If previously awarded SE or EQ credit is to be denied, strike through the circled numeral. If credit which was previously denied is now to be awarded, circle the appropriate numeral, initial and date each change and record documentation to substantiate the change(s) in the space provided.
5. For Domain V, credit is automatically awarded unless documentation justifies denial.
6. For each domain, record the total credits earned during the appraisal period (SE + EQ) in the space provided.

OTHER APPRAISER(S):

1. After each formal observation, an EVALUATION RECORD must be completed for Domains I-IV. Record the date on which the ER is completed in the space provided in the upper right hand corner of this form.
2. For each indicator observed and/or credited, circle the numeral 1. Evidence concerning the basis on which credit for an indicator has been denied must be documented in the space provided.
3. For each criterion for which Exceptional Credit is awarded, circle the numeral 3. Evidence concerning the basis for awarding EQ credit must be documented in the space provided.
4. For each domain, record the total credits earned in the space provided.
5. If the teacher's supervisor has scored the teacher's performance in Domain V less than satisfactory, review documentation and score Domain V.

I. Instructional Strategies

		SE
1. Provides opportunities for students to participate actively and successfully.	a. varies activities appropriately	1
	b. interacts with group(s) appropriately	1
	c. solicits student participation	1
	d. extends responses/contributions	1
	e. provides time for response/ consideration	1
	f. implements at appropriate level	1
	Exceptional Quality	3
2. Evaluates and provides feedback on student progress during instruction.	a. communicates learning expectations	1
	b. monitors student performance	1
	c. solicits responses/demonstrations for assessment	1
	d. reinforces correct response/ performances	1
	e. provides corrective feedback/ clarifies/none needed	1
	f. reteaches/none needed	1
	Exceptional Quality	3

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

II. Classroom Management and Organization

Teacher _____

3. Organizes materials and students.

a. secures student attention	1
b. uses procedures/routines	1
c. gives clear administrative directions/none needed	1
d. maintains appropriate seating/grouping	1
e. has materials/aids/facilities ready	1
Exceptional Quality	3

4. Maximizes amount of time available for instruction.

a. begins promptly/avoids waste at end	1
b. implements appropriate sequence of activities	1
c. maintains appropriate pace	1
d. maintains focus	1
e. keeps students engaged	1
Exceptional Quality	3

5. Manages student behavior.

a. specifies expectations for behavior/none needed	1
b. prevents off-task behavior/none needed	1
c. redirects/stops inappropriate/disruptive behavior/none needed	1
d. applies rules consistently and fairly/none needed	1
e. reinforces desired behavior when appropriate	1
Exceptional Quality	3

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

III. Presentation of Subject Matter

6. Teaches for cognitive, affective, and/or psychomotor learning.

a. begins with appropriate introduction	1
b. presents information in appropriate sequence	1
c. relates content to prior/future learning	1
d. defines/describes concepts; skills, attitudes, interests	1
e. elaborates critical attributes	1
f. stresses generalization/principle/rule	1
g. provides for application	1
h. closes instruction appropriately	1
Exceptional Quality	3

III. Presentation of Subject Matter *(continued)*

Teacher _____

7. Uses effective communication skills.

a. makes no significant errors 1
b. explains content/task(s) clearly 1
c. stresses important points/dimensions 1
d. uses correct grammar 1
e. uses accurate language 1
f. demonstrates written skills 1

Exceptional Quality 3

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

IV. Learning Environment

8. Uses strategies to motivate students for learning.

a. relates content to interests/ experiences 1
b. emphasizes value/importance of activity/content 1
c. reinforces/praises efforts 1
d. challenges students 1

Exceptional Quality 3

9. Maintains supportive environment.

a. avoids sarcasm/negative criticism 1
b. establishes climate of courtesy 1
c. encourages slow/reluctant students 1
d. establishes and maintains positive rapport 1

Exceptional Quality 3

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

V. Professional Growth and Responsibilities

10. Plans for and engages in professional development.

a. progresses in growth requirements or none needed 1
b. stays current in content taught 1
c. stays current in instructional methodology 1

V. Professional Growth and Responsibilities *(continued)*

Teacher _____

11. Interacts and communicates with parents.

a. initiates communications with parents as appropriate 1
b. conduct conferences with parents in accordance with local policy 1
c. reports student progress to parents 1
d. maintains confidentiality 1

12. Complies with policies, operating procedures, and requirements.

a. follows TEA requirements 1
b. follows district/campus policies/ procedures 1
c. performs assigned duties 1
d. follows promotion procedures 1

13. Promotes and evaluates student growth.

a. participates in goal-setting 1
b. plans instruction 1
c. documents progress 1
d. maintains records 1
e. reports progress 1

FOR EVALUATION RECORD
DOMAIN CREDIT TOTAL _____ (SE)

Comments:

Teacher Signature/Date Received OR _____ Appraiser Signature/Date Completed OR _____ Date of Conference (if any) _____

Teacher Signature/Date Received ER _____ Appraiser Signature/Date Completed ER _____ Date of Conference (if any) _____

(The signature of the teacher indicates that he/she has reviewed and received a copy of this record.)

Original Copy—Central Office
Copy #2—Teacher's Supervisor
Copy #3—Teacher

APPENDIX B

DEMOGRAPHIC DATA

DEMOGRAPHIC DATA

I. Size of District (ADA)
1. Less than 500 2. 501 - 1500 3. 1501 - 5000
4. 5001 - 7500 5. 7501 - 10,000 6. 10,001 - 25,000
7. 25,001 and over

II. Years of Administrative Experience _____

III. Previous teaching Experience/Certification
1. Elementary 2. Secondary

IV. Present Level as an Appraiser
1. Elementary 2. Secondary

V. Age
1. 25 or under 2. 26-35 3. 36-45 4. 46+

VI. Race
1. Caucasian 2. Black 3. Hispanic 4. Other

VII. Present Position
1. Elementary Principal 2. Secondary Principal
3. Elementary Assistant Prin. 4. Secondary Assistant Prin.
5. Director - Elem./Sec. 6. Supervisor/Coordinator
7. Dean of Instruction 8. Classroom Teacher
9. Department Head 10. Diagnostician/Counselor
11. Curriculum Director 12. Administrative Intern
13. Outside Appraiser 14. Service Center
15. Superintendent 16. Assistant Superintendent

VIII. Type of Training
1. 40-hour training 2. Update training

IX. Social Security Number (last 4 digits) _____
(This will be used for computer purposes.)

X. Sex
1. Male 2. Female

Exceptional Quality Data

Scores on OR/ER: Please put a check by the criterion in which you awarded exceptional quality points.

1. ___ 2. ___ 3. ___ 4. ___ 5. ___ 6. ___ 7. ___ 8. ___ 9. ___

Note: Completion of this questionnaire means that you agree to participate in this study. Thank you for your help.

REFERENCES

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), Handbook of teacher evaluation (pp. 110-145). Beverly Hills, CA: Sage Publications.

American Association of School Administrators. (1979). Staff dismissal: Problems and solutions. Reston, VA: Author.

Bloom, B. S. (1956). Taxonomy of educational objectives--Handbook I cognitive domain. New York: Longman.

Borich, G. D. (1977). The appraisal of teaching: Concepts and process. Reading, MA: Addison-Wesley.

Borich, G. D., & Malitz, D. (1972). Convergent and discriminant validation of three classroom observation systems: A proposed model. Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Brophy, J. E. (1973). Stability of teacher effectiveness. American Educational Research Journal, 19, 245-252.

Brophy, J., & Evertson, C. (1974). Product-product correlations in the Texas Teacher Effectiveness Study: Final report (Report No. 74-4). Austin, TX: The University of Texas, Research and Development Center for Teacher Education. (ERIC Document Reproduction Service No. ED 091 394)

Brophy, J. E., & Evertson, C. M. (1976). Learning from teaching: A developmental prospective. Boston: Allyn and Bacon.

Broudy, H. S. (1956). Craft or profession? The Educational Forum, 21, 175-184.

Brown, B. B. (1968). The experimental mind in education. New York: Harper and Row.

Centra, J. A., & Potter, D. A. (1980). School and teacher effects: An interrelational model. Review of Educational Research, 50(2), 273-291.

Cole, R. W., Jr. (1979). Minimum competency tests for teachers: Confusion compounded. Phi Delta Kappan, 61 (December), 233.

Coleman, J., Campbell, E. A., Hobson, C. J., McPartland, J., Mood, A., Weinfeld, F. D., & York, R. L. (1966). Equality of education opportunity. Washington, DC: U.S. Government Printing Office.

Commission on Elementary Schools. (1970). Guide to conducting programs of school improvement. Atlanta, GA: Southern Association of Colleges and Schools.

Cronbach, L. J. (1963). Course improvement through evaluation. Teachers College Record, 64, 672-683.

Cronbach, L. J., & Snow, R. E. (1977). Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research, 63(3), 285-328.

Davidoff, S. H. (1970). The development of an instrument designed to secure student assessment of teaching behaviors that correlate with objective measures of student achievement. Philadelphia, PA: The School District of Philadelphia, Office of Research and Evaluation.

Deneen, J. (1971). Black teachers: Uses and abuses of tests. National Council on Measurement in Educational News, (Fall).

Doyle, W. (1978). Paradigms for research on teacher effectiveness. In L. S. Shulman (Ed.), Review of Research in Education, 5 (pp. 163-198). Itasca, IL: Peacock.

Dunkin, M. J., & Biddle, B. J. (1974). The study of teaching. New York: Holt, Rinehart, and Winston.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psycometrika, 16, 407-424.

Educational Commission of the States: A Task Force in Education for Economic Growth. (1983). Action for excellence: A comprehensive plan to improve our nation's schools. Denver, CO: Author.

Eisner, E. W. (1978). On the uses of educational connoisseurship and criticism for evaluating classroom life. Teachers College Record, 78, 345-358.

Ellett, C. D., Capie, W., & Johnson, C. E. (1980). Assessing teaching performance. Educational Leadership, 38(3), 219-220.

Evertson, C. M., & Holley, F. M. (1981). Classroom observation. In J. Millman (Ed.), Handbook of teacher evaluation (pp. 90-109). Beverly Hills, CA: Sage Publications.

Finlayson, H. J. (1979). Incompetence and teacher dismissal. Phi Delta Kappan, 61(September), 69.

Frick, T., & Semmel, M. I (1978). Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 48(1), 157-184.

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

Gagne, R. M., & Briggs, L. (1979). Principles of instructional design (2nd ed.). New York: Holt, Rinehart, and Winston.

Gallup, G. H. (1979). The 11th annual Gallup Poll of the public's attitude toward the public schools. Phi Delta Kappan, 61, 33-45.

Garawski, R. A. (1980). Successful teacher evaluation not a myth. National Association of Secondary Principals Bulletin, 64, 1-7.

Guthrie, J. (1970). Survey of school effectiveness studies. In A. Mood (Ed.), Do teachers make a difference? (pp. 160-182). Washington, DC: U.S. Government Printing Office.

Haefele, D. L. (1980). How to evaluate thee, teacher-let me count the ways. Phi Delta Kappan, 61(5), 349-352.

Haefele, D. L. (1981). Teacher interviews. In J. Millman (Ed.), Handbook of teacher evaluation (pp. 73-90). Beverly Hills, CA: Sage Publications.

Haggard, E. A. (1958). Intraclass correlations and the analyses of variance. New York: Dryden Press.

Hardebeck, R. J. (1973). A comparison of observed and self-reported individualization of instruction by vocational, academic, and special education teachers in Texas. Unpublished doctoral dissertation, The University of Texas, Austin.

Harris, B. M. (1979). Orientation on branching diagram analysis. Studies in Educational Evaluation, 5, 157-162.

Harris, B. M. (1986). Developmental teacher evaluation. Boston, MA: Allyn and Bacon.

Harris, W. U. (1981). Teacher command of subject matter. In J. Millman (Ed.), Handbook of teacher evaluation (pp. 58-72). Beverly Hills, CA: Sage Publications.

Herbert, J., & Attridge, C. (1975). A guide for developers and users of observation systems and manuals. American Educational Research Journal, 12(1), 1-20.

Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hammerlynch, L. C. Hardy, & E. J. Mash (Eds.), Behavior change (pp. 160-188). Champaign, IL: Research Press.

Lewis, A. (1982). Evaluating educational personnel. Arlington, VA: American Association for School Administrators.

Lewis, J. (1973). Appraising teacher performance. New York: Parker Publishing Co.

Madeus, G. F., Kellaghan, T., & Rakow, E. A. (1979). Within school variance in achievement: School effects or error? Studies in Educational Evaluation, 5, 101-107.

Mazur, J. L., & Peterson, D. D. (1978). Lesson organization: A system for observing related teacher behaviors. Unpublished master's thesis, University of South Florida, Tampa.

McDonald, F. J. (1976). Summary report: Beginning teacher evaluation system, phase II. Princeton, NJ: Educational Testing Service.

McDonald, F. J., & Elias, P. (1976). Executive summary report: Beginning teacher evaluation study, phase II. Princeton, NJ: Educational Testing Service.

McGraw, B., Wardrop, J. L., & Bunda, M. A. (1972). Classroom observation schemes: Where are the errors? American Educational Research Journal, 9(1), 13-27.

McIntyre, K. E. (1979). Evaluation of teaching: Can a formula be found? Texas Association of School Boards Journal, 5, 12-16.

McNeil, J., & Popham, W. (1973). The assessment of teacher competence. In R. M. Travers (Ed.), Second handbook of research on teaching (pp. 220-231). Chicago: Rand McNally.

Medley, D. M. (1982). Teacher competency testing and the teacher educator. Charlottesville: University of Virginia, Association of Teacher Educators and the Bureau of Education Research.

Medley, D. M., Coker, H., & Soar, R. S. (1984). Measurement-based evaluation of teacher performance: An empirical approach. New York: Longman.

Medley, D. M., & Mitzel, H. E. (1958). Application of analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior. Journal of Experimental Education, 27, 23-35.

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching (pp. 247-328). Chicago: Rand McNally.

Medley, D. M., & Norton, D. F. (1971). The concept of reliability as it applies to behavior records. Paper presented at the meeting of the American Psychological Association, Washington, DC.

Millman, J. (Ed.). (1981). Handbook of teacher evaluation. Beverly Hills, CA: Sage Publications.

Mitchell, D. E., & Kerchner, C. T. (1983). Collective bargaining and teacher policy. In L. S. Shulman & G. Sykes (Eds.), Handbook of teaching and policy (pp. 19-29). New York: Longman.

National Education Association. (1979). Teacher opinion poll. Washington, DC: Author.

National Study of School Evaluation. (1973). Elementary school evaluative criteria. Arlington, VA: National Study of School Evaluation.

Peterson, K., & Kauchak, D. (1982). Teacher evaluation: Perspective, practices, and promises. Salt Lake City: University of Utah, Center for Educational Practice.

Peterson, P. L. (1976). Interactive effects of student anxiety, achievement orientation, and teacher behavior on student achievement and attitude. Unpublished doctoral dissertation, Stanford University, California.

Peterson, P. L. (1979). Direct instruction reconsidered. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching (pp. 57-69). Berkeley, CA: McCutchan.

Phi Delta Kappan. (1979). Gallup says public wants greater productivity. News, Notes, and Quotes, 24, 3.

Popham, W. J. (1971). Designing teacher evaluation systems. Los Angeles, CA: The Instructional Objectives Exchange.

Quirk, T. J., Witten, B. J., & Weinberg, S. F. (1973). Review of studies of the concurrent and predictive validity of the National Teacher Examination. Review of Educational Research, 43, 89-114.

Redfern, G. B. (1980). Evaluating teachers and administrators: A performance objective approach. Boulder, CO: Westview Press.

Rosenshine, B. (1970). The stability of teacher effects upon student achievement. Review of Educational Research, 40, 647-662.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R. M. Travers (Ed.), Second handbook of research on teaching (pp. 122-183). Chicago: Rand McNally.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. C. Wittcock (Ed.), Handbook of research on teaching (3rd ed.) (pp. 85-94). New York: MacMillan.

Rossmiller, R. A. (1983). Resource allocation and achievement: A classroom analysis. In A. Odden & L. D. Webb (Eds.), School finance and school improvement: Fourth annual yearbook of the American Education Finance Association (pp. 213-226). Cambridge, MA: Ballinger Publishing.

Rowley, G. (1974). Reliability of observational measures. American Educational Research Journal, 13, 51-59.

Sellitz, C., Wrightsman, L. S., & Cool, S. W. (1976). Research methods in social relations. New York: Holt, Rinehart, and Winston.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teacher behavior. Review of Educational Research, 40, 553-612.

Shavelson, R., & Russo, N. A. (1977). Generalizability of measures of teacher effectiveness. Educational Research, 19, 171-183.

Shavelson, R., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. Review of Educational Research, 51, 455-498.

Shine, W. A., & Goldman, N. (1980). Reply to Fred G. Burke. Educational Leadership, 38, 201.

Soar, R. S. (1972). Follow through classroom process measurement and pupil growth. Gainesville: University of Florida, Institute for Development of Human Resources.

Texas Education Agency. (1989). Texas teacher appraisal system: Appraiser's manual. Austin, TX: Author.

Thiagarajan, S. (1973). Instructional systems for interactional systems. Classroom Interaction Newsletter, 9(1), 13-22.

Thorndike, R. L. (1950). Reliability. In E. F. Lundquist (Ed.), Educational measurement (pp. 356-442). Washington, DC: American Council on Education.

United States National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. A report to the nation and the Secretary of Education, U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Veldman, D. J., & Brophy, J. E. (1974). Measuring teacher effects on pupil achievement. Journal of Educational Psychology, 66, 319-324.

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). Teacher evaluation: A study of effective practice. Santa Monica, CA: Rand.