1980

A comparison of school administrators' ratings of teaching performance utilizing varied instructional materials in the assessment process

Sarah Jane Frudden
Iowa State University

FRUDDEN, SARAH JANE

A COMPARISON OF SCHOOL ADMINISTRATORS' RATINGS OF TEACHING PERFORMANCE UTILIZING VARIED INSTRUCTIONAL MATERIALS IN THE ASSESSMENT PROCESS

Iowa State University PH.D. 1980

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs _____
2. Colored illustrations _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Print shows through as there is text on both sides of page _____
6. Indistinct, broken or small print on several pages ✓
7. Tightly bound copy with print lost in spine _____
8. Computer printout pages with indistinct print _____
9. Page(s) _____ lacking when material received, and not available from school or author
10. Page(s) _____ seem to be missing in numbering only as text follows
11. Poor carbon copy _____
12. Not original copy, several pages with blurred type _____
13. Appendix pages are poor copy _____
14. Original copy with light type _____
15. Curling and wrinkled pages _____
16. Other _____

A comparison of school administrators' ratings of teaching performance utilizing varied instructional materials in the assessment process

by

Sarah Jane Frudden

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Department: Professional Studies in Education
Major: Education (Educational Administration)

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1980

TABLE OF CONTENTS

LIST OF TABLES

Page

CHAPTER I. INTRODUCTION

Teacher evaluation has become an integral part of our educational system, and even though the process is a common one, no standardized procedure has evolved (Knezevich, 1977). While the approaches to the appraisal process are many and diverse (Haefele, 1980), the consensus is that improved evaluation techniques logically should lead to improved instruction, and finally, to improved educational benefits for children and youth (Thomas, 1979).

What precisely is to be evaluated in the instructional process has not been definitively agreed upon by evaluators (Robinson, 1978), but almost all would concur that sound planning of instruction, which is usually recorded in lesson plans and objectives for the instructional sequence, is an essential condition to good instruction (Popham, 1969; Taba, 1962; Tyler, 1950). And, if good planning is where good instruction begins, is it not reasonable to assume that it is also where good evaluation should begin? It is widely known that many teacher evaluators will evaluate a teacher and the lesson without considering the congruence of instruction to the objectives or the triviality of the objectives (Berliner, Note 1; Scriven, 1977) and, furthermore, will not examine the intended lesson plans or instructional objectives for the ensuing instruction (Robinson, 1978).

Brighton (1965) and Kowalski (1979) have stated that educational administrators often feel inadequate to advise teachers about curricular matters, including the appropriateness of lesson plans and objectives,

for assessment or improvement of instructional performance. They believe that some administrators may abdicate their roles as instructional leaders because they lack the skills and knowledge for recommending specific educational strategies to assist the teacher. While the literature makes reference to the administrator/evaluator's perceived lack of ability to deal with curricular matters, few studies have been conducted to investigate the situation. Little research has been undertaken to see if administrators are skilled at the identification of appropriate educational planning and the execution of appropriate instructional strategies.

Statement of the Problem

This study will investigate the problem of the educational administrator's ability to accurately assess a teacher's preinstructional planning, and to determine if the use of a performance criteria instrument will assist in more accurate assessment of both preinstructional materials and classroom procedures. The investigation is also designed to ascertain if the utilization of low inference items on the criteria rating instruments will facilitate accurate evaluation.

The research will also determine what influence the appraiser's study of preinstructional materials has upon his accuracy in assessing teacher performance. In other words, will accuracy be affected by whether or not an appraiser previews the teacher's planning before observing and evaluating the instruction. Also, data will be gathered to determine if an appraiser who more accurately assesses the preinstructional materials

similarly will more accurately assess the teacher's classroom performance.

Finally, the study will provide information to specify if the amount of teacher appraisal training an evaluator has had, the administrative job position of the evaluator, or the regional origin of the evaluator, will affect evaluative accuracy of preinstructional materials and classroom procedures of the teacher.

The Hypotheses

Certainly the teacher appraiser's examination of lesson plans, objectives, activities, materials, and evaluation techniques (the preinstructional materials) in relation to the course outline should provide evidence of greater curriculum-congruent instruction when assessing a teacher's performance. It is to the joining of curriculum content to teacher behavior that this study is directed and can more specifically be defined by the following statements which are the hypotheses to be investigated:

1. Will teacher appraisers more accurately assess teacher performance given preinstructional materials for study during the preobservation period as compared with those who are not provided with preinstructional materials for study?
2. Will teacher appraisers more accurately assess teacher performance given preinstructional materials and evaluative criteria for assessing those materials as compared with those who are provided the preinstructional materials but not the evaluative

criteria?

3. Will teacher appraisers more accurately assess teacher performance given preinstructional materials and an evaluative instrument utilizing low inference descriptors on the rating scale as compared with those provided with preinstructional materials and an evaluative instrument utilizing high inference descriptors on the rating scale?
4. Will teacher appraisers more accurately assess preinstructional materials utilizing an evaluative instrument with low inference descriptors on the rating scale as compared with those utilizing an evaluative instrument with high inference descriptors on the rating scale?
5. Is there a relationship between the accuracy of the rater judging the preinstructional materials and the accuracy of the rater judging teacher performance?
   5a. If the rater utilized a high inference evaluative instrument, did those who more accurately judged the preinstructional materials also more accurately judge the teacher's performance?
   5b. If the rater utilized a low inference evaluative instrument, did those who more accurately judged the preinstructional materials also more accurately judge the teacher's performance?
6. Will previous training in teacher appraisal affect the rater's accuracy when rating preinstructional materials?
7. Will previous training in teacher appraisal affect the rater's accuracy when rating teacher performance?

8. Will there be regional differences among teacher appraisers in their rating accuracy of preinstructional materials?
9. Will there be regional differences among teacher appraisers in their rating accuracy of teacher performance?
10. Will there be differences in rating accuracy of preinstructional materials among administrative job positions?
11. Will there be differences in rating accuracy of teacher performance among administrative job positions?

Definitions

SUPERVISOR--The school officer designated to evaluate, appraise, guide, and direct the work of teachers, usually with the expectation that such instruction will be of better quality because of the supervision.

PERFORMANCE (instructional or educational) OBJECTIVES--Statements of what the learner will be able to do after completing the course of study assigned; may be in terms of observable action or a product.

PREINSTRUCTIONAL MATERIALS--The materials produced as a result of the teacher's planning prior to instruction. These may include the lesson plans, instructional objectives, educational activities, materials, worksheets, and instructional evaluation techniques.

EVALUATIVE CRITERIA--Any of a wide variety of items which are or can be used as standards for judging the performance of a teacher or the appropriateness and quality of preinstructional materials.

LOW INFERENCE ITEMS ON EVALUATIVE CRITERIA--The increments on the evaluative scale which are fully described so assigmment of a rating on the

criteria will be facilitated by these descriptors.

HIGH INFERENCE ITEMS ON EVALUATIVE CRITERIA--The increments on the evaluative scale of the criteria which are open to conjecture and subjective judgments by the evaluator.

Delimitations

The information to be gathered in this study will come from the utilization of a simulated teacher assessment exercise using appropriate preinstructional materials, a videotaped instructional sequence, and an associated evaluative criteria for both of these components. The experiment will be conducted as a part of a workshop in teacher performance evaluation which will be given in Iowa and other places across the country. The simulation will take approximately one and one-half hours with appropriate amounts of time being given to study of preinstructional materials and their evaluative rating, viewing of the taped instructional session, and evaluation of the classroom sequence. The project will be limited by the fact that subjects will be allowed only one opportunity to view the videotape and preinstructional materials, and will be granted only one opportunity to rate the evaluative criteria. This may be a factor in the reduction of reliability of these data.

Only one videotaped sequence will be used, an instructional session of an intermediate grade level language arts class, and this will be restricted only to that specific grade level and subject matter. Also, the session is filmed in such a manner that not all of the children in the class are shown. This narrows the observer's capability to get an optimum

perception of the classroom's functioning.

The subjects in this investigation will be educational administrators (certified professionals including principals, superintendents, supervisors, and central office administrators) involved in various professional improvement programs, i.e., workshops, conferences, or college classes, and therefore, will demonstrate a more positive disposition to the study. The participants will be all volunteers who in no manner are required to be an accessory to the research. This attitude is favorable to the ends of the study, and not typical of a total population.

The educational administrators to be sampled will come largely from groups seeking training and assistance in teacher performance appraisal, and generalizations cannot be made outside of that population.

CHAPTER II. REVIEW OF LITERATURE

> When teacher evaluation is discussed, contemplated, or written about, there is always agreement on one point - the need for it. However, opinions on why the need exists are as divergent as the groups that have the differing views. To further complicate this issue, opinions differ and often conflict on the goals and purposes of evaluation, on the criteria for evaluation, how it should be done, who should do it, and even on whether it is possible to have meaningful evaluation. (Wicks, 1973, pp. 42-43)

A History of Teacher Evaluation and Supervision

Teacher evaluation is not a new concept in this country. The observation of teachers has been conducted in American classrooms since the formation of schools by law in 1642 (Spears, 1953, pp. 37-38) when the first inspections were usually conducted by citizens of the community, including a member of the clergy, for the purpose of control and inspection. The emphasis was on ascertaining that the values of the school's teachings were consistent with the standards and beliefs of the patrons (Lamb and Swick, 1975, p. 240). These committees were less interested in improving deficient teachers, and were more interested in dismissing them from the system.

The growth of the urban areas as well as the increased desire for education began to swell the school populations. A demand resulted for greater numbers of teachers, and consequently the responsibilities for monitoring schools, which had previously been assumed by lay people, now were transferred to the person appointed as head teacher, or the person

filling that position (Barr and Burton, 1938, p. 3). By 1870, there were twenty-nine superintendents of schools serving as chief administrative officers, with the supervision of instruction as one of their duties. They worked for the improvement of instruction of deficient teachers, rather than dismissal (Lucio and McNeil, 1979, p. 4).

The Nineteenth Century saw the emergence of the school principal who was destined to assume a major share of the task of the supervision of teachers (Davis, 1964; Piper and Elgart, 1979). Inspection and control continued to be the major purpose of observation, but the matter became complex as an increased number of new subject areas, such as music, art, home economics, and physical education were introduced into the schools. Consequently, a new supervisory officer, the specialist, was utilized in a number of cities to share the supervisory responsibilities (Lucio and McNeil, 1979).

Confusion resulted when two supervisory officials--administrators and specialists--were overseeing, inspecting, and enforcing school regulations. Lowry, in 1908, however, placed the responsibility quite squarely on the principal when he said, "Whether in the regular or in the special subjects (responsibility) should be his and not that of the visiting supervisor, no matter how expert she may be in her particular line" (Lowry, 1908, p. 19).[1]

---

[1]Lowry's use of the impersonal pronoun is particularly revealing!

The scientific management era

The supervision and evaluation of teachers underwent a metamorphosis in the early 1900s with the rise of the scientific management age. This concept was developed by Fredrich Winslow Taylor in the business sector, and its byword was "efficiency" as emphasis was placed on standardization and systematization in the production of materials. Under scientific management, the focus of educational instruction was concentrated on the ends: development of the student, and the manipulation of the process was the means (Davis, 1964; Lucio, 1967).

School leaders began proposing the application of these principles to school supervision and calling for clear definitions of educational ends as well as coordination of all who worked to attain them. Teachers were held to certain standards of performance and rated accordingly. Emphasis was placed on product, measurement, and testing, and the attainment of set standards by pupils and teachers became the rule (Davis, 1964; Lucio, 1967; Piper and Elgart, 1979).

It was held that the important purpose of supervision was to discover the laws of teaching and learning, and to require teachers to apply these laws under direction. Lucio and McNeil state that during this period "supervisors were to: (1) discover best procedures in the performance of particular educational tasks, and (2) give these best methods to the teachers for their guidance" (Lucio and McNeil, 1979, p. 9). The assumptions underlying these views held that schools were staffed by persons who had to be led because "it was useless to look to teachers for sources of new ideas, that the problems of schools were known, but known to only

a few select individuals" (Lucio, 1967, p. 4).

It became common for supervisors to construct instruments to aid them in the description of teacher and student behaviors in the classroom. Indeed, the practice became so pronounced by the 1920s as to be accepted as a mechanical process. The use of rating scales for evaluating the efficiency of teachers was so widespread that they were acknowledged as "standard equipment" in a 1925 report by the Department of Classroom Teachers of the National Education Association (Spears, 1953, pp. 39-40).

The human relations era

In the 1930s, the Western Electric studies on worker motivation provided impetus for a new direction in educational supervision. Meeting the needs for social satisfaction of the employee (or teacher) was held to be motivating, and subsequently more productive to the organization. Educational supervisors endeavored to set relaxed environments where wide participation of staff was encouraged, and the improvement of the entire faculty became the goal. Cooperative planning and shared responsibilties were utilized to arrive at mutually agreed upon solutions.

Rating instruments continued to be in use, but whereas the devices had previously been utilized for control and inspection, they now were constructed to describe classroom behaviors. The data collection was for the purpose of research and the literature teems with citations of classroom experimental findings; for example, the study of classroom behavior conducted through sociometric techniques (Reemers, 1963), observations to describe the effect of teacher domination and integrative behaviors (Anderson, 1939), observing and recording bad health habits of

students (Urban, 1943), and an observation study describing small instructional group interaction (Bales, 1950). Perhaps one of the most noteworthy is the widely used observation instrument developed in the 1950s known as the Flanders System of Interaction Analysis (Flanders, 1965). This observational system was utilized to measure facets of teacher's indirect and direct influence on students, and many of the instruments created after this time reflected the influence of Flander's work both in categories of observation and methodology.

The revisionists era

Since 1950, these two viewpoints, the Scientific Management and the Human Relation, have undergone reconciliatory attempts wherein the unrealistic aspects of the Human Relationists approach is eliminated and the positive aspects of the Scientific Management theory are retained. The Revisionists attempted to consider both the individual's and organization's goals in their proper perspective while holding work is a natutal activity of man, organizational goals can be motivating to teachers, lack of control in an organization is undesirable, and teacher participation in decision-making is congruent with the accountability required of administrators.

Lucio claims that "A more balanced perspective of school supervision may result from the application of Revisionist suggestions concerning individual and instructional purposes and needs. Unitary emphasis on the purpose and demands of either the school or the individual would be 'deweighted' but neither would be 'devalued' at the expense of the other" (Lucio, 1967, pp. 8-9). Further, Lucio states,

> New ways of viewing organizational functioning, including the decentralization and realignment of decision-making processes, increased professional accountability for determining educational goals, planning teaching strategies, and the assessment of learning outcomes, suggest the following directions for supervision:
>
> 1. Supervision by objectives
> 2. Teacher - supervisor point responsibility
> 3. Differentiated supervision
>
> (Lucio, 1967, pp. 10-11)

The accountability era

The 1960s heralded in another era in education and it was predicated upon the restlessness and dissatisfaction of a society which acted upon those sentiments by demanding greater accountability in the schools (Harris, 1976; Thomas, 1979).

After the Soviets launched the Sputnik satellite in 1957, the American public began to have doubts that their educational system was keeping abreast with national needs. The disenchantment was further nutured by other circumstances such as the population explosion which was engorging school enrollments, the crash programs in teacher preparation attempting to meet enrollment pressures, and the poor selection and evaluation systems leaving no tenable alternatives (Tolor, 1973). Lessinger, called by many the father of accountability, felt that increased and accelerating costs, the poor scholastic performance of minority children, and the inconclusive results of massive compensatory programs also were contributing factors (Peach, 1977).

Lessinger (1970a) thought that citizens had three implicit rights in the educational process. First, they retain the right to be taught basic intellectual skills to become productive citizens. Second, taxpayers have the right to be informed of the educational results of educational

expenditure, and finally, school personnel have the right to draw from the communities' talent pool.

Further, Lessinger (1970b) believed that the educational designers must specify in measurable terms what is to be accomplished so there can be an independent audit of results. In his words:

> Accountability is the product of a process. At its most basic level, it means that an agent, public or private, entering into a contractual agreement to perform a service will be held answerable for performing according to agreed upon terms, within an established time period, and with a stipulated use of resources and performance standards. This definition of accountability requires that the parties to the contract keep clear and complete records and this information be available for outside review. It also suggests penalties and rewards; accountability without redress of incentive is mere rhetoric. (Lessinger, 1970b, p. 217)

Borich (1977) states that the concept of accountability also had impact upon congressional legislation. Documentation of effectiveness was required by federal agencies implementing innovative social and educational programs. This concept frequently filtered down and settled on the doorstep of city groups who pressed for concrete evidence of the effectiveness of their local schools, just as they had earlier supported accounting of the federally funded programs. Another force, state government, was influenced by the federal accountability and community concern, and began to enact legislation requiring the appraisal of school personnel. The most conspicuous piece of such legislation was passed by the General Assembly of California in 1971, and was known as the Stull Act. This act provided for the professional evaluation of all certified employees, from superintendents to certified aides, and specified the legal grounds for the dismissal of these employees.

Thus, conditions were ripe to induce the search for improved teacher assessment techniques and encourage the efforts of researchers to address the problem of identifying the effective teacher, an essential core to valid teacher assessment.

The Measurement of Teacher Effectiveness: Process, Characteristics, and Product

The quest for the prototype of the effective teacher has been a long one commencing with research by Barr et al. (1961), Morsh and Wilder (1954), and Tomlinson (1955). The studies have continued with many contributors; Cruickshank (1976), Gage (1963), Manatt, Palmer, and Hidlebaugh (1976), Mitzell (1960), Rosenshine and Furst (1971), and Ryans (1960) to name a few.

Traditionally, these studies have been grouped into three broad rubics (Brody, 1977; Lucio and McNeil, 1979). These three categories are:

1. Process. Teacher behavior is appraised against some standard of performance or set of actions (overt teaching acts) assumed or inferred to be related to effective teaching performance. If the teacher performs certain specific acts, pupil behavior then can be predicted. In this view, teacher performance may be described, rated, or observed in terms of factors such as: (a) how teachers structure learning situations (time and motion analysis), (b) extent and kind of pupil-teacher and/or teacher-pupil responses, and (c) analysis of teacher behavior by diverse systems, such as learner-centered versus teacher-centered behaviors, or various "psychiatric criteria" for assessing "good" or "bad" classroom pupil and/or teacher behavior.

2. Teacher characteristics. A variety of characteristics such as intelligence, personality traits, personal appearance, verbal skills, quality of speech, health, and other personal

attributes of teachers are assumed to be measures related to or predictive of effective teaching. Rating instruments, observation inventories, and reporting devices (containing indices assumed to relate to teaching ability) have been used widely to assess teacher performance.

3. Product (pupil-behavior change). Appraisal of teacher performance (instructional behavior) is focused on assessing defined changes in pupil behavior (on outcomes of teaching acts) rather than on the act itself or on teacher characteristics assumed to relate to pupil behavior. Thus the act of teaching is viewed as that which brings about a change in the learner. Appraisal of teaching by this criterion is concerned with the degree to which defined behavior or results are achieved by pupils instead of depending upon a teacher's congruence to some hypothetical model. (Lucio and McNeil, 1979, pp. 248-249)

There are those who would support teacher effectiveness as measured by product, pupil behavior change, as having the greatest promise for accountability in our schools. Many deem the definition of learning to incorporate the concept of pupil change. Indeed, Popham states, "the only reason a teacher is in a classroom is to modify his learner, to change him somehow" (Popham, 1969, p. 4). He continues, "I can't believe that a teacher could be considered successful who doesn't change his student" (Popham, 1969, p. 4). Smith and Smith support this notion when they say, "Every other type of accountability is secondary to accountability for student performance" (Smith and Smith, 1976, p. 189).

The specification of pupil change via educational objectives has become a boon to the advocates of educational accountability. According to Smith and Smith:

> Accountability demands that the teacher state specific educational objectives for the student, estimate the time necessary to achieve these outcomes, select the most educationally appropriate and highly motivating student activity for achieving the goals, and evaluate the degree to which the objectives have been achieved by each student. (Smith and Smith, 2976, p. 189)

Educational objectives give substance to the evaluation of teachers. Doherty and Hathaway comment, "As anyone who has attemped to evaluate teaching knows, absence of well-stated instructional goals has been the chief deterrent to teacher evaluation" (Doherty and Hathaway, 1972, p. 5). They continued by expressing their belief that the accountability for teacher performance in education necessitates the joining forces with the behavioral objective movement.

McNeil's perceptions are similar regarding assessment of teachers, and he believes, "A central concept of educational accountability is that the teacher produce evidence regarding the quality of his teaching--facts about what happens to pupils under his direction. Supervisors then must respond to this evidence, making decisions with respect to programs for teacher improvement, salary increase, teacher assignment and dismissal" (McNeil, 1972, p. 2).

Popham sees the evaluation of instruction being synonymous with the evaluation of instructional objectives. He feels that the last step in the instructional model is evaluation, not of the student, but of the quality of the instruction. He says that evaluation in this instance is the evaluation of the teacher and the merit of the instructional decisions made by the teacher, and "in general, unachieved objectives reflect instructional inadequacies, not deficiency in learners" (Popham, 1969, p. 5).

Instructional Planning

The need to establish and instruct from predetermined objectives is well-supported in educational circles (Popham and Baker, 1970a; Taba, 1962; Tyler, 1950).

Friedman remarks, "A good teacher would no more teach a class without a plan than would an engineer build a bridge without a blueprint" (Friedman, 1974, p. 440). However, he goes on to say that lesson plans are unpopular with teachers who often endure the task because it's "the supervisors' hangup" (p. 441).

This perspective is shared by Morine as she notes many educators regard lesson plan writing as a useless activity, and still she asserts, "It seems sensible to expect that improved planning for instruction will improve the quality of teaching that ensures. What is remarkable is that there is so little research evidence to support such a sensible notion" (Morine, 1973, p. 135).

How teachers plan

The virtue of planning is so well-accepted that hardly anyone has asked what difference it makes whether a teacher plans or not, or how a teacher plans, or how teachers state their goals, decide classroom events, or allow for instructional flexibility (Morine, 1976).

Zahorik (1970) did an interesting inquiry regarding a teacher's decision-making and planning. He observed that teachers who planned, in contrast to those who did not, when teaching the same lesson, more frequently "summarized pupils' responses, repeated original solicitations,

and tried to shape pupils' responses to reflect the teacher's views" (p. 148). In Zahorik's perception, teachers who planned exhibited less honest, or authentic use of pupils' ideas.

Morine (1976) states that teacher planning may not be a deliberate problem-solving activity as assumed, but more likely a routine activity of reporting pages to be covered. Zahorik (1970) conducted a survey from which he concluded that teacher planning centered first around content to be taught, the selection of activities was mentioned frequently, the choice of materials was reported by half of the teachers surveyed, and decisions about evaluation, diagnosis, and organization were cited by one-third of the teachers, and finally, specification of learning objectives was infrequently mentioned. Goodlad, Klein, and Associates (1974) found in their investigations that most teachers neither planned nor taught with specific objectives in mind, but were concerned with covering material. Similarly, Joyce and Harootunian (1964) found that few elementary science teachers prepared objectives preferring to depend on instructional materials for content and sequence. Popham and Baker (1970b) concluded that, without specific training, few teachers established behavioral objectives which were tied to instruction or evaluation devices.

The need for sound planning and decision-making is apparent. Hunter makes this point when she says "the human who teaches the student has potential power for exceeding that of any inanimate factor. Only recently, however, have we been able to establish that an important source of the human's power to make a learning difference lies in professional decisions and the behaviors that result from those decision" (Hunter, 1976, p. 162).

Teaching is viewed as a decision-making process with the teacher striving to reach some goal. The actual specification of these goals obviously is a topic of some contention. Williams (1976) feels that without an awareness of educational goals, purposeful decisions cannot be made.

Planning by objectives

"The argument in favor of objectives is a strong one. In essence, it is that if teachers are to guide learning, then they must have some criteria for guiding one way rather than another; they must have some ideas in mind which helps them in making the decisions they are constantly called upon to make" (Harlen, 1972, p. 225). Harlen maintains that the controversy is over whether or not educational goals should be articulated as statements of intended changes in student behaviors; the generality or the specificity of the articulated objectives lies at the heart of the argument. Tyler (1966) says the level of generality appropriate for objectives is probably the most puzzling question about objectives that curriculum workers must face.

Many educators are most determined to have objectives stated behaviorally to serve as a goal reference model of instruction to determine if the intended behavior changes have taken place in the learner. Among those advocates would be Bloom (1956), Mager (1962), Popham (1969), Taba (1962), and Tyler (1950).

Others would argue that we can never know or identify beforehand the outcomes of a learning experience, and frequently a learning situation leads to unanticipated situations which present more numerous outcomes

than could have been encompassed by more specific behavioral statements (Harlen, 1972). Friedman (1974) supports the idea that the specific format of lesson plans as well as the amount of detail required will vary from one teacher to another and will be influenced by their background, experience, and ability. He does, however, hold the teacher responsible for preparing an adequate plan.

Instructional objectives and pupil gain

Lucio (1973), who concurs with others in the belief that learning is evidenced by change--behavior change, contends that change in pupil behavior becomes the criterion for evaluating teacher performance. An experiment by McNeil (1967) was conducted in a preservice teacher training setting to explore this contention. He assigned 77 student teachers for two consecutive days to public school secondary placements. After selection of control and experimental groups, the experimental subjects and their supervising teachers were told to obtain agreement in advance as to what would constitute evidence of successful teaching in terms of pupil change. The experimental group submitted a "contract" of objectives to be achieved to the training teacher, whereas the control group submitted a copy of the two-day lesson plans.

Consistent with McNeil's hypothesis, more of the experimental group were viewed as achieving greater success in teaching as demonstrated by pupil gain ($t=3.0$, $p > .01$). Also, those students who negotiated agreement on criteria for successful pupil gain tended to be viewed by their supervisors as more successful in the principles of learning (71% As to 59% As in the respective experimental and control groups).

Justiz (1968), who studied under McNeil at the University of California, Los Angeles, designed a doctoral study which confirmed that teaching ability could reliably be measured by student achievement. Two high school placements were utilized in his study with 10 student teachers assigned to the first school and 7 student teachers at the other. The teacher trainees were charged with teaching unfamiliar subjects in a realistic teaching situation, independent of their supervising teachers. Each student teacher was given a packet containing the two subjects in "kit" form which supplied an objective, related subject matter, and practice exercises. Each trainee in the first high school selected 18 experimental pupils from his supervising teacher's class which were deposited, two at a time, in 9 adjourning classrooms in a testing area. Each student teacher was then assigned to the one classroom which did not contain any of his own students. All trainees instructed for 30 minutes in the first subject and after completion were allowed, under close supervision, to administer a 15-minute paper and pencil posttest. All pupils were then reconstituted a second time, and the same 45-minute procedure was used for the second subject area.

In the second school, the identical procedure was used with 7 student teachers, and 12 pupils per class, but without the second reconstitution of pupils. Each student teacher was ranked according to the mean score of his class. Two rankings were then correlated, one for each subject field, to determine the reliability of the student teacher rankings, using Spearman Rank-Difference Correlation. The correlations were significant to the .05 level at both schools. Justiz concluded that it was the first

experiment to reliably measure general teaching ability for student teachers. It would be interesting to see this study replicated with a larger sample size, teacher trainees assigned to elementary schools, and inservice teachers.

Moffett (1966) also did a study comparing the performance of student teachers evaluated on the basis of attaining agreed-upon educational objectives with that of student teachers evaluated by means of a rating scale. Utilizing a pretest-posttest procedure, he found that pupils in the experimental group did better on their posttest of geography skills, the experimental group student teachers were more satisfied with their supervisory help and midterm grades, and they expressed preference for having their teaching performance evaluated on pupil achievement rather than a rating scale measure.

An intriguing investigation was conducted by Popham (1971) exploring pupil progress and instruction by objectives to ascertain teacher performance. He employed two groups of instructors, one comprised of professionally trained and experienced secondary teachers, and one of non-teachers including college students, garage mechanics, television repairmen, and electronics workers. The investigator constructed an instructional packet for these subjects, which included specific performance objectives to be measured by a posttest, and a set of resource materials which could be used by the teachers in planning the instructional sequence to accomplish the objectives. The two types of instructors were randomly assigned students, and taught short units of instruction based upon planning from the provided instructional objectives and resource material.

The results revealed that the experienced teachers did not markedly outperform the nonexperienced on the measure of pupil growth on objectives. Popham concluded that "experienced teachers are not particularly skilled in bringing about specific behavior changes in learners" (Popham, 1971, p. 602).

Morine (1976) substantiates the concept that more precise planning of instruction contributes to student growth. She found that teachers with low to average pupil gain in math made more general statements in writing their math lesson plans. Conversely, the tendency of teachers with high pupil gain scores in math was to attend more to cognitive aspects of the lesson in planning daily instruction.

A case for student achievement tied to the use of criterion-referenced measures in math was reported by Stow (1979). She states that the use of performance objectives in all the fourth-grade math classes in the West Des Moines, Iowa schools, resulted in impressive performance on the criterion-referenced tests as well as on the Iowa Test of Basic Skills, where the students in the district made 22 months' gain in one year of study.

Briggs (1977) supports the contention that performance objectives have value. He believes that performance objectives help teachers to plan their instruction, to evaluate student performance, and to improve communication with students regarding instructional intent. Performance objectives also aid teachers in their communication with other teachers, curriculum planners, administrators, and parents. Efforts to justify new and current courses, and to defend the legitimacy of existing courses

and curriculum, are also facilitated through the use of performance objectives.

While Briggs (1977) finds performance objectives to be desirable, he does not feel that they conclusively have been shown to promote pupil achievement. Whereas objectives "have been shown to facilitate learning in a limited number of studies, such a facilitating effect has not been consistent across the vast majority of investigations" (Briggs, 1977, p. 87). He comments that, based upon his review of literature, he must conclude that studies of the effects of teachers utilizing objectives suggest that it makes little difference to student achievement whether teachers have objectives.

It is confounding to note that one of the studies Briggs cites as producing inconclusive evidence about student gain using performance objectives is one conducted by McNeil in 1967. This investigation was reported by McNeil as providing evidence to support pupil achievement through the use of instructional objectives. He reported his findings saying:

> Pupils taught by teachers in the experimental group (emphasis on results) achieved more than those taught by teachers in the control group (emphasis on procedures) with respect to their previously identified area of deficiency. More than that, pupils whose teachers were told to emphasize results tended to show greater achievement than their matched peers in the remaining types of punctuation skills as measured by the standard test. (McNeil, 1967, p. 71)

Supervision and Evaluation of Teacher Effectiveness

There is virtual agreement among educators that teacher evaluation has two major functions: (1) to improve instructional and, (2) to provide a basis for making personnel decisions (Oliva, 1976; Redfern, 1963; Robinson, 1978; Thomas, 1979). More specifically, these related but different practices would include evaluation to provide data for the improvement of the individual teacher's skills, and for the intent of gathering information to assist in administrative decisions on hiring, firing, promotion, transfer, tenure, or reallocation of resources (Brighton, 1965).

Supervisory practices

These determinations are frequently made solely or partially upon the assessment made following a classroom visitation by the supervisor. Robinson (1978) reports that in his survey, 82 percent of the principals and 88 percent of the supervisors usually had a conference with the teacher following the visit. The previsitation conference was reported as not widely in use and many of the supervisors stated that they did not prepare themselves before observing a class "by either reviewing lesson plans, having a conference with the teacher, or reviewing previous observation reports" (Robinson, 1978, pp. 24-25). This procedure for evaluation of instruction is unacceptable to many educators and has provided motivation for improved supervisory practices.

The preobservational conference

Professor Robert H. Anderson at Harvard has pioneered the development of a program aimed at improvement of supervision. His model employs a four-phase cycle for supervising teachers which is: (1) the preobservational conference, (2) observation, (3) analysis and strategy, and (4) the postobservational conference.

The implementation of the preobservation, or preinstructional, conference has been gaining favor among supervisors and administrators. McNeil (1972) supports the concept of the preinstructional conference when he states, "Supervisors are not only concerned with finding out what the pupils have achieved, they want to know that the results are desirable. The preinstructional conference is a mechanism by which the supervisor helps teachers clarify and justify instructional intents before instruction commences" (McNeil, 1972, p. 4). He also affirms that, "Without instructional objectives, it is difficult to assess whether teachers have or have not justifiable intents" (p. 4), and the preinstructional conferences using instructional objectives is likely to reveal those teachers who do not have a sense of direction or are teaching to trivial ends. Identifying such teachers in the preinstructional conference is a "first step in the direction of accountability by supervisors" (p. 5).

Models of supervision

The preinstructional conference, as used to clarify instructional intent, is incorporated into a few supervision paradigms, perhaps one of the better known being clinical supervision (Denham, 1977). Clinical supervision is a concept described by Cogan (1973) and Goldhammer (1969).

This supervisory technique utilizes a three-step procedure including: (1) preconference, (2) observation, and (3) postconference.

Clinical supervision has been defined as a procedure for observing in the clinic of the classroom. Here supervision is "up close" so that direct feedback can be provided to the teacher. It is also a tenent of the concept that improvement of instruction can only come from those aspects of the teaching which are of concern to that teacher, and not of items on an evaluation form or a supervisor's pet concern (Reavis, 1976). Cogan further defines clinical supervision as a "rationale and practice designed to improve the teacher's classroom performance. It takes its principal data from the events in the classroom" (Cogan, 1973, p. 9). This notion does not preclude the desirability for preinstructional input in this process; Cogan also advocates the supervisor's planning with the teacher before instruction. He suggests that these "plans commonly include specification of outcomes, anticipated problems of instruction, materials and strategies of teaching, processes of learning, and provisions for feedback and evaluation" (p. 14).

Another supervisory model utilizing the preinstructional conference to determine educational outcomes is supervision by objectives (SBO) as conceived by Lucio and McNeil. This concept of supervision is more structured and precise than clinical supervision and places greater emphasis on student product. They define supervision by objectives as shifting the evaluation of teachers from how they teach and from their particular characteristics to the results teachers obtain with learners. The advantage cited for this approach is that, "once teachers know that their

evaluation does not rest upon the opinion of an administrator, but upon the extent to which they achieve the instructional objectives they themselves agreed upon, they voluntarily make more fundamental improvements in instruction" (Lucio and McNeil, 1979, pp. 106-107).

Within the SBO model, the preobservation conference calls upon the teacher to present to the supervisor the instructional objectives for the lesson. In addition, the teacher indicates the measurement (tests, product, and situations) by which the pupils will demonstrate acquisition of specified outcomes. The teacher also establishes what will be accepted as a satisfactory level of performance on the measuring instrument. Following this designation, the teacher justifies the selection of objectives to the supervisor and agrees upon attainability, appropriateness of objectives, criteria measures, and teacher criteria. The preobservation conference also provides the opportunity to clarify the supervisor's role in the process (Lucio and McNeil, 1979; McNeil, 1971).

The use of "contract plans" as outlined by McNeil and Popham is an alternative for assessing teacher competence. Contract plans are based upon student gain and rest on the premise that the ends of instruction must be agreed upon before teacher competency can be assessed. The technique involves the development of objectives for the pupil, and the supervisor and the teacher agree in advance what they will accept as evidence that the teacher has been successful in changing the skills, competence, or attitude of the pupils. The agreement drawn up before the teacher instructs, is "designed to counter the prevailing practice of trying to make an ex post facto judgment about the desirability of ends" (McNeil

and Popham, 1977, p. 270). The strength of the contract plan is that it allows teachers to establish outcomes and standards that are deemed most appropriate for the lesson, and instead of comparing teachers on the basis of normative criteria, this plan permits teachers to serve as their own control.

Haefele (1980) has summarized a dozen approaches to teacher evaluation. They are:

1. Teacher competence is measured by performance of the teacher's classes on standardized tests given at the end of the year. Year end performance is compared with established norms.

2. Standardized tests are administered to students to determine how much they increase their learning over time. The amount of desired gain is established in advance by school personnel, teachers, and an independent evaluator.

3. Students in each grade or subject-matter area are tested at the beginning and end of each semester or school year. Gain scores are computed to contrast class performance (gain or loss) with classes of comparable ability. Teacher effectiveness is measured by proportion of "gainers" to "losers."

4. Informal observations and ratings of the teacher are conducted by the principal and/or other supervisory personnel. Comments by students, parents, and colleagues are incorporated in the final evaluation.

5. Systematic observation of the teacher is conducted by the principal and/or supervisor, using a rating form that lists characteristics of good teachers. The teacher's evaluation score is compared to a school or district standard.

6. The teacher is systematically observed and rated by peers on the extent to which he exhibits important characteristics of good teaching. A predetermined school or district standard is the criterion.

7. The teacher's students use a rating form to judge the extent to which the teacher exhibits important characteristics of good teaching. The teacher must meet a predetermined school or district standard of effectiveness.

8. Teachers are required to take the National Teacher Examination (NTE) and achieve a predetermined standard composite score.

9. Periodically, the teacher is provided with an instructional objective, a sample test item measuring that objective, and information about the content it covers. A small group of students is assigned to that teacher randomly (to balance abilities) and is instructed by the teacher on the objective for one to 10 lessons. After instruction, the students are tested on the objective. Teacher effectiveness is determined on the basis of how well the students achieved the objective.

10. The Teacher Perceiver Interview is administered to teachers. Teacher effectiveness is based on how well the teacher meets a predetermined criterion or norm-referenced score.

11. The teacher is given written descriptions and/or shown films of typical classroom problems. The teacher's effectiveness is judged on the basis of answer quality.

12. The teacher, together with the principal and/or curriculum supervisor, establishes mutually agreed-upon (negotiated) instructional goals and objectives for the year. Observation data and other sources of information gathered at regular intervals during the year are used to monitor and evaluate the attainment of goals. (Haefele, 1980, pp. 349-352)

The final method described by Haefele is global, and approaches evaluation and supervision via mutual goal setting of objectives during a preinstructional meeting. No particular name is given to this system but the process prescribes a meeting between the teacher and supervisor where instructional objectives are negotiated and agreed upon recognizing that teachers also have individual differences. A criterion for success is established, and after observation, a feedback loop is utilized for the teacher's self-improvement. Administrative support throughout helps teachers achieve their objectives.

Haefele states that the notable strengths of this supervision model are that "(1) teachers and administrators work together for the benefit of the students, (2) the goal is improvement, and (3) teacher self-

evaluation is an important feature of this system" (Haefele, 1980, p. 352).

The Teacher Evaluator's Capacity to Evaluate

The question of who is to be charged with the responsibility of teacher evaluation must be defined. The task of supervision, for improvement of instruction or personnel decisions, may fall to persons with varying titles. Supervisory personnel, those involved in the evaluation of teachers, in the central office of urban school districts are usually the assistant superintendent, director, supervisor, evaluator, coordinator, or consultant. These positions are not clear-cut, e.g., and administrator in one school system may be called a director, and a person performing the same job in another school may be called a supervisor (Lucio and McNeil, 1979). Levin (1979) states that most commonly the principal is the one doing teacher evaluation, and Tuckman, Steber, and Hyman (1979) agree with this assessment of the principal's functioning. Lucio (1969) believes that supervisors are all persons whose unique or primary concern is instructional leadership.

It is assumed in the assignment of instructional leadership to these professionals that they are well-equipped and competent to judge good teaching. Worcester (1968) notes that it has been accepted that a person with a certain title, i.e., supervisor, superintendent or principal, is capable to assess good instruction and points out the implication that this is a general ability requiring little or no training. He sardonically affirms:

> Many principals give little time to systematically observing teaching procedures, examining testing methods, and otherwise really evaluating the results of teacher's activities. The judgment of raters are frequently based upon personal attractiveness, willingness of the teacher to participate in extracurricular activities, her presence in town over the weekend, the frequency with which pupils are sent to the office for discipline, but not on the evidence supported by tests of pupil progress. (Worcester, 1968, p. 124)

Given to reliance on rather subjective criteria of who is an effective teacher, administrators may lack agreement upon that determination. Tuckman, Steber, and Hyman (1979) conducted a study which suggests that elementary principals are more likely to perceive a teacher as effective if she is warm and accepting. Intermediate principals are influenced by creative teachers, and high school principals favor dynamic teachers as the more effective. Since teachers receive their supervision and evaluation from building principals, such perceptions could hold consequences for all instructional personnel.

Rater effect

Regardless of an appraiser's preference for personality types, care with which instructional objectives are studied for curriculum congruence, attention upon pupil gain in the class, or scores considered from the National Teacher Examination, the teacher evaluator must gather some sort of evidence to support appraisal judgments. One of the most widely used methods for obtaining data is the classroom observation.

McIntyre (1980) contends that there is a body of research studying rater effect which should be of concern to teacher evaluators utilizing classroom observations in the teacher assessment process. McIntyre believes that the presence of an observer in the classroom (for the purpose

of teaching evaluation) affects the teacher's usual teacher behavior, and data collected during such an observation is not truly representative of the teacher's usual functioning, but only a sample of how that teacher performs when being observed.

Samph (1968) conducted a study with 10 female teachers as the subjects of observation. The analysis of data demonstrated the existence of an observer effect, and the teachers became more "indirect" when an observer was present whether they were informed of the visit or not. The teachers in this study used more praise and acceptance of student ideas and less criticism of students during the duration of the observation period.

In 1974, Ragosta collected data using two observation instruments: the Florida Climate and Control System (FLACCS), and the Teacher Practices Observation Record (TPOR). Data from four series of observations on 289 Follow Through classes were collected. The results showed the presence of an observer in a classroom resulted in positive teacher affective behaviors and in higher ratings of classroom attitude. Conversely, the absence of an observer resulted in less positive teacher affective behaviors and in lower ratings of classroom attitude.

A study by Di Martino (1974) was undertaken with 24 secondary school teachers in DeKalb County, Georgia, to determine rater effect in classroom climate utilizing the Reciprocal Category System. The data indicated that the presence of the observer influenced the teacher's verbal behavior to become warm and less cool toward students.

The attendance of an observer may be expected to produce self-conscious reactions from teachers and students, but this may be less serious

than charged. It's been reported that most much-observed subjects quite readily become accustomed to the observer's presence and the passage of time brings about an accommodation to the observer's presence and produces a tranquilizing effect on the observed (Jersild and Meigs, 1939). Indeed, Masling and Stern (1969) did a correlational study comparing units of observation, and they, too, determined that the effect of the observer gradually diminishes over time.

Gage (1963), however, relates "Observation by expert judges of teaching probably cannot be used for administration appraisals. Observers are hard enough to ignore when they are friends or researchers, whose impressions will not affect one's standing. But when the teacher knows he is being looked over by someone whose opinion will determine his promotion or salary, his performance may depend more on his nerve than on his teaching skill" (p. 172).

Keeping all this in mind, McIntyre (1980) recommends that, because observations may not provide evidence of the teacher's "true" behavior, teacher evaluation programs should not be based solely on classroom observation. He suggests that the appraisal system that emphasizes classroom observation should require a series of observations over a length of time, and that to reduce any aura of threat, administrators work to interact with faculty to develop a trusting, solid relationship. These efforts would work to improve the problems encountered with rater effect in the collection of data via classroom observation.

Rater error

Evaluation of personnel performance is much the same for the private sector as it is for the public sector. For a supervisor to be competent in the assessment of teacher effectiveness, he must engage in the observation of behavior and, consequently, observer accuracy becomes a factor. The same is true in the evaluation of personnel in business and industry, and Spool (1978) says that "for the observer to be more accurate in observing, he or she must, among other things, be able to use the observer system with ease and be able to make observations in accordance with some standard of criteria" (p. 854).

According to Borman (1975), performance ratings are almost inevitably contaminated by rater errors which he specifies as leniency error and halo error. Latham, Wexley, and Pursell (1975) concur with the problem of rater error caused by halo effect, but cite additional rating errors of similarity, or similar-to-me effect, contact effect caused by subsequent events, and first impressions.

Researchers in the private sector have utilized training programs to reduce rater error on observation and rating scales used for evaluation of personnel. Although behavior scaling strategies seem conceptually to offer promise for more reliable and error-free ratings, Borman and Dunnette (1975) note:

> Getting raters to observe work-related behaviors more systematically and representatively is a potentially fruitful approach. Raters are seldom skilled in making systematic work-related behavior observations. They need to become adept at observing and recording relevant job behaviors so that they may be better equipped with the information necessary for making accurate evaluations of employee performance. (p. 563)

Borman and Dunnette (1975) go on to state that if raters can first be trained to observe work-related behavior more competently, and second to use the scales more accurately, it is possible that more error-free portrayals of performance can be made.

Borman (1975) reports on training as a method to reduce rater error and found that brief training sessions of 5 or 6 minutes in which raters are introduced to the halo error, and urged not to commit this error, significantly reduced its occurrence. He also found that training to discriminate greater distinctions among performance categories may lead to a reduction in interrater reliability.

Bernardin (1978) investigated the diminishing effects (over time) of a comprehensive versus an abbreviated training for raters on leniency and halo errors. He found there was no significant difference in the two groups (one trained for 5 minutes and the other for one hour) when four months later they were retested for leniency and halo errors in their ratings. The investigator noted that another confounding variable was intelligence which was related to the rater's ability to do error reduced rating (p. 307).

Latham et al. (1975) reported that only an intensive training workshop was effective in training raters. They prescribed sessions giving subjects a chance to practice observing and rating videotaped performances with immediate feedback regarding the accuracy of their ratings. Utilizing this approach, it was found that after 6 months the trained raters were still able to function as error-free raters.

The evaluator's feelings of inadequacy

The reality remains that, while supervisors and administrators are responsible for the assessment of teachers and their instruction, many feel ill-prepared for the task. Kowalski (1979) reports that research substantiates the difficulty principals feel in striving to achieve the goal of leadership in the functional and theoretical aspects of curriculum, thus making the assessment of objectives a frustrating experience for them.

Professional training does not appear to be preparing administrators to competently perform in this area of their job responsibility. Of the principals surveyed by Kowalski (1979), 79 percent perceived their course work in curriculum to be insufficient to meet their assessment needs, and 93 percent believed their experiences in curriculum were inadequate. These findings indicate a high degree of insecurity among principals performing in the area of instructional leadership.

Brighton (1965) observes that the principal often fears lack of respect because his role as evaluator places upon him certain expectations from the faculty. He must cope with the realization that the assessment process may force him into positions where his own lack of knowledge will become obvious. "He may fear such disclosures will diminish the respect he must have from his staff if he is to serve as their administrative officer" (p. 21).

These feelings of inadequacy may incapacitate an administrator in the role of teacher evaluation. It has been stated that principals often do not know what action to prescribe to a teacher to improve instruction.

Rather than reveal their own lack of expertise, the principal will feign approval of the teacher's instruction and employ a type of ritualism that tends to be, not critical, but lauditory in tone. This form of escapism is called "ceremonial congratulations" by Guthrie and Willower (1973).

These kinds of findings can hardly be comforting when it has been demonstrated that better supervisors are tougher raters. Kirchner and Reisberg (1962) have found that less-effective supervisors are more lenient, particularly with poorer subordinates. The less-effective supervisors tend to rate all subordinates much more alike (ceremonial congratulations) than do better supervisors. Better supervisors are more discriminating in their rating and show more "spread" and variation in their ratings. "Since a basic objective of rating is to differentiate performance, the better supervisors are doing the more effective job" (p. 299).

Summary

In summarizing the literature, the review noted the "who", "why", and "how" of teacher evaluation in the United States. Looking at the "who", it was seen that the person first charged with the appraisal of teachers and quality of education began as the duty of laymen in the community and usually included some member of the clergy to guarantee the school's teachings were consistent with prevailing religious philosophy. The responsibility for assessment shifted to school administrators, namely the superintendents, who were in the employ of school boards and answered to their directives. When the position of the school principal emerged, teacher supervision and evaluation more regularly fell to the person

functioning in that role, and as the occurrence of the specialist became more prevalent in education, there emerged new developments for the evaluator and the evaluated. The assessment of these specialists caused many teacher appraisers to feel inadequate about evaluating areas of expertise in which they lacked knowledge, yet on the other hand, it also brought into being a resource pool from which administration could borrow to evaluate specialists in the same field. Currently, we find that the building principal is most generally called upon to do teacher evaluation, although the superintendent, supervisor, or specialist may also be making major contributions in these determinations.

The rationale for teacher appraisal, the "why" of evaluation, has changed over time beginning with a belief that inspection and control of teachers were required to make certain the ongoing instructional process reflected the values and standards of the community. Yet, with the rise of scientific management, a focus on the ends of instruction produced a more child-centered approach to teacher evaluation. It became important that teachers were imparting academic knowledge to students and the quantification of this process was pursued. Product, measurement, and testing were emphasized to ascertain student change, and rating scales came into vogue for measuring teacher effectiveness. These measurement techniques have aided educational administrators in working for the improvement of instruction as well as evaluation to assist in personnel decisions of hiring, firing, and promotion. Today these practices may be questioned as to their effectiveness, with the public serving as a severe critic, but it is a calculated move to meet the demand for accountability

in education.

How teacher evaluation should be conducted is open to conjecture; however, all would agree that decisions should be based upon evidence and this evidence should be the concern of appraisers. It is to this end, that of evaluating the evidence prescribed in the teacher's planning, that this investigation is aimed.

While there is universal acceptance of the need for teachers to do appropriate planning and decision-making prior to good instruction, this aspect of teaching has received relatively little notice within the evaluation process. Lesson plans and the associated instructional materials, which are the products of the teacher's preparatory planning, have not been widely used as evidence of teacher proficiency.

The present study will address the problem by investigating the supervisor's ability to accurately evaluate the lesson plans, including objectives, media, worksheets, and evaluation techniques, and to ascertain if supervisors more accurately evaluating preinstructional materials will also more accurately rate instructional classroom actions. The study may lead to conclusions about where best to provide assistance and training to supervisors and administrators in their teacher assessment responsibilities. Thomas (1979) supports such endeavors when he affirms:

> In the future it may be possible to develop more sophisticated methods for performance evaluation. We look to Colleges, Universities, and Research Agencies to help us to do this. Evaluation has been and will continue to be a complex area of human activity. It is not however, beyond solution. (p. 7)

CHAPTER III. METHODS

The purposes of this investigation were: 1) to determine if educational administrators and supervisors had the ability during the teacher appraisal process to accurately assess preinstructional materials, 2) to ascertain if utilization of evaluative criteria featuring low or high inference descriptors on the evaluation instrument facilitated appraisers to make increasingly more accurate ratings, 3) to learn the influence which preinstructional materials had upon the teacher appraiser's assessment of the teacher's classroom performance, and 4) to find out if appraisers who were more skillful at assessing the preinstructional materials were also more skillful at evaluating teacher performance. Additional functions of this study were: 5) to ascertain if previous training in teacher assessment influenced the educational administrator's skill in teacher appraisal, 6) to determine if regional differences in perception of preinstructional materials and teacher performance were evident in the sampled populations, and finally, 7) to learn if appraisers, in varying job positions of educational administration, evaluated more or less accurately.

Population

The population for this study was comprised of 529 subjects, all involved in some form of professional improvement program. Thirty-two of the participants were involved in an Iowa AEA 14 (Area Educational Agency) workshop on teacher performance evaluation. Other

subjects attending teacher performance evaluation workshops and participating in this experiment from outside of Iowa were the following: 63 from Philadelphia, Pennsylvania, 107 from Portage, Indiana, 66 from Oyster Bay, New York, 48 from Detroit, Michigan, 28 from Novato, California, 112 from Toronto, Canada, and 73 from Albuquerque, New Mexico. The eight geographical locations sampled were diverse and widespread including the western seaboard (California), the eastern seaboard (New York), the desert southwest (New Mexico), and the industrial midwest (Michigan) and east (Pennsylvania), as well as the great plains state of Iowa. Also represented in the sample was a foreign country, our northern neighbor, Canada. These locations represented industrial and agricultural business centers and typified rural, urban, and suburban areas.

The persons sampled were superintendents, principals and supervisors. A category of "others" included college teachers, coordinators, assistant principals, teachers, consultants, and specialists. The training these people had in teacher performance appraisal varied from none to one day or less, one to three days, three to five days, more than five days, and input from a college or university course.

All data were collected within a short period of time, April 17, 1980 to May 20, 1980. These dates, their locations, and numbers of participants are recorded in Table 1.

Table 1. Population, dates, and locations of data-gathering experiment

Location	Data	Number requesting exclusion from the study	Number included in the study
Philadelphia, Pennsylvania	April 17, 1980	3	63
Detroit, Michigan	April 23, 1980	6	48
Ontario, Canada	May 5, 1980	5	112
Portage, Indiana	May 6, 1980	10	107
Oyster Bay, New York	May 8, 1980	7	66
Novato, California	May 9, 1980	2	28
Creston, Iowa (AEA 14)	May 13, 1980	4	32
Albuquerque, New Mexico	May 20, 1980	6	73
Total		43	529

The Instruments

The materials used in the study, the preinstructional materials, the videotape instructional scene, and both of the evaluative criteria instruments were all taken from the Georgia Teacher Assessment Project (TAP). These materials were developed to be used in the training of teacher

appraisers under a Title IV-C grant and were in the public domain, i.e., not copyrighted. In a telephone communication to TAP personnel, it was determined that the materials have not been widely used in subsequent studies of teacher assessment with the exception of one dissertation now under investigation on the training of evaluators in the rating process (Anderson, Note 2).

The Georgia Teacher Assessment Project was a joint project between the University of Georgia (Athens) and the Georgia Department of Education which began in April 1976, to initiate a performance-based student teacher/beginning teacher certification program (Johnson et al., 1978).

The TAP materials feature teachers' authentic preinstructional materials, which include lesson plans, instructional objectives, worksheets, and evaluation measures that the first-year teachers participating in the research taught in their classes. This was followed by the taping of the accompanying actual instructional lessons. The Project collected a wide variety of preinstructional materials and teaching sessions from many grade levels (K to 9) and from numerous subject areas including Social Studies, Mathematics, Science, Reading, and Language Arts. The results of this collection provided real instructional samples from which appraisers might learn and improve assessment skills.

Teacher Performance Assessment Instruments (TPAI)

TAP conducted a study which sought to verify generic teacher competencies believed to be essential for teachers regardless of subject or level. Competency statements from numerous sources were classified, combined, and edited. This resulted in a survey instrument containing 52 competency statements. Stratified sampling was employed with a

population of teachers, teacher educators, and administrative personnel in which the total number of respondents was 4,668 (Teacher Assessment Project, Note 3). Five subgroups of teaching specialties were identified and all agreed as to the most important competencies and ranked them in approximately the same order of priority. The five subgroups were defined as:

1. Teaching Plans and Materials (TPM)
2. Classroom Procedures (CP)
3. Interpersonal Skills (IS)
4. Professional Standards (PS)
5. Student Perceptions (SP)

Instruments (TPAI) were developed for measurement of the competencies for each of these five subgroups. The instrument focusing on subgroup one (TPM) relates to the preparation of instruction, planning of instruction, selection of objectives, and choosing materials and equipment. In the present investigation, this component is referred to as "preinstructional materials" and the accompanying evaluative criteria instrument is used (see Appendix A).

The second subgroup (CP) is also germane to this study and utilizes another instrument which is concerned with actual classroom practice and seeks to ascertain how effective are the teacher's practices, methods, and techniques in the classroom. In the present research, this subgroup topic was evidenced in the videotaped session and subsequent evaluative criteria included in the component (see Appendix B).

The remaining three subgroups, Interpersonal Skills, Professional Standard, and Student Perceptions, were not dealt with in this investigation.

The Georgia Teacher Assessment Project reported the reliability of their instruments saying "investigations conducted via videotape and other standardized means have shown reliability estimates in the range of .70 to .90 in most instances" (Teacher Assessment Project, Note 3).

Validity of the TPAI instruments was established in a variety of ways. Content validity was determined through verification studies utilizing experts' opinions. Concurrent validity was determined by comparing teachers with various years teaching experience, and finding that more experienced teachers tended to score higher on the instruments. Finally, construct validity was examined through factor analysis of competencies, and indicators, correlating TPAI assessments with the Purdue Observation Rating Scale (Teacher Assessment Project, Note 3).

All TAP competency statements were refined into items to be used to assess the competencies. These items were articulated as evaluative criteria, and generally more than one item or criteria was used to assess any single competency.

This study concerned itself only with those competencies and associated criteria which accompanied the Teacher Plans and Materials and Classroom Procedures.

For the purpose of this study, nine criteria were selected to measure the preinstructional materials, and eleven criteria were chosen to measure the classroom procedures featured in the videotaped sequence. All criteria were restated into question format for the study (see Appendices A and B).

The TAP program supplied a scoring rationale for each of these criteria (see Appendix C). A panel of 40 persons: 20 classroom teachers, 12 administrators, and 8 college faculty members provided ratings to be

used as a measure determining rating proficiency (Blake, Note 4). The suggested ratings from the panel of experts were used in this study as the standard for accuracy by which rater proficiency was measured.

Research Design

An experimental technique of posttest-only control group design was used in this study. The steps in this particular approach are as follows: 1) the subjects are randomly assigned to experimental and control groups, 2) the treatments are administered to the experimental groups but not the control groups, and 3) the posttest is administered to both groups (Borg and Gall, 1976; Van Dalen, 1966).

Random assignment was accomplished in this investigation by arranging experimental kits in such a fashion that each of the four treatments was sequentially distributed to members. In this manner, every fifth person had the same packet. The kits were unobtrusively numbered on the corner of the envelope so that sequence was easily monitored, and an approximately equal number of each treatment kit was given out. This randomization procedure was utilized to eliminate initial differences between control and experimental group members without the administration of a pretest procedure. The relatively large number of subjects participating in this study made random assignment a specially effective method while eliminating pretest examination.

Because a randomized control group was used and no pretest was given, this design controls, but does not measure, the main effects of history, maturation, and pretesting effect. This makes the design desirable.

There are disadvantages, however, because some types of analyses and statistical tests cannot be utilized without pretest information, and these techniques produce more powerful findings.

After the subjects of this study were randomly assigned to groups, and the experimental groups were exposed to the treatment, and posttest results were gathered to determine the effect of the treatment, these data were subjected to the appropriate tests of significance to determine if the differences were greater than might have occurred by chance.

Data processing was conducted at the Iowa State Computer Center using the Statistical Package for the Social Sciences wherein the eleven hypotheses were subjected to varying statistical treatments. The first three hypothetical questions were submitted to examination by a one-way analysis of variance with contrasts for testing each individual hypothesis. The fourth hypothesis was also given a one-way analysis of variance treatment, and hypothesis 5 was given the test for comparison or correlation coefficient analysis. The remaining questions, hypotheses 6, 7, 8, 9, 10, and 11 were all examined by Crosstabs for Chi-square analysis to determine if the variables were independent of each other.

Procedures

The format for this experiment was a simulated exercise in teacher performance appraisal. The simulation, which was conducted as a portion of a workshop offered on teacher evaluation, involved the participants in the study and evaluation of a teacher's preinstructional planning, and viewing and evaluation of a videotaped instructional sequence on language

arts for intermediate grade youngsters. These components, namely Packet One and Packet Two, were made available to the subjects via a kit enclosed in a large manila envelope. Packet One envelopes were inconspicuously numbered one through four, and were made up of four different treatments; these were: 1) a journal article discussing teacher assessment procedures (see Appendix D), 2) the preinstructional materials for the lesson about to be viewed on the videotape (see Appendix E), 3) the same preinstructional materials as supplied to group two as well as a high inference rating instrument by which the materials were evaluated (see Appendix F), and finally, 4) the exact preinstructional materials as given to groups two and three including the evaluative scale which, in this case, utilized low inference items on the criteria scale (see Appendix G). A rating form was included for groups three and four of Packet One as responses were required on the evaluative criteria instrument (see Appendix H).

Packet Two envelopes all contained the same contents, an eleven-item evaluative criteria for assessing the videotaped class (see Appendix I), and a rating form for recording responses and supplying personal information regarding job title, amount of teacher evaluation training, and a request form giving them the option to have their responses withheld from the study (see Appendix I).

The exercise was introduced by a presenter who briefly explained the nature of the simulation, the kinds of materials and activities to be used, and the type of interaction requested from the audience. After the kits were distributed, twenty minutes were allotted for studying and recording evaluative responses to the contents of Packet One. When Packet One was returned to the kit envelope, participants were instructed to take

out Packet Two and for five minutes study the eleven criteria to be used to evaluate the taped instructional session.

The subjects then viewed the taped classroom sequence which lasted for twenty minutes, and finally, recorded their responses to the criteria on the rating form that had been supplied in Packet Two. The rating form was made up of two attached sheets, one white and the other yellow, of pressure sensitized paper. This allowed for responses recorded on the first sheet (white) to be duplicated on the second (yellow) sheet of paper. In this manner, when ratings were completed the participants could tear off the yellow sheet and retain this copy of their responses. The white copy and the evaluative criteria instrument were returned to the kit envelope, and with Packet One, were returned to the leader of the simulation activity. The leader then led a feedback session and participants, referring back to their responses recorded on the yellow rating sheet, were able to review their ratings and compare them with the normed responses of other groups of trainees. Time was then allowed for open discussion and comment from the subjects regarding these rating scores. This entire sequence required approximately one and one-half hours of the subjects' time.

A few unsolicited written responses from subjects are interesting to report. These comments were always anonymous and usually were written on margins of the rating forms or the blank paper included for narrative responses from treatment groups one and two. Typical reactions from subjects in treatment group one (those reading a journal article and no preview of preinstructional materials) were feelings of inadequacy to

respond and make judgments about teacher performance without knowing the teacher's objective. These subjects were typified by the following quote, "couldn't do effective job because didn't know objectives (Albuquerque subject)." Again, some of treatment one group members related they enjoyed reading the journal article and one commented by stating, "These materials would be helpful as part of a preconference (Oyster Bay subject)."

There were some replies from the Indiana subjects regarding the audio system for the videotape sequence. One said, "I have answered only those (on rating form) I could honestly make a judgment. Sound was poor." Another related, "I was unable to comprehend the video accurately," and left the rating form blank.

Some participants requested their responses not be included in the research project. Various reasons were cited, and again, an Indiana subject said hearing the sound took away accuracy, another opted out, "Because my job specifies that duties do not include evaluation of performance (Albuquerque subject)." And finally, a terse statement, "The film was a waste of my time. The teacher was totally incomprehensible. I resent being used for your profit from research (Indiana subject)."

An Oyster Bay subject commented upon the black male teacher featured on the tape saying, "There are very few black educators in the schools represented here today. Discount this if you wish, but I believe that there are latent prejudices among our group. This should be taken into consideration during evaluation." Another opinion regarding the teacher came from a Canadian who said, "If this fellow was teaching on our staff,

I'd have a full time job trying to teach him how to teach."

A final response to be noted came from a Philadelphia participant commenting on the teacher's preinstructional planning. This person said, "A better than averege grade teachers [sic] lesson plans."

These many unsolicited remarks were informational to review. The most useful feedback was in the way of comment on the poor audio system for the videotape. Subsequently, precautions were taken to prevent the reoccurrence of the problem by careful testing of the audio system before beginning a presentation.

CHAPTER IV. ANALYSIS OF DATA AND FINDINGS

Analysis of Data

The data for this experiment were generated from the manipulation of numerous variables. These variables included: 1) four treatment groups, 2) eleven scores from rating teacher performance, 3) nine scores from rating preinstructional materials, 4) accuracy of these scores, 5) high and low inference items on the evaluative instruments, 6) amount of training in the teacher appraisal process, 7) the evaluator's job description, and 8) the geographical regions from which the scores were taken. More specifically, these variables can be described as follows:

1. Four treatment groups - These groups were given four differing interactions previous to observing the classroom instruction. These interactions were:
   a. Read a journal article on teacher evaluation (Placebo).
   b. Studied lesson plans that accompanied the lesson (Lesson plans only).
   c. Studied lesson plans and rated them on nine criteria with a high inference rating instrument (Lesson plans and high inference).
   d. Studied lesson plans and rated them on nine criteria with a low inference rating instrument (Lesson plans and low inference).

For frequency distribution of treatment groups see Appendix J,

Table J.1 and for standard deviations, see Appendix J, Table J.2.

2. Eleven scores from rating teacher performance - These scores were derived from ratings on the instrument which measured the teacher's performance in the classroom. For composite scores, see Appendix J, Table J.3.

3. Nine scores from rating preinstructional materials - These scores were acquired from ratings on the instrument which measured the teacher's preinstructional materials. For composite scores, see Appendix J, Table J.4.

4. Accuracy of scores - Proficiency on rating of teacher performance and preinstructional materials was measured against accurate scores as given by the TAP scoring rationale (see Appendix C). Accuracy on rating instructional materials was coded as AIM, and accuracy on teacher performance was coded as ATP. In the statistical analysis of this study, an accurate score was recorded as 0, one score higher or lower than the indicated accurate score was given a 1, and a rating score two higher or lower than the accurate score was given a 2. Direction from 0 was not determined. A five-point scale was used continuing from low (1) to high (5) ratings. It should be noted that in some instances, TAP accepted either of two adjacent scores as accurate/correct, and in these cases, a midpoint score was programmed to measure accuracy.

5. High and low inference items - The evaluative criteria for the nine-item preinstructional material instrument utilized two types of increments on the criteria. The high inference criteria gave no rationale for selection of a low (1) or high (5) rating

(see Appendix A). The low inference instrument provided detailed descriptors to assist the choice of low (1) or high (5) ratings along the scale (see Appendix B).

6. Amount of training in the teacher appraisal process - The previous training an evaluator had experienced in teacher appraisal was placed in one of six categories. These were:
   a. No training (1)
   b. One day or less (2)
   c. One to three days (3)
   d. Three to five days (4)
   e. More than five days (5)
   f. Component of a college course (6)

   For frequency distributions by training, see Appendix J, Table J.5 and for standard deviations, see Appendix J, Table J.6.

7. Job description - The participants chose one of four categories to describe their jobs. These were:
   a. Superintendent (1)
   b. Principal (2)
   c. Supervisor (3)
   d. Other (4)

   For frequency distributions by job, see Appendix J, Table J.7 and for standard deviations, see Appendix J, Table J.8.

8. Geographical regions - The scores from the eight regions were purposefully segregated to determine if regional differences could be detected. These geographical regions were:

a. Philadelphia, Pennsylvania (1)

b. Detroit, Michigan (2)

c. Ontario, Canada (3)

d. Portage, Indiana (4)

e. Oyster Bay, New York (5)

f. Novato, California (6)

g. Creston, Iowa (7)

h. Albuquerque, New Mexico (8)

For frequency distributions by region, see Appendix J, Table J.9 and for standard deviations, see Appendix J, Table J.10.

Eleven hypotheses were posited in this research to study varying combinations and interactions of the above variables. All data obtained from this study were analyzed using SPSS procedures, FREQUENCIES, CROSS TABS, ONE-WAY, or PEARSON CORR. Statistical options used included the Chi-square test of significance which is part of the CROSS TABS package. FREQUENCIES was used primarily to tabulate raw data and percent of responses by selected groupings. CROSS TABS was utilized to compare rating accuracy scores by amount of training, job, and geographical region. The ONE-WAY was used to determine if significant differences in means occurred between groups with the CONTRAST program determining which groups were to be compared. The PEARSON CORR established the degree of relationship on rating accuracy between groups.

All statistical tests were required to fall within the .05 level of significance to reject a given hypothesis, and failure to support any one of the eleven major null hypotheses required that 75 percent of the

criteria variables (AIM or ATP) must have been rejected. This determination was made by the investigator and was arbitrarily set.

Findings

The first three hypotheses of the study investigated the differences between the four treatment groups and their accuracy of rating teacher performance. A one-way analysis of variance with contrasts was utilized to determine whether the mean of any of the groups differed significantly from the mean of any other. These hypotheses and their findings are as follows:

Hypotheses 1. There will be no difference in teacher appraisers' rating accuracy of teacher performance given preinstructional materials for study during the preobservational period, as compared with those who are not provided preinstructional materials for study.

The findings of the one-way analysis of variance with contrast (3-1-1-1) which compares the mean of group one with the averaged means of groups two, three, and four (see Appendix J, Table J.11) are summarized in Table 2, and reveal that of the eleven ATP criteria only one exceeds the critical area set for .05 level of significance. This was ATP 4 which indicates that there was a difference in the mean of group one and the averaged group means, those provided with preinstructional materials and those not provided with preinstructional materials when judging teachers' performance of utilizing a variety of teaching methods, such as would occur less than five times out of one hundred.

This analysis provided evidence that the null hypothesis may not be rejected as 75 percent of the ATP criteria did not fall within the rejection area. Therefore, it may be stated that there is no difference in teacher appraisers' rating accuracy of teacher performance between

Table 2. Analysis of variance summary on accuracy of teacher performance (ATP) contrasting placebo group vs. lesson plans only, lesson plans with high inference, and lesson plans with low inference

	F	F probability	Contrast t	t probability
ATP 1 (Use of AV)	1.16	.33	.77	.44
ATP 2 (Practice on objectives)	2.09	.10	.74	.08
ATP 3 (Logical sequence)	1.93	.12	.70	.48
ATP 4 (Variety of methods)	4.39	.01	-2.23	.03*
ATP 5 (Appropriate methods)	.53	.67	-1.17	.24
ATP 6 (Group sizes)	1.26	.29	-0.85	.40
ATP 7 (Learner involved)	1.72	.16	.89	.38
ATP 8 (Maintain involvement)	.25	.86	-0.13	.90
ATP 9 (Knowledge of subject)	2.74	.04	.11	.91
ATP 10 (Accurate information)	1.82	.14	-0.33	.74
ATP 11 (Purpose of topic)	1.20	.31	1.38	.17

*$p < .05$.

those studying preinstructional material before observation and those not studying preinstructional materials before observation.

Statistical results for each individual criteria are recorded in tables and may be found in Appendix J, Table J.12 through Table J.22.

Hypotheses 2. There will be no difference in teacher appraisers' rating accuracy of teacher performance given preinstructional materials and evaluative criteria for assessing those materials, as compared with those who are provided the preinstructional materials but not the evaluative criteria.

Summary Table 3 reports the one-way analysis of variance with contrast (2-1-1) comparing the averaged means of groups one and two and the averaged means of groups three and four (see Appendix J, Table J.11), and the finding that three of the criteria were rejected at the designated .05 level of significance.

These criteria are ATP 3, ATP 9, and ATP 10. Criterion ATP 3 relates to the teacher's organization of activities in a logical sequence, ATP 9 refers to the teacher's knowledge of the subject area, and finally, ATP 10 has to do with the teacher presenting accurate information about the topic being taught. On these three criteria, the averaged means of the two groups, those provided with evaluative criteria and those not provided with evaluative criteria, are observed to be significantly different such as would occur less than five times in one hundred.

The null hypothesis must be retained, however, because 75 percent of the total ATP criteria did not meet the .05 level of significance. The statement can be maintained that there will be no difference in teacher evaluators' rating accuracy of teacher performance given preinstructional materials and evaluative criteria for their assessment as opposed to those who study the preinstructional materials but do not utilize an evaluative criteria.

The individual ATP criteria findings are recorded as tables and may be found in Appendix J, Tables J.12 through Table J.22.

Table 3. Analysis of variance summary on accuracy of teacher performance (ATP) contrasting placebo group and lesson plans only vs. lesson plans with high inference and lesson plans with low inference

	F	F probability	Contrast t	t probability
ATP 1 (Use of AV)	1.16	.325	-0.36	.72
ATP 2 (Practice on objectives)	2.09	.100	-1.69	.09
ATP 3 (Logical sequence)	1.93	.124	-2.29	.02*
ATP 4 (Variety of methods)	4.39	.005	-1.82	.07
ATP 5 (Appropriate methods)	0.53	.665	-0.40	.69
ATP 6 (Group sizes)	1.26	.287	-1.43	.15
ATP 7 (Learners involved)	1.72	.162	-1.24	.38
ATP 8 (Maintain involvement)	0.25	.861	-0.33	.45
ATP 9 (Knowledge of subject)	2.74	.043	-2.33	.02*
ATP 10 (Accurate information)	1.82	.142	-2.31	.02*
ATP 11 (Purpose of topic)	1.20	.310	0.88	.38

*$p > .05$.

Hypothesis 3. There will be no difference in teacher appraisers' rating accuracy of teacher performance given preinstructional materials and an evaluative instrument utilizing low inference descriptors on the rating scale, as compared with those provided with preinstructional materials and an evaluative instrument utilizing high inference descriptors on the rating scale.

The one-way analysis of variance and the contrast (-1-1) which compares the mean of group three with the mean of group four (see Appendix J, Table J.11) revealed that of the eleven criteria tested, only one, ATP 4 rejected at the .05 level of significance. The summary Table 4 reports this and other findings.

This one criterion meeting the level of significance requirement (ATP 4) relates to the teacher's ability to conduct lessons using a variety of teaching methods. It may be stated that there was observed a difference between the means of the two groups, those using high inference descriptors on the evaluative instrument and those using low inference descriptors on the evaluative instrument, such as would occur less than five times out of one hundred.

Again, the criterion of 75 percent of the ATP being rejected at .05 level was not met to reject the major null hypothesis. It appears that rating accuracy of teacher performance will not be affected by studying preinstructional materials and using a low inference evaluative criteria to rate these materials as compared to evaluators who use a high inference instrument to evaluate.

The complete tables of individual criteria ATP 1 - ATP 11 findings are to be found in Appendix J, Tables J.12 through Table J.22.

The fourth hypothesis was tested by a one-way analysis of variance to test if a significant difference in means existed between two treatment groups and their rating accuracy of preinstructional materials. A rejection level of .05 significance was set.

Table 4. Analysis of variance summary on accuracy of teacher performance (ATP) contrasting lesson plans with high inference vs. lesson plans with low inference instruments

	F	F probability	Contrast t	t probability
ATP 1 (Use of AV)	1.16	.325	1.66	.10
ATP 2 (Practice on objectives)	2.09	.100	0.52	.60
ATP 3 (Logical sequence)	1.93	.124	0.07	.94
ATP 4 (Variety of methods)	4.39	.005	2.24	.03*
ATP 5 (Appropriate methods)	0.53	.665	0.23	.82
ATP 6 (Group sizes)	1.26	.287	1.02	.31
ATP 7 (Learners involved)	1.72	.162	1.67	.10
ATP 8 (Maintain involvement)	0.25	.861	0.40	.69
ATP 9 (Knowledge of subject)	2.74	.043	1.66	.10
ATP 10 (Accurate information)	1.82	.142	-0.17	.86
ATP 11 (Purpose of topic)	1.20	.310	0.98	.33

*$p > .05$.

Hypothesis 4. There will be no difference in teacher appraisers' rating accuracy of preinstructional materials given an evaluative instrument utilizing low inference descriptors on the rating scale, as compared with those provided high inference descriptors on the rating scale.

The statistical findings of the one-way analysis of variance reported in Table 5 indicate that five criteria met the .05 level of rejection standard. These criteria were AIM 3, AIM 5, AIM 6, AIM 7, and AIM 8. These criteria relate to the following areas:

1. AIM 3 - Organization of learning activities in a logical sequence.
2. AIM 5 - Use of teaching methods appropriate for objectives and learners.
3. AIM 6 - Ability to work with individuals, small groups, and large groups.
4. AIM 7 - Use of procedures which initially involve learners in the lesson.
5. AIM 8 - Maintains learner involvement in lessons.

It may be stated that there is a difference in the means of these two groups, those using a high inference instrument and those using a low inference instrument, when judging the above five items. This would occur by chance less than five times out of one hundred.

Whereas the number of rejected AIM criteria is approaching the required 75 percent quota, it is still insufficient to reject the fourth major null hypothesis. The statement can now be made that there will be no difference in teacher evaluators' rating accuracy of preinstructional materials given an evaluative instrument utilizing low inference descriptors on the rating scale as compared with evaluators using an evaluative criterion with high inference descriptors.

Table 5. A one-way analysis of variance contrasting lesson plans with low inference (group 3) vs. lesson plans with high inference (group 4) instruments

	N_3 [a]	$\bar{X}_3$ [b]	S_3 [c]	N_4 [d]	$\bar{X}_4$ [e]	S_4 [f]	t	t Probability
AIM 1	126	1.27	1.04	127	1.14	.97	-1.46	.14
AIM 2	126	.96	.75	127	1.06	.79	1.53	.15
AIM 3	126	1.04	.73	127	.71	.77	-5.08	.00*
AIM 4	126	1.11	.70	127	1.14	.87	.45	.66
AIM 5	126	1.11	1.05	127	1.46	1.14	3.70	.00*
AIM 6	126	1.13	1.01	127	1.37	1.16	2.49	.01*
AIM 7	126	1.02	.77	127	1.27	.86	3.55	.00*
AIM 8	126	.99	.76	127	1.34	.80	5.09	.00*
AIM 9	126	1.26	.72	127	1.32	.79	.81	.42

[a] N_3 = members in treatment group 3 (high inference instrument)

[b] $\bar{X}_3$ = mean of treatment group 3.

[c] S_3 = standard deviation of group 3.

[d] N_4 = members in treatment group 4 (low inference instrument).

[e] $\bar{X}_4$ = mean of treatment group 4

[f] S_4 = standard deviation of treatment group 4.

* $p > .05$.

The fifth hypothetical proposition deals with the relationship of variables of accuracy and the use of high inference or low inference descriptors on the evaluation instrument. To test this relationship, a correlation coefficient was run with a .05 level of significance required for rejection.

Hypothesis 5. There will be no relationship between the accuracy of teacher appraiser's judging preinstructional materials and the accuracy of those judging teacher performance.

A Pearson Product-Moment correlation coefficient was run and the statistical results from this procedure are recorded in Table 6 revealing that the composite of teacher performance rating accuracy and the composite on instructional materials rating accuracy correlates highly at a .00 significance level. This will reject the null hypothesis and establish that there is a positive relationship between rating accuracy of evaluators' judging preinstructional materials and the rating accuracy of those judging teacher performance. In other words, those who more accurately rated the lesson plans also more accurately rated the teacher's performance.

Table 6. Correlation coefficient of rating accuracy on preinstructional materials criteria

	Rating accuracy		
Criteria	Composite	High inference	Low inference
N	249	122	127
r	.25	.38	.15
p	.00*	.00*	.05*

*$p < .05$.

Hypothesis 5a. If the teacher appraisers received a high inference evaluative instrument, there will be no relationship between those who more accurately judged the preinstructional materials and those who more accurately judged teacher performance.

Table 6 reports the correlation coefficient findings for hypothesis 5a which meets the .05 level of significance. This is sufficient to reject the null hypothesis and establish that there is a relationship between those evaluators who more accurately judged the preinstructional materials and those who more accurately judged teacher performance.

Hypothesis 5b. If the teacher appraisers received a low inference evaluative instrument, there will be no relationship between those who more accurately judged the preinstructional materials and those who more accurately judged teacher performance.

The statistical findings in Table 6, those from the analysis of scores collected on ratings of low inference evaluative instruments, indicate that there is a correlation of .15 which is a significant finding. This substantiates that there is a relationship, of those participants utilizing low inference evaluative criteria, between those more accurately judging the preinstructional materials and those more accurately judging teacher performance than would occur by chance less than five times out of one hundred.

The remaining hypotheses were all submitted to the Chi-square (x^2) nonparametric statistical test of significance. This test is used to determine if variables are independent of each other. Frequency counts are placed into categories as observed cases and are examined to determine if this count differs from the frequencies that would be expected by chance. A qualification is outlined by Ott, Mendenhall, and Larson

which states that "the expected cell satisfy the requirement that no expected cell count is less than 1 and no more than 20% are less than 5" (1978, p. 318). If such a condition exists, the Chi-square analysis will be invalid. Ott, Mendenhall, and Larson also say that adjacent cells may be combined as a correction to increase the expected frequencies, however, this may introduce an arbitrary component into the analysis and it must be used with caution.

This condition of expected cell frequencies being less than five in more than 20 percent of the cells was found in this investigation. The researcher carefully considered feasibility of correction by collapsing adjacent cells. Determination was made on an individual basis and correction was undertaken only if it appeared to contribute meaning to the study. Therefore, when corrections are not undertaken the invalid Chi-square analyses are allowed to prevail, these cases will be reported as descriptive contingency tables.

The subsequent hypothesis and the Chi-square findings are as follows:

Hypothesis 6. Teacher appraisers' rating accuracy on preinstructional materials will be independent of previous training in teacher performance appraisal.

A Chi-square analysis was run cross-tabbing a table on rating accuracy of preinstructional materials with previous training in teacher appraisal. Table 7 summarizes the AIM score analyses and it will be noted that none of the items met the .05 level of significance. For a complete recording of the contingency tables cross-tabbing accuracy of

rating instructional materials (AIM) with training in teacher appraisal (TATRAIN), see Appendix J, Tables J.23 through Table J.31.

Table 7. Chi-square summary analysis on accuracy on instructional materials (AIM) by training (TATRAIN)

	X^2	df	Probability
AIM 1 (Appropriate objectives)	See Table 8[a]		
AIM 2 (Procedure for objectives)	12.81	15	.62
AIM 3 (Specific content)	9.31	15	.86
AIM 4 (Appropriate for mastery)	18.10	15	.26
AIM 5 (Cognitive levels)	See Table 9[a]		
AIM 6 (Learner capabilities)	See Table 10[a]		
AIM 7 (Learning styles)	7.13	15	.95
AIM 8 (Rates of learning)	7.62	15	.94
AIM 9 (Evaluation material)	17.02	15	.32

[a] X^2 not appropriate because of low cell frequencies.

Three of the AIM criteria produced data in which 20 percent of the cells had expected frequencies of less than five. The first, AIM 1 on judging appropriate instructional objectives, has 25 percent of the expected cell frequencies less than five, and therefore, no valid Chi-square can be reported for this criterion. The frequency contingency table for AIM 1 is registered in Table 8.

Table 8. Observed frequencies and expected frequencies for AIM 1 (appropriate objectives) by training

Obs. \ Exp.	1	2	3	4	5	6
.00	23.4 17	9.3 13	12.8 15	3.2 2	5.0 5	21.0 23
1.00	25.8 27	10.0 8	14.0 20	3.6 1	5.5 5	23.5 22
2.00	105.7 109	42.0 43	57.7 47	14.7 19	22.0 22	96.0 99
3.00	9.9 12	3.9 2	5.0 8	1.0 1	2.0 3	9.0 6

6 out of 24 (25%) have expected cell frequencies less than 5.

The frequencies for AIM 5 are recorded in Table 9 and indicate that 30 percent of the expected cell frequencies are less than five making the Chi-square findings invalid.

Table 9. Observed frequencies and expected frequencies for AIM 2 (cognitive levels) by training

Obs. \ Exp.	1	2	3	4	5	6
.00	22.0 19	8.9 11	12.0 17	3.1 4	4.7 3	20.0 17
1.00	26.3 26	10.7 11	14.16 14	3.7 2	5.7 10	24.0 23
2.00	18.0 14	7.2 10	9.8 11	2.5 2	3.8 3	16.4 18
3.00	8.7 12	3.5 2	4.7 4	1.2 1	1.8 1	7.9 8
4.00	8.9 94	35.6 32	48.6 44	12.0 14	18.9 18	81.0 84

9 out of 30 (30%) have expected cell frequencies less than 5.

The contingency table recording frequencies on AIM 6 cross-tabbed with training are listed in Table 10. Another instance of expected cell frequencies less than five exceeding the 20 percent standard is found here. Item AIM 6 reveals a 30 percent incidence of expected cell frequencies less than five, and again, no Chi-square statistic is deemed valid in this circumstance.

Table 10. Observed frequencies and expected frequencies for AIM 6 (learner capabilities) by training

Obs. \ Exp.	1	2	3	4	5	6
.00	22.0 19	8.8 10	12.0 14	3.0 3	4.7 5	20.0 20
1.00	12.8 31	11.4 15	15.6 19	4.0 4	6.0 4	26.0 19
2.00	17.0 11	6.8 6	9.3 12	2.3 2	3.6 5	15.6 19
3.00	7.8 9	3.0 3	4.0 2	1.0 0	1.6 3	7.0 8
4.00	89.0 95	35.6 32	48.6 43	12.0 14	18.9 18	81.0 84

9 out of 30 (30%) have expected cell frequencies less than 5.

In conclusion, the major null hypothesis cannot be rejected because 75 percent of the nine criteria could not meet the .05 level of significance. It can now be stated that teacher evaluators' rating accuracy on preinstructional materials is independent of previous training in teacher performance appraisal.

Hypothesis 7. Teacher appraisers' rating accuracy of teacher performance will be independent of previous training in teacher performance appraisal.

This null hypothesis was examined by Chi-square with cross-tabbing to produce contingency tables (see Appendix J, Tables J.32 through Table J.42) and Table 11 capsulizes these data.

Table 11 indicates, that of the eleven criteria tested, only ATP 8 is rejected at the .05 level of significance. Such a finding allows the statement to be made that teacher appraisers' rating accuracy is dependent upon training in teacher assessment when judging the maintenance of student involvement with the lesson's objectives.

The researcher's requirement that 75 percent of any given set of AIM or ATP scores must be rejected to allow rejection of the major null hypothesis was not met for hypothesis 7, therefore, it may be stated that teacher appraisers' rating accuracy of teacher performance is independent of the amount of training the evaluator has had in the teacher appraisal process.

Hypothesis 8. The teacher appraisers' rating accuracy on preinstructional materials will be independent of the region of the country in which the appraiser resides.

The Chi-square, which cross-tabbed rating accuracy with region, is recorded in summary Table 12, and the complete contingency tables for the ATP computations are in Appendix J, Tables J.43 through Table J.51. Summary Table 12 reports the nine criteria findings and indicates that none reached .05 level for rejection.

Table 11. Chi-square summary analysis on accuracy of teacher performance (ATP) by training (TATRAIN)

	X^2	df	Probability
ATP 1 (Use of AV)	6.32	10	.78
ATP 2 (Practice on objectives)	9.80	10	.46
ATP 3 (Logical sequence)	11.65	10	.31
ATP 4 (Variety of methods)	12.92	15	.61
ATP 5 (Appropriate methods)	10.23	15	.81
ATP 6 (Group sizes)	10.61	10	.39
ATP 7 (Involves learners)	17.83	10	.06
ATP 8 (Learner involvement)	22.48*	10	.01*
ATP 9 (Knowledge of subject)	9.45	10	.49
ATP 10 (Accurate information)	9.98	5	.08
ATP 11 (Purpose of topic)	7.29	10	.70

*$p < .05$.

Table 12. Chi-square summary analysis on accuracy on instructional materials (AIM) by regions (REG)

	X^2	df	Probability
AIM 1 (Appropriate objectives)	See Table 13[a]		
AIM 2 (Procedures for objectives)	24.93	21	.25
AIM 3 (Specific content)	16.75	21	.73
AIM 4 (Appropriate for mastery)	23.62	21	.31
AIM 5 (Cognitive levels)	See Table 14[a]		
AIM 6 (Learner capabilities)	See Table 15[a]		
AIM 7 (Learning styles)	18.24	21	.63
AIM 8 (Rates of learning)	27.52	21	.15
AIM 9 (Evaluation material)	11.18	21	.96

[a] X^2 not appropriate because of low cell frequencies.

Three of the AIM criteria revealed expected cell frequencies, with 20 percent of the expected frequencies below five. This makes interpretation spurious in these instances; however, correction was not considered in the best interests of data interpretation, and the frequencies are submitted as descriptive evidence of the findings. Criterion AIM 1 is recorded in Table 13 showing nine out of thirty, or 28 percent, of the expected cell frequencies are less than five. Table 14 records AIM 5 with 27 percent of the expected cell frequencies as less than five, and

Table 13. Observed frequencies and expected frequencies for AIM 1 (appropriate objectives) by region

Obs. \ Exp.	1	2	3	4	5	6	7	8
0.00	9.0 7	6.9 3	15.7 17	15.0 22	9.0 9	3.0 4	4.5 5	10.2 8
1.00	10.0 10	7.6 7	17.0 17	16.7 16	10.0 12	4.3 7	5.0 4	11.3 10
2.00	41.0 44	31.0 34	71.0 67	68.0 64	42.0 41	17.9 17	20.5 22	46.0 50
3.00	3.8 3	2.9 5	6.7 10	6.4 5	3.9 4	1.6 0	1.9 1	4.3 4

9 of 32 (28.1%) of the expected cell frequencies are less than 5.

Table 14. Observed frequencies and expected frequencies for AIM 5 (cognitive levels) by region

Obs. \ Exp.	1	2	3	4	5	6	7	8
0.00	8.6 7	6.5 5	14.9 16	14.3 20	8.8 7	3.7 1	4.2 4	9.6 11
1.00	10.4 13	7.9 5	18.0 22	17.4 16	10.7 13	4.5 5	5.2 7	11.7 5
2.00	7.0 8	5.0 7	12.0 11	11.7 11	7.0 8	3.0 5	3.5 4	7.8 4
3.00	3.4 2	2.6 4	5.9 8	5.0 4	3.4 2	1.5 1	1.6 1	3.8 6
4.00	34.6 34	26.5 28	60.0 54	59.9 56	35.7 36	15.0 16	17.0 16	39.0 46

11 out of 40 (27.5%) of the expected cell frequencies are less than 5.

finally, AIM 6 exhibits 27.5 percent of the expected cell frequencies as less than five and are listed in Table 15.

The findings for hypothesis 8 indicate conclusively that 75 percent of the nine criteria did not meet the .05 level of rejection, and this being so, it may be said that teacher evaluator's rating accuracy of pre-instructional materials is independent of the region of the country in which the evaluator lives.

Hypothesis 9. The teacher appraisers' rating accuracy on teacher performance will be independent of the region of the country in which the appraiser resides.

The process of cross-tabbing rating accuracy of teacher performance with regions has yielded eleven contingency tables on the individual ATP variables (see Appendix J, Tables J.52 through Table J62). The findings of these eleven tables are summarized in Table 16. It can be noted there that five of the eleven criteria reached the .05 level of significance. These were:

1. ATP 1
2. ATP 3
3. ATP 4
4. ATP 5
5. ATP 10

Of these five items, corrections were undertaken on ATP 1 and ATP 10. It should also be recognized that corrections were done on ATP 2 and ATP 11 although they did not meet the .05 level of significance. Cell corrections were achieved by collapsing cells with adjacent cells. This decision was made after consultation with Anton Netusil, Department

Table 15. Observed frequencies and expected frequencies for AIM 6 (learner capability) by region

Obs. \ Exp.	1	2	3	4	5	6	7	8
0.00	8.6 6	6.5 7	14.9 17	14.3 17	8.8 11	3.7 3	4.2 3	9.6 7
1.00	11.0 13	8.20 7	19.3 21	18.6 19	11.4 12	4.8 2	5.6 8	2.5 10
2.00	6.6 11	5.0 4	11.5 12	11.1 12	6.8 4	2.9 5	3.3 3	7.4 4
3.00	3.0 0	2.3 3	5.2 7	5.0 4	3.0 3	1.3 1	1.5 2	3.0 5
4.00	34.0 34	26.0 28	60.0 54	57.0 55	35.0 36	15.0 17	17.0 16	38.9 46

11 of 40 (27.5%) of the expected cell frequencies are less than 5.

Table 16. Chi-square summary analysis on accuracy on teacher performance (ATP) by region

	X^2	df	Probability
ATP 1 (Use of AV)	72.23	14	.01*
ATP 2 (Practice on objectives)	13.53	14	.49
ATP 3 (Logical sequence)	25.88	14	.03*
ATP 4 (Variety of methods)	48.73	21	.00*
ATP 5 (Appropriate methods)	34.56	21	.03*
ATP 6 (Group sizes)	22.77	14	.06
ATP 7 (involves learners)	20.86	14	.11
ATP 8 (Learner involvement)	20.29	14	.12
ATP 9 (Knowledge of subject)	See Table 17[a]		
ATP 10 (Accurate information)	17.68	7	.01*
ATP 11 (Purpose of topic)	23.19	14	.06

[a] X^2 not appropriate because of low cell frequencies.

* $p < .05$.

of Professional Studies in Education, Iowa State University, and determination was made to collapse the cross-classification row with extremely low observed frequencies.

Criterion 9 (ATP 9) reported 25 percent of the valid cells having expected frequencies less than five. It was deemed inadvisable to collapse the cells in the row of the cross-classification table, therefore, no Chi-square is submitted for this criterion as it would be deemed invalid. The contingency table for criteria 9 (ATP 9) is reported in Table 17.

In summarizing the findings on hypothesis 9, it appears that even though six of the eleven criteria could meet the .05 level of significance, and criterion 9 was found to be invalid by statistical testing, there is not sufficient evidence upon which to reject the major null hypothesis. In studying these data, it would be untenable to conclude that rating accuracy is dependent upon region.

Hypothesis 10. Teacher appraisers' rating accuracy on preinstructional materials will be independent of the appraisers' job positions.

The nine AIM criteria were examined by a Chi-square analysis and were reported in cross-tabulation frequency tables (see Appendix J, Tables J.63 through Table J.71). Summary Table 18 reiterates these results and indicates that only AIM 7 meets the required .05 level of significance. This conclusion reveals that teachers' rating accuracy on judgments of the teachers' accommodation of differences in student rates of learning is, in fact, associated with the appraisers' job position.

Table 17. Observed frequencies and expected frequencies for ATP 9 (knowledge of subject) by region

Obs. \ Exp.	1	2	3	4	5	6	7	8
0.00	42.5 38	31.9 21	73.8 77	71.0 85	43.8 38	18.6 18	21.2 22	47.8 52
1.50	17.4 21	13.0 22	30.2 28	29.0 19	18.0 21	7.6 7	8.7 8	19.6 18
2.50	4.0 5	3.0 5	6.9 6	6.6 3	4.1 7	1.7 3	2.0 2	4.5 2

6 out of 24 (25%) of the valid cells have expected cell frequency less than 5.

Table 18. Chi-square summary analysis on accuracy on instructional materials (AIM) by job

	X^2	df	Probability
AIM 1 (Appropriate objectives)	See Table 19[a]		
AIM 2 (Procedures for objectives)	11.50	9	.24
AIM 3 (Specifies content)	See Table 20[a]		
AIM 4 (Appropriate for mastery)	15.93	9	.07
AIM 5 (Cognitive levels)	See Table 21[a]		
AIM 6 (Learner capabilities)	See Table 22[a]		
AIM 7 (Learning styles)	18.23	9	.03*
AIM 8 (Rate of learning)	16.63	9	.05
AIM 9 (Evaluation materials)	See Table 23[a]		

[a] X^2 not appropriate because of low cell frequencies.

* $p < .05$.

Five of the criteria were again disqualified from Chi-square analysis because the condition prevailed wherein the expected cell frequencies of less than five exceeded 20 percent. The first of these five is AIM 1 which can be found in Table 19. Table 20 records findings on AIM 3, Table 21 reveals the computations on AIM 5, whereas AIM 6 is listed in Table 22, and finally, Table 23 records the resulting analysis on AIM 9.

The major null hypothesis 10 cannot be rejected as the researcher's criterion that 75 percent of the AIM criteria must meet the .05 level of

Table 19. Observed frequencies and expected frequencies for AIM 1 (appropriate objectives) by job

Obs. \ Exp.	1	2	3	4
0.00	2.3 3	36.7 35	7.1 3	27.6 33
1.00	2.6 2	40.0 41	7.8 3	30.0 35
2.00	10.8 8	168.0 174	32.8 42	126.0 114
3.00	1.0 4	15.9 11	3.1 3	11.9 14

4 out of 16 (25%) have expected cell frequencies less than 5.

Table 20. Observed frequencies and expected frequencies for AIM 3 (specifies appropriate content) by job

Obs. \ Exp.	1	2	3	4
0.00	2.7 2	41.0 42	8.0 7	31.0 33
1.00	3.8 6	59.0 56	11.0 10	44.0 48
2.00	1.2 2	19.0 16	3.8 1	14.5 20
3.00	9.0 7	140.0 147	27.0 33	105.0 95

4 out of 16 (25%) of the expected cell frequencies are less than 5.

Table 21. Observed frequencies and expected frequencies for AIM 5 (variety of cognitive levels) by job

Obs. \ Exp.	1	2	3	4
0.00	5 / 2.2	36 / 34.8	2 / 6.8	27 / 26.0
1.00	2 / 2.7	46 / 42.0	9 / 8.0	28 / 31.7
2.00	0 / 1.8	24 / 28.3	5 / 5.5	28 / 21.2
3.00	3 / 0.9	7 / 3.9	1 / 2.7	17 / 10.4
4.00	7 / 9.2	148 / 141.6	34 / 27.7	96 / 106.4

5 out of 20 (25%) of the expected cell frequencies are less than 5.

Table 22. Observed frequencies and expected frequencies for AIM 6 (accommodate learner capabilities) by job

Obs. \ Exp.	1	2	3	4
0.00	3 / 2.2	36 / 34.8	5 / 6.8	26 / 26.1
1.00	5 / 2.9	47 / 45.2	6 / 8.8	33 / 33.9
2.00	0 / 1.8	23 / 26.8	5 / 5.2	26 / 20.0
3.00	2 / 0.8	8 / 12.4	1 / 2.4	14 / 9.3
4.00	7 / 9.2	147 / 141.7	34 / 27.7	97 / 106.4

5 out of 20 (25%) expected cell frequencies are less than 5.

Table 23. Observed frequencies and expected frequencies for AIM 9 (evaluation material) by job

Obs. \ Exp.	1	2	3	4
0.00	1.1 1	17.9 20	3.5 2	13.4 13
1.00	3.6 3	56.0 48	10.9 11	42.1 51
2.00	3.0 6	45.7 46	8.9 4	34.3 36
3.00	9.2 7	141.0 147	27.3 34	106.0 96

4 out of 16 (25%) expected cell frequencies are less than 5.

significance was not accomplished. Therefore, it may be stated that teacher appraisers' rating accuracy on preinstructional materials is independent of appraisers' job positions.

Hypothesis 11. Teacher appraisers' rating accuracy of teacher performance will be independent of the appraisers' job positions.

Chi-square analysis was undertaken on the eleven ATP criteria and the findings of these individual tests are reported in Appendix J, Tables J.72 through Table J.82. Summary Table 24 restates these computations and, it will be noted, AIM 1 is the only criterion that meets the .05 level of significance. This finding indicates that rating accuracy on the teacher's performance, when utilizing audiovisual aids, is not independent of the evaluator's job.

The APT 9 criterion, pertinent to judging the teacher's knowledge of subject, registered 20 percent of the cells as having less than five in the expected cell frequencies. A descriptive report of this criterion

Table 24. Chi-square summary analysis on accuracy of teacher performance (ATP) by job

	X^2	df	Probability
ATP 1 (Use of AV)	14.40	6	.03*
ATP 2 (Practice on objectives)	6.17	6	.40
ATP 3 (Logical sequence)	4.57	6	.60
ATP 4 (Variety of materials)	5.09	9	.83
ATP 5 (Appropriate methods)	14.24	9	.11
ATP 6 (Group sizes)	6.59	6	.36
ATP 7 (Involves learners)	5.40	6	.49
ATP 8 (Learner involvement)	8.39	6	.21
ATP 9 (Knowledge of subject)	See Table 25[a]		
ATP 10 (Accurate information)	2.39	3	.50
ATP 11 (Purpose of topic)	7.59	6	.27

[a] X^2 not appropriate because of low cell frequencies.

* $p < .05$.

is listed in Table 25.

The judgment on hypothesis 11 must be made according to the required standard that 75 percent of the criteria be rejected in order to reject the null hypothesis. With this in mind, the hypothesis cannot be rejected and it can be stated that teacher evaluator's rating accuracy of teacher performance is independent of the part of the country where the appraiser lives.

Table 25. Observed frequencies and expected frequencies for ATP 9 (knowledge of subject) by job

Obs. \ Exp.	1	2	3	4
0.00	11.3 13	174.0 179	33.0 30	130.0 128
1.50	4.6 3	70.0 72	13.5 15	53.0 52
2.50	1.0 1	15.9 10	3.0 5	11.9 16

3 out of 12 (25%) expected cell frequencies are less than 5.

CHAPTER V. SUMMARY AND CONCLUSIONS

Summary

The hypothetical concept for this study was generated from the need to address certain unanswered questions regarding educational administrators' capabilities to function effectively as teacher appraisers. The literature review discloses that many administrators are overwhelmed by this responsibility, and Kowalski (1979) reports that many administrators feel inadequately prepared in their educational training programs, and too long and too far removed from their own classroom experiences to do suitable teacher evaluation.

This becomes an important issue as, more and more, administrators are being called upon to do valid, reliable, and legally discriminating appraisal to substantiate their decisions on promotion, tenure, merit, and termination. Recent trends have extended the role responsibility of the appraiser, not only to assess the teacher's performance, but also to prescribe actions which, if followed by the teacher, will result in improved instruction. This heightens the anxiety levels of administrators, because their knowledge of effective teaching now becomes conspicuous to the instructional staff. Administrators fear their educational leadership image will diminish if they fail to make appropriate recommendations for improvement of instruction.

These apprehensions, as expressed by educational administrators, are the focus of this research which was undertaken to determine if such feelings were based upon imagined or real inabilities to evaluate the instructional process. The study began by looking at the administrator's

capability to assess the teacher's instructional planning phase, which was referred to throughout the investigation as preinstructional planning.

The research was designed to test the administrator's ability to accurately evaluate the teacher's preinstructional planning by requiring participants (excluding the placebo group) to assess a teacher's lesson plans, objectives, instructional materials, and evaluation techniques. The administrator's ability to make more accurate assessments of the teacher's performance by utilizing the knowledge gained from studying the preinstructional materials was examined. The question was, do evaluators take the information gained from the teacher's plan, which designates the intended educational outcomes and actions, and use this input to make more accurate judgments about the curriculum congruence of the instruction when it is played out in the classroom?

Furthermore, there was a desire to know if educational administrators could be assisted to make more accurate evaluations by using an evaluation instrument describing what was to be observed and assessed. If the use of an evaluative instrument contributed to increased rating accuracy, rather than ignoring or superficially reviewing the preinstructinal materials, would there be greater improvement if criteria statements on the instruments were more specifically expressed? Also, would the use of such low inference descriptors on the evaluative scale increase rater accuracy when judging the preinstructional materials and the classroom performance of the teacher? Finally, the thought that rating accuracy could be dependent upon the rater's geographic location, job position,

or training in the teacher evaluation process was investigated.

A simulation exercise was set up to gather data relevant to these questions. Various teacher evaluation workshops conducted for educational administrators across the country included this exercise in the program. The simulation included a two-step process. The first required the study and, for some, the evaluation of a teacher's lesson plans, objectives, instructional materials, and evaluation procedures (preinstructional materials) for an intermediate grade level language arts class. The second part required the evaluation of a videotape of the teacher's classroom performance as prescribed by the preinstructional materials. All ratings, whether on an instrument for assessing the preinstructional materials or the classroom performance, used one of two types of items; a high inference scale with nonspecific incremental levels, or a low inference scale with very precisely described rating levels.

All instruments and simulation exercise materials were adapted from those developed, normed, and validated by the Georgia Teacher Assessment Project. Because these items had been tested, rating accuracy of the subjects could be determined by comparing them with those established in the Georgia Project.

These data were collected in a five-week period from eight widely spread geographic locations. Five hundred twenty-nine subjects released their ratings for research analysis. The subjects included within their numbers superintendents, principals, supervisors, and various other professional certified staff members. Their previous training and background in the teacher appraisal process ranged from no training to extensive

training.

Randomization was used to assign subjects to a control or experimental group and a posttest-only research design was utilized. Data processing was conducted at Iowa State University using the Statistical Package for the Social Sciences (SPSS), and the analysis techniques included Chi-square, one-way analysis of variance, and correlation coefficient. Eleven operational hypotheses were generated from the initial research concepts.

The statistical analysis of the data yielded the results summarized below.

1. It appears that teacher appraisers who studied a teacher's preinstructional plans did not utilize this information to more accurately assess the teacher's classroom performance. The only instance where accuracy was significantly increased by previewing the teacher's planning was when the appraiser was assessing the teacher's ability to use a variety of methods in instruction.

2. The use of an evaluative criteria instrument to aid in evaluation of the teacher's classroom performance did not significantly increase the rater's accuracy except in the following cases:

   a. When judging a teacher's organization of learning activities in a logical sequence,

   b. When assessing the teacher's knowledge of the subject area, and

   c. When evaluating the accuracy of information on the topic being taught.

3. The rating accuracy of teacher appraisers did not increase when using an evaluation instrument with well-defined rating levels to evaluate a teacher's classroom procedures. The only time the low inference instrument contributed significantly to rating accuracy was when assessing the teacher's use of a variety of teaching methods in the instructional process.

4. The use of an evaluative instrument with low inference descriptors contributed significantly to accuracy of rating preinstructional materials in five instances. These were:

   a. When determining if plans specified appropriate content, materials, and media to accomplish the stated objectives,

   b. When ascertaining if instruction was planned to accommodate a variety of cognitive levels,

   c. When determining if instruction was organized to accommodate differences in learner capabilities,

   d. When deciding if instruction was organized to accommodate differences in learning styles, and

   e. When determining if instruction was planned to meet differences in learners' rates of learning.

5. Teacher appraisers who were more accurate in the assessment of the lesson plans and objectives in the preinstructional materials packet were also more accurate in their judgments of the teacher's classroom performance. This ability to rate accurately was significant whether the rater was using a low inference or high inference instrument, although it is interesting to note that raters using a high inference instrument were a little more likely to be accurate.

6. The amount of training an educational administrator had had in the teacher assessment process did not affect rating accuracy when judging the teacher's preinstructional materials.

7. When assessing the teacher's classroom performance, training in teacher assessment contributed to significantly more accurate rating when judging the maintenance of student involvement. In this instance, the following conditions were determined using 0 as the accurate score, 1 as one score away from the accurate score, and 2 as two scores away from the accurate score:

   a. There were more accurate ratings than would be expected by those with no training,

   b. There were more ratings 2.5 points away from accuracy than could be expected by those with one to three days of training,

   c. There were less ratings 2.5 points away from accuracy than could be expected by those with three to five days of training, and

   d. There were more ratings 2.5 points from accuracy than could be expected by subjects trained in a university course.

8. The accurate rating of preinstructional materials was not influenced by the geographical origin of the appraiser.

9. The appraiser's rating accuracy of teacher performance was, at times, dependent upon the geographical location of the rater. The following findings are reported:

   a. Rating was significantly more accurate when judging the use of audiovisual equipment. This dependency appears to come from more ratings than could be expected one point from accuracy and less ratings than could be expected two points from accuracy from subjects in Philadelphia, Pennsylvania. More ratings than could be expected also occurred two points from accuracy from the Canadians.

   b. Rating was significantly more accurate by region when the appraiser was assessing the teacher's logical sequencing of instruction. This geographic dependency appears to come from more ratings than could be expected 1.5 points from accuracy and less scores than could be expected 2.5 points from accuracy from the Iowa subjects. The New Mexican participants had less scores 1.5 points from accuracy and more scores 2.5 from accuracy than could be expected.

   c. Rating accuracy by region was significantly more accurate when assessing the variety of teaching materials. The dependency on accuracy by region here appears to come from Philadelphia where more scores than could be expected were three points from accuracy; Canada, where less accurate scores and more scores two points from accuracy occurred than could be expected; and New Mexico, where less scores than could be expected were two points from accuracy.

   d. Rating accuracy was significantly dependent upon the region when assessing the appropriateness of methods. This dependency appears to occur where more scores than could be expected three points from accuracy came from Canadian appraisers, and from New Yorkers and New Mexicans, who tallied less scores than could be expected three points from accuracy.

   e. Rating accuracy was dependent upon location when judging the use of accurate information on the topic. It appears here that less scores than could be expected 1.5 points from accuracy occurred in Philadelphia and more scores than could be expected 1.5 points from accuracy came from the New York and New Mexico participants.

10. The educational position held by the appraiser did not influence

rating accuracy of preinstructional materials except when judging the teacher's plans to accommodate differences in learning style. In this case, it appears that there are more scores two points from accuracy than could be expected when analyzing superintendents, less scores than could be expected two points from accuracy by supervisors, and more scores two points from accuracy than could be expected by the group designated as "others".

11. Rating accuracy was associated with job position when assessing use of audiovisual equipment. The dependency appears to occur with supervisors when more ratings are accurate than could be expected, more ratings are one point from accuracy than could be expected, and less scores than could be expected are recorded two points from accuracy.

Conclusions

The conclusions which may be made from these findings are as follows:

1. The study of preinstructional materials by teacher appraisers does not associate with better teacher performance evaluation.

2. The use of evaluation instruments to assist in structuring the evaluation did help when judging the teacher's use of logical sequencing, the teacher's knowledge of subject matter, and accuracy of the topic information.

3. The use of an evaluation instrument with rating levels specifically described on the scale did improve the evaluator's assessment, but only when judging the teacher's use of a variety of instructional materials.

4. The evaluation instrument with a low inference descriptor scale did increase rater accuracy when judging preinstructional materials as related to the following:

   a. The appropriateness of content methods and media for objectives,

   b. The determination of instruction meeting a variety of cognitive levels,

   c. The determination of learning capabilities being met by instruction,

d. The judging of learning styles being considered in planning, and

e. The judging of planning to accommodate rates of learning in the instruction.

5. The appraisers who were better at judging preinstructional materials were better at judging teacher performance.

6. The amount of training an appraiser had had in teacher appraisal did not influence the judging of preinstructional materials.

7. The degree of training in teacher assessment did effect the judging of student involvement with the instructional objectives. However, this effect was scattered and revealed no directional trends.

8. The rater's geographical location did not effect the accuracy of assessing preinstructional materials.

9. At times, the rater's location was associated with the rater's ability to accurately judge teacher performance. However, the results are scattered with no apparent trends being established.

10. The job assignment of the rater effected the judging of lesson plans when looking for accommodation of learning styles. Here, again, the scatter of findings prevents practical interpretation.

11. Accuracy of rating the teacher's performance when assessing use of audiovisual equipment was better when the rater was an educational supervisor.

Limitations

Research studies usually are confronted with circumstances that are limiting to the investigation. These situations may, or may not, have been foreseen before the study was undertaken. However, after a project is completed, these limitations become more obvious and it is then recognized that alternatives may have been taken to make valuable contributions to the study. As this study progressed, numerous limitations became apparent.

The evaluation instrument utilized in this study introduced some limitations. This measurement device used a five-point scale which restricted spread and narrowed the range of scores. The scale, even when using the low inference descriptors, tended to encourage selection of scores in the middle range and, therefore, minimized the discrimination power of the analyses. The utilization of a seven-point incremental scale might have influenced participants to place more ratings on either side of the median, and therefore, facilitate improved discrimination power.

Furthermore, the use of accuracy as a standard was limiting because it constricted the range of scores, and it did not identify directionality of the responses. In other words, were the subjects' ratings higher or lower than the accuracy scores adapted from the Georgia Teacher Assessment Project? The use of raw scores for analysis would have overcome this limitation by expanding the range of scores and demonstrating directionality from the accepted accurate score. In this way, it would have been possible to ascertain if the Georgia scores were more or less stringent than those assigned by the subjects of the study.

The subjects in the study received no training in the use of the evaluation instrument. They were required to give their best effort after simply hearing and reading the appropriate directions. Any difficulties associated with the use of the instrument could have gone unnoticed and possibly had implications for the validity of the data. A training session, and provision for a practice period with these particular instruments, may have been useful for improving the results of the study.

The simulation exercise used only one videotaped sequence, that of an intermediate language arts class. Again, this was a restriction because of student age group and subject matter viewed by the audience. Not all participants were equally able to work with the age group and subject presented. Providing elementary and secondary classes, as well as other subject matter areas, would have allowed for participants to choose those components they felt most competent to evaluate.

Finally, the sample was taken entirely from educators enrolled in professional improvement workshops. They volunteered to participate in the research, and therefore, had a more positive disposition to the project. It may be assumed that this attitude would not generalize outside of the sample and may have influenced the findings of the investigation.

Recommendations for Further Research

This study was an effort to investigate the educational administrator's ability to appraise numerous components of the instructional process.

This complex study touched upon many major issues which are worthy of research projects themselves. From this examination, it would appear that further work needs to be done with the Georgia Teacher Assessment evaluation instruments. These instruments have been carefully developed, normed, validated, and used. However, the weighting of the individual criterion items would be a valuable addition to the already

creditable instruments. The criteria would profit if a panel of experts, including university education faculty, classroom teachers, educational consultants, and instructional supervisors, would make some determinations as to the relative importance of the criteria. If this panel utilized a forced ranking selection procedure, the more educationally valued items would emerge. The input could be utilized to assign weights to the criteria and add discrimination power to the instrument.

The present study could be replicated using the same methodology except for a change in the age level of student and the content area used within the simulation exercise. Educators align themselves in areas of expertise with training in elementary or secondary levels, and with curricular specialties such as social studies, history, reading, math, or science. This being the case, many subjects felt uncomfortable because they were forced to evaluate content and age groups outside of their specialty. If treatment groups could be set up assigning subjects to their preferred age level, and with a closer match to their content areas, perhaps the connection between the teacher's planning and execution would have been more fully understood and rating accuracy may have been enhanced.

Another suggestion for consideration might be the replication of the present study but with an expanded variety of subjects involved. This investigation identified four groups of participants which included superintendents, principals, supervisors, and a category for "others" containing mostly teachers and consultants. Future studies should identify all subjects as female or male, and enlarge the category of "others" to

include special education teachers, elementary and secondary teachers, counselors, consultants, and university education faculty. This more refined identification of subjects would provide an opportunity to explore rating accuracy of more diverse groups of educators.

The availability of the Georgia Teacher Assessment Project's accuracy scores provided this study with a somewhat extraordinary comparative measure, and therefore, the major efforts of the investigation were planned around the rater's relationship to that accuracy score. Future research may want to look at other relationships, and these may be: 1) a comparison of rater accuracy by treatments to other raters in the study using interrater reliability measures, 2) a comparison of rater's mean scores, even though the mean score is not the accurate score, by the calculation of variances and standard deviations, and 3) ratings of teacher performance and the interaction, if any, with the study of pre-instructional materials, the use of evaluation instruments, training in teacher appraisal, geographical location, or job position of the appraiser. The determination of interaction effects through examination by analysis of variance would reveal any combinations of variables which contribute significantly to effective teacher evaluation.

Recommendations for Practice

Looking at this study and interpreting the findings opens up the discussion to conjecture. The subjects' general inability to transfer any useful information from the study of the lesson plans to making more accurate judgments of the teaching act is disheartening. It is possible

that the lesson plans were so extensive that the appraisers were unable to assimilate the contents. In discussion with subjects, it was often found that they failed to comprehend what they were reading. For example, in the lesson plans (see Appendix E) an objective stated that the teacher would introduce certain vocabulary words. Appropriately trained educators would quickly note that this educational objective should specify learner outcomes and not teacher action. The fact that this identification was not commonly made would indicate that administrators are not proficient at recognizing such obvious flaws in the teacher's statement of intent, and apparently indicates major deficits in their ability to judge appropriate education objectives.

Another instance, illustrating lack of critical reviewing not related to educational expertise, occurred when studying lesson plans designating the Emperor's New Cothes as the topic for a creative writing assignment. In actuality, the teacher used Peter Pan in the lesson, and yet not one participant in the experiment commented on this incongruity. It may be that administrators cannot process so much information and quickly judge its worth in the time allotted. On the other hand, they may not be reading conscientiously enough to pick out those items which are discrepant. Still, if educational administrators are to play a proactive role in the improvement of instruction, they must become more skillful in the assessment of good teacher planning. This provides the evaluator with the opportunity to improve instruction before students are involved, and possibly before they are victimized by poor teaching. To accomplish this, the appropriate training needs to be obtained by the teacher appraiser.

The professionals responsible for the training of educational administrators have traditionally concentrated on educational administration theory, leadership theory, school laws and other worthwhile subjects. However, this theoretical base has not prepared administrators to function effectively as teacher evaluators. An experiential base of training would be a valuable addition to the existing administrator training program.

For maximum effectiveness, the practicum should provide some input previous to the actual onsite evaluation experience. The needed skills would include those required to determine good planning for student acquisition which would include objective setting and provision for input to accomplish that goal, maintenance of learning through guided and independent practice, and generalization of the learning to increase problem solving in other areas.

Evaluators need to be trained in data gathering techniques. The analysis of teaching must be based upon evidence, and evaluators need data capturing skills to assemble observable and measurable data upon which to base their decisions. If appraisers are inept at the data gathering process, and unskilled in the use of evaluative instruments, the data they have may be invalid for substantiating the assessment of the teacher.

Experiential training would give evaluators input to master these skills and allow for monitored application in an actual appraisal experience. Appraiser trainees would learn how to judge good instructional planning, to capture objective data, to use appropriate measurement

instruments as well as to develop communication skills which will improve their interaction with teachers. Teacher evaluators must be able to convey information in such a way that it is not perceived as criticism against the person, but rather directed solely at the functioning in the classroom. Improvement of instruction will be greatly facilitated if appraisers can present themselves as nonthreatening instructional associates. The complete practicum experience would be under the surveillance of the training program and would allow educational administrator/teacher evaluator trainees to benefit from the continuous input of the program trainer.

This type of training program would create confidence in administrators as they developed those competencies which would allow them to become effective teacher evaluators as well as capable agents for improvement of instruction. Until such training programs are available, it may be wishful thinking to expect administrators to function as accurate appraisers of teachers' preinstructional planning or classroom performance.

REFERENCE NOTES

1. Berliner, David. Successful classroom teaching and learning. Unpublished paper for Educational Forum, 1978. (Available from U.S. Office of Education, Washington, D.C.)

2. Anderson, Joy. Georgia Teacher Assessment Project, Athens, Georgia. Personal telephone communication, April 29, 1980.

3. Teacher Assessment Project. The role of the teacher performance assessment instruments in Georgia's beginning teacher certification program. Unpublished paper, University of Georgia, 1978.

4. Blake, Joy T. Northwest Georgia Regional Assessment Center, Cedartown, Georgia. Personal communication, November 12, 1979.

REFERENCES

Anderson, Harold H. The measurement of domination and of socially integrative behavior in teachers' contacts with children. Child Development, 1939, 10, 73-89.

Bales, Robert F. Interaction process analysis. Reading, Mass.: Addison-Wesley Publishing Co., 1950.

Barr, A. S. and Burton, William H. Supervision. New York, N.Y.: Appleton-Century Co., 1938.

Barr, A. S., Worchester, D. A., Abell, Allen, Beecher, Clarence, Jensen, Leland E., Peronto, Archie, Ringness, Thomas A. and Schmid, John, Jr. Wisconsin studies of the measurement and prediction of teacher effectiveness: A summary of investigations. Journal of Experimental Education, September, 1961, 30, 151-152.

Bernardin, H. John. Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 1978, 63(3), 301-308.

Bloom, B. D. (Ed.). Taxonomy of educational objectives. Handbook I: Cognitive Domain. New York, N.Y.: David McKay, 1956.

Borg, Walter R. and Gall, Meredith D. Educational research. New York, N.Y.: David McKay Company, Inc., 1976.

Borich, Gary D. (Ed.). The appraisal of teaching. Reading, Mass.: Addison-Wesley, 1977.

Borman, Walter C. Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60(5), 556-560.

Borman, Walter C. and Dunnette, Marvin D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60(5), 561-563.

Briggs, Leslie J. Instructional design. Englewood Cliffs, N.J.: Educational Technology Publications, 1977.

Brighton, Strayner F. Increasing your accuracy in teacher evaluation. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1965.

Brody, J. A. The good teacher is harder to define than find. American School Board Journal, July 1977, 164, 15-28.

Cogan, Morris L. Clinical supervision. Boston, Mass.: Houghton Mifflin Co., 1973.

Cogan, Morris L. Rationale for clinical supervision. Journal of Research and Development in Education, 1976, 9(2), 3-19.

Cruickshank, Donald R. Synthesis of selected recent research on teacher effects. Journal of Teacher Education, Spring 1976, 27(1), 57-60.

Davis, Hazel. Evolution of current practices in evaluating teacher competence. In Bruce J. Biddle and William J. Ellena (Eds.), Contemporary Research on Teacher Effectiveness. New York, N.Y.: Holt, Rinehart and Winston, 1964.

Denham, Alice. Clinical supervision: What we need to know about its potential for improving instruction. Contemporary Education, Fall 1977, 49(1), 33-37.

Di Martino, Carl Joseph. Observer effects on teacher verbal behavior as perceived by students. Unpublished dissertation, University of Georgia, 1974.

Doherty, V. M. and Hathaway, W. E. Course goals, general introduction, and introduction to the critique drafts of the course goal collections. Multnomal County Intermediate Educational District, Portland, Oregon Tri-County Goal Development Project, U.S. Department of Health, Education and Welfare, Office of Education. November 1972. ERIC ED 073 552.

Dow, Clyde W. Handbook for supervising teachers. Michigan State University, East Lansing, Michigan, College of Education, 1968. ERIC ED 028 977.

Flanders, N. A. Teacher influence, pupil attitudes and achievement. Cooperative research monograph No. 12. Washington, D.C.: U.S. Government Printing Office, 1965.

Friedman, Aaron. Lesson planning: A basic professional tool. The Clearing House, March 1974, 48(7), 441-443.

Gage, N. L. (Ed.). Handbook of research on teaching. Chicago, Ill.: Rand McNally, 1963.

Georgia State Department of Education. Teacher performance assessment instructions. Atlanta, GA: State Department of Education, 1978.

Goldhammer, Robert. Clinical supervision. New York, N.Y.: Holt, Rinehart, and Winston, 1969.

Goodlad, J. I., Klein, M. F. and Associates. Looking behind the classroom door. Worthington, Ohio: Charles A. Jones Publishing, 1974.

Guthrie, Harold D. and Willower, Donald J. The ceremonial congratulations. The High School Journal, March 1973, 41(6), 289.

Haefele, Donald L. How to evaluate the teacher - Let me count the ways. Phi Delta Kappan, January 1980, 61(5), 349-352.

Harlen, Wynne. Formulating objectives - problems and approaches. British Journal of Education Technology, October 1972, 3(3), 223-235.

Harris, Ben M. Supervisor competence and strategies for improving instruction. Educational Leadership, February 1976, 33(5), 229-232.

Hunter, Madeline. Teacher competency: Problem, theory, and practice. Theory Into Practice, 1976, 15(2), 162-171.

Jersild, Arthur T. and Meigs, Margaret F. Direct observation as a research method. Review of Educational Research, December 1939, 9(5), 473-583.

Johnson, Charles E., Okey, James R., Capie, William, Ellett, Chad, and Adams, Pearce T. Identifying and verifying generic teacher competencies. Athens, Georgia: University of Georgia, 1978.

Joyce, B. R. and Harootunian B. Teaching as problem solving. Journal of Teacher Education, 1964, 15, 420-427.

Justiz, Thomas B. A method for identifying the effective teacher. Unpublished doctoral dissertation, University of California, Los Angeles, 1968.

Kirchner, Wayne K. and Reisberg, Donald J. Difference between better and less effective supervisors in appraisal of subordinates. Personnel Psychology, 1962, 15(3), 295-302.

Klein, Stephen P. and Alkin, Marvin C. Evaluating teachers for outcome accountability. In Gary D. Borich (Ed.), The Appraisal of Teaching. Reading, Mass.: Addison-Wesley Publishing, 1977.

Knezevich, Stephen J. Designing performance appraisal systems. In Gary D. Borich (Ed.), The Appraisal of Teaching. Reading, Mass.: Addison-Wesley Publishing, 1977.

Kowalski, Theodore J. The principal's role in curriculum development: What are the barriers? Contemporary Education, Spring 1979, 50(3), 159-161.

Lamb, Morris L. and Swick, Kevin J. A historical overview of classroom teacher observation. The Educational Forum, January 1975, 39(2), 239-247.

Latham, Gary P., Wexley, Kenneth N., and Pursell, Elliot D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60(5), 550-555.

Lessinger, Leo M. Every kid a winner: Accountability in education. Palo Alto: Science Research Associates, 1970. (a)

Lessinger, Leo M. Engineering accountability for results in public education. Phi Delta Kappan, December 1970, 52(4), 217. (b)

Levin, Benji. Teacher evaluation - A review of research. Educational Leadership, December 1979, 37(3), 240-245.

Lowry H. The relationship of superintendents and principals to the training and professional improvement of their teachers. National Society for the Scientific Study of Education, Seventh Yearbook, Part II. Chicago, Ill.: University of Chicago Press, 1908.

Lucio, William, H. Supervision: Perspectives and propositions. Washington, D.C.: National Education Association, 1967.

Lucio, William H. (Ed.). The supervisor: New demands, new dimensions. Washington, D.C.: Association for Supervisors and Curriculum Development, 1969.

Lucio, William H. Pupil achievement as an index of teacher performance. Educational Leadership, October 1973, 31(1), 71-77.

Lucio, William H. and McNeil, John D. Supervision in thought and action. New York: McGraw-Hill Book Company, 1979.

Mager, Robert F. Preparing instructional objectives. Palo Alto, California: Fearon Publishers, 1962.

Manatt, Richard P., Palmer, Kenneth L., and Hidlebaugh, Everett. Evaluating teacher performance with improved rating scales. NASSP Bulletin, September 1976, 60(401), 21-24.

Masling, Joseph and Stern, George. Effect of the observer in the classroom. Journal of Educational Psychology, 1969, 60(5), 351-353.

McIntyre, John D. Teacher evaluating and the observer effect. NASSP Bulletin, March 1980, 64(434), 36-40.

McNeil, John D. Concomitants of using behavioral objectives in the assessment of teacher effectiveness. Journal of Experimental Education, Fall 1967, 36(1), 69-74.

McNeil, John D. Supervision of instruction: The relationship of theory and practice to accountability. Ohio State Department of Education, Columbus Ohio, Bureau of Elementary and Secondary Education, Washington, D.C., 1972. ERIC ED 064 803.

McNeil, John D. and Popham, W. James. The assessment of teacher competence. In Robert M. W. Travers (Ed.), Second Handbook of Research on Teaching. Chicago, Ill.: Rand McNally and Co., 1973.

McNeil, John D. and Popham, W. James. A critique of widely used criteria in assessing teacher competency. In Gary Borich (Ed.), The Appraisal of Teaching. Reading, Mass.: Addison-Wesley Publishing Co., 1977.

Mitzell, Harold E. Teacher affectiveness. Encyclopedia of Educational Research (3rd edition). New York, N.Y.: The Macmillan Co., 1960.

Moffett, George. Use of instructional objectives in the supervision of student teachers. Unpublished doctoral dissertation, University of California, Los Angeles, 1966.

Morine, Greta. Planning skills: Paradox and parodies. Journal of Teacher Education, 1973, 24(2), 135-141.

Morine, Greta. A study of teacher planning: Beginning teacher evaluation study. Far West Laboratory for Educational Research and Development, San Francisco, California, March, 1976. ERIC ED 146 160.

Morsh, J. and Wilder, E. Identifying the effective instructor: A review of the quantitative studies 1900-1952. Research Bulletin AFPTRC-T-54-54. San Antonio, Texas: Lackland Air Force Base, 1954.

Oliva, Peter F. Supervision for today's schools. New York, N.Y.: Thomas Y. Crowell Company, 1976.

Ott, Lyman, Mendenhall, William and Larson, Richard F. Statistics: A tool for the Social Sciences. North Scituate, Mass.: Duxbury Press, 1978.

Peach, J. W. Accountability in education: A North American viewpoint. In P. K. Chippendale and Paula V. Wilkes (Eds.), Accountability in Education. St. Lucia, Queensland: University of Queensland Press, 1977.

Piper, Terrence J. and Elgart, Denise B. Teacher supervision through behavioral objectives. Baltimore, Maryland: Paul H. Brookes, Publishers, 1979.

Popham, W. James. Behavioral objectives and teaching skills. Forum, October 1969, 8(1), 4-7.

Popham, W. James. Teaching skill under scrutiny. Phi Delta Kappan, June 1971, 52(10), 599-602.

Popham, W. James and Baker, Eva I. Systematic instruction. Englewood Cliffs, N.J.: Prentice-Hall, 1970. (a)

Popham, W. James and Baker, Eva I. Establishing instructional goals. Englewood Cliffs, N.J.: Prentice-Hall, 1970. (b)

Ragosta, Marjorie. Reliability in classroom observation: Observer effects and stability of behavior. Unpublished doctoral dissertation, University of Florida, 1974.

Reavis, Charles A. Clinical supervision: A timely approach. Educational Leadership, February 1976, 33(3), 360-363.

Redfern, George B. How to appraise teaching performance. Columbus, Ohio: School Management Institute, Inc., 1963.

Reemers, H. H. Rating methods in research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago, Ill.: Rand McNally and Co., 1963.

Robinson, John J. The observation report - A help or a nuisance? NASSP Bulletin, March 1978, 62(416), 22-26.

Rosenshine, B. and Furst, N. Research on teacher performance criteria. In B. O. Smith (Ed.), Research in Teacher Education: A Symposium. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Ryans, David G. Characteristics of Teachers. Washington, D.C.: American Council on Education, 1960.

Samph, Thomas. Observer effects on teacher behavior. Unpublished doctoral dissertation, University of Michigan, 1968.

Scriven, Michael. The evaluation of teachers and teaching. In Gary Borich (Ed.), The Appraisal of Teaching. Reading, Mass.: Addison-Wesley Publishing Co., 1977.

Smith, Edward B. and Smith, Sharon Clare. Accountability in the classroom. Contemporary Education, Summer 1976, 47(4), 189-194.

Spears, Harold. Improving the supervision of instruction. New York, N.Y.: Prentice-Hall, Inc., 1953.

Spool, Mark D. Training programs for observers of behavior: A review. Personnel Psychology, Winter 1978, 31(4), 853-888.

Stow, Shirley B. Using effectiveness research in teacher evaluation. Educational Leadership, October 1979, 37(1), 55-58.

Taba, Hilda. Curriculum development, theory and practice. New York, N.Y.: Harcourt, Brace and World, 1962.

Thomas, Donald M. Performance evaluation of educational personnel. Bloomington, Indiana: Phi Delta Kappa Educational Foundation, 1979.

Tolor, Alexander. Evaluation of perceived teacher effectiveness. Journal of Educational Psychology, 1973, 64(1), 98-104.

Tomlinson, Loren R. Recent studies in the evaluation of teachers. Educational Research Bulletin, October 1955, 34, 151-157.

Tuckman, Bruce W., Steber, James M., and Hyman, Ronald T. Judging the effectiveness of teaching styles: The perceptions of principals. Educational Administration Quarterly, Winter 1979, 15(1), 104-115.

Tyler, Ralph W. Basic principles of curriculum and instruction. Chicago, Ill.: University of Chicago Press, 1950.

Tyler, Ralph W. New dimensions in curriculum development. Phi Delta Kappan, 1966, 48, 25-28.

Urban, John. Behavior changes resulting from a study of communicable diseases. New York, N.Y.: Teachers College, Columbia University, 1943.

Van Dalen, Deobold B. Understanding education research. New York, N.Y.: McGraw-Hill Book Company, 1966.

Wicks, Larry E. Opinions differ: Teacher evaluation. Today's Education, 1973, 63(3), 42-43.

Williams, Ray C. In the beginning . . . goals. Theory Into Practice, April 1976, 15(2), 86-89.

Worcester, D. A. Some assumptions, explicitly and implicitly made in the investigations here summarized. Wisconsin Studies of the Measurement and Prediction of Teacher Effectiveness: A Summary of Investigations. Madison, WI.: Dembar Publication, Inc., 1968.

Zahorik, John A. The effects of planning on teaching. The Elementary School Journal, December 1970, 71(3), 143-151.

APPENDIX A: TPM EVALUATION INSTRUMENT

TEACHER INSTRUCTIONAL PLANS AND MATERIALS ASSESSMENT SCALE

Please do not mark on this assessment scale. A rating form has been enclosed for that purpose.

The following criteria are to be rated after a thorough study of the lesson plans and materials (worksheets, etc.).

1. Do the lesson plans clearly and specifically state appropriate objectives indicating what the learner will be able to do after successful mastery of the instruction?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
No objectives listed or are teacher behaviors.	Objectives listed are inappropriate or in broad terms.	Objectives with a few exceptions are appropriate to topics & learners.	All objectives are appropriate to topic and learners.	In addition to #4, objectives reflect a sequence for the objectives.

2. Do the lessons plans specify procedures to appropriately accomplish objectives of the lesson?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
Plans do not specify teaching procedures.	Teaching procedures are specified but some seem inappropriate for objectives and/or learners or are not referenced.	Teaching procedures are referenced to objectives. Most are appropriate for objectives and/ or learners.	Teaching procedures are referenced to objectives. All are appropriate for objectives and/or learners	In addition to #4, a variety of appropriate procedures ranging from teacher-centered to learner centered approaches are indicated.

3. Do the lesson plans specify appropriate instructional content, materials, and media for accomplishing the objectives of the lesson?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
No content, materials, or media are listed.	Content materials and media are listed but not referenced to learning activities.	Content materials and media are listed and referenced to intended objectives and learning.	In addition to #3, many appropriate resources have been included.	In addition to #4, the teacher has included an imaginative use of available resources or original materials.

4. Do the lesson plans specify appropriate materials and plans for assessing the learners mastery of the objectives?

Check one of the following ratings:

1.	2.	3.	4.	5.
No assessment plans or materials are included in the instructional unit.	Tests to assess learners are included in the unit; however many of the items are inappropriate to the instructional objectives and/or the learners.	Tests are included and are appropriate for evaluating the objectives and the learners.	In addition to #3, procedures such as interviews, questionnaires, or self-tests, are included to assess learners and/or objectives.	In addition to #4, plans and material are included to assess attitudes of learners toward the topic.

5. Do the lesson plans outline instruction that is planned to accommodate a variety of cognitive levels?

Descriptions for the above.

A)	B)	C)	D)
Learners have an opportunity to acquire factual information and to explain or summarize.	Learners have an opportunity to apply the information to particular situation.	Learners have an opportunity to identify and clarify parts of complex ideas or synthesize knowledge by integrating information.	Learners have an opportunity to judge the value and importance of ideas and information.

Check one of the following ratings:

1.	2.	3.	4.	5.
None of the descriptions is evident.	One of the descriptions is evident.	Two to the descriptions were evident.	Three of the descriptions were evident.	Four of the descriptions were evident.

6. Do the lesson plans outline instruction that is organized to accommodate the differences in learner capabilities?

Descriptions for the above.

A)	B)	C)	D)
Current instruction is presented initially to entire group at a success level and then individualized capabilities.	Assignments for learners are differentiated according to ability.	Materials compatible with the range of learner abilities are available to achieve a given objective.	Remedial and/or enrichment activities are available for the unit.

Check one of the following ratings:

1.	2.	3.	4.	5.
None of the descriptions is evident.	One of the descriptions is evident.	Two of the descriptions are evident.	Three of the descriptions are evident.	Four of the descriptions are evident.

7. Do the lesson plans indicate instruction that is organized to accommodate the differences among learners in their learning styles?

Descriptions for the above.

A)	B)	C)	D)
Alternative resources are available for achieving a given objective.	Alternative means (group structure or teacher presentation methods) are available for achieving a given objective.	Learners are given options in responding to a given assignment (e.g., writing, drawing, verbalizing).	Learners are matched with resources and procedures for a portion of the instructional sequence.

Check one of the following ratings:

1.	2.	3.	4.	5.
None of the descriptions is evident.	One of the descriptions is evident.	Two of the descriptions are evident.	Three of the descriptions are evident.	Four of the descriptions are evident.

8. Do the lesson plans indicate instruction that is organized to accommodate differences among learners in their rates of learning?

Description for the above.

A)	B)	C)	D)
Learners may work at their own rate for some of the objectives some of the time.	Special assistance (e.g., teacher or peer tutoring) is available.	Fast learners are allowed to work on topic-related enrichment activities some of the time.	Objectives are divided between those that all learner can achieve; those that are enrichment objectives which are desirable but not essential.

Check one of the following racings:

1.	2.	3.	4.	5.
None of the descriptions is evident.	One of the descriptions is evident.	Two of the descriptions is evident.	Three of the descriptions is evident.	Four of the descriptions is evident.

9. Do the lesson plans or material indicate teacher-made or teacher-selected evaluation material that are to be used to obtain information about learner progress?

Check one of the following ratings:

1.	2.	3.	4.	5.
No evaluations of learner progress are indicated.	Only end-of-unit or summative evaluations are used to determine learner progress.	In addition to #2, additional progress checks are used to determine learner progress.	In addition to #3, evaluations indicate individual learner progress on specific objectives.	In addition to #3 & #4, the teacher uses affective measures to assess learner attitudes.

APPENDIX B: CP EVALUATION INSTRUMENT

PACKET TWO

You have completed the first portion of the program and now are ready to participate in the viewing and evaluation of a videotaped instructional lesson.

Included in Packet Two are the following:

1) The Classroom Procedure Assessment Scale which you will use to evaluate the lesson. Take a few minutes to familiarize yourself with these items before you view the videotape.

2) Remember, you will not be marking directly on the assessment scale. A rating sheet is provided for recording your rating of the criteria items. Also included are a few personal information questions for you to complete.

After viewing the videotape, complete all items on the rating sheet. When you have finished, retain one copy of the rating sheet for your information. Return the remainder of the materials to the envelope and return to the program leader.

You will be given an opportunity to review your responses with the group and the program leader.

CLASSROOM PROCEDURE ASSESSMENT SCALE

Please do not mark on this assessment scale. A rating form has been enclosed for that purpose.

The following criteria are to be rated after viewing the videotaped instructional session.

1. Does the teacher use audiovisual and other instructional equipment?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
Audiovisual (e.g., projectors) or other (e.g., duplicator, transparency maker) equipment that is available and appropriate is not used.	Occasionally uses available equipment. Has trouble with it when doing so. Equipment used does not always fit planned lesson.	Usually uses audiovisual equipment at appropriate times in lesson. Shows evidence of mastering the operation of most equipment.	Highly skillful use of audiovisual equipment at appropriate times. Media presented blends smoothly with other kinds of instruction.	In addition to the items in #4, the teacher shows evidence of skillfully preparing original materials for use with audiovisual or other equipment.

2. Does the teacher use instructional materials that provide appropriate practice on the objectives?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
Materials chosen are irrelevant to the topic or objectives.	Materials chosen are related to the topics being studied but not to the objectives.	Most materials chosen provide for practice on specific objectives. Some of the practice may be insufficient in quantity to	Materials chosen are relevant to the objectives. Learners are given ample opportunity to practice and achieve the objectives.	In addition to the items in #4, the teacher uses formal or informal progress assessment techniques to determine whether the practice

parency maker) equipment that is available and appropriate is not used.	when doing so. Equipment used does not always fit planned lesson.	Shows evidence of mastering the operation of most equipment.	times. Media presented blends smoothly with other kinds of instruction.	fully preparing original materials for use with audio-visual or other equipment.

2. Does the teacher use instructional materials that provide appropriate practice on the objectives?

Check one of the following ratings:

1.	2.	3.	4.	5.
Materials chosen are irrelevant to the topic or objectives.	Materials chosen are related to the topics being studied but not to the objectives.	Most materials chosen provide for practice on specific objectives. Some of the practice may be insufficient in quantity to achieve the objectives.	Materials chosen are relevant to the objectives. Learners are given ample opportunity to practice and achieve the objectives.	In addition to the items in #4, the teacher uses formal or informal progress assessment techniques to determine whether the practice individual learners receive is sufficient.

3. Does the teacher organize learning activities in a logical sequence?

Check one of the following ratings:

1.	2.	3.	4.	5.
Activities used in the classroom are unrelated to one another or to the objectives.	Activities relate to a common topic within lessons but many activities seem out of sequence.	Activities are arranged to present ideas so that one builds on another. Only occasionally are these problems of sequence.	No instances of problems in sequencing learning activities are noted.	In addition to the items in #4, activities provide an opportunity to acquire prerequisites if learners have not already done so.

4. Does the teacher demonstrate ability to conduct lessons using a variety of teaching methods?

Teaching methods such as the following may be observed: drill, inquiry, discussion, role playing, demonstration, problem solving, etc.

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
No teaching method is used acceptably.	One method is used acceptably.	Two methods are used acceptably.	Three methods are used acceptably.	Four methods are used acceptably.

5. Does the teacher use teaching methods appropriate for objectives and learners?

Descriptions for the above.

A. Teaching methods are matched to objectives.

B. Teaching methods are matched to learners.

C. A smooth transition is made from one method to another.

D. The classroom environment is arranged for the teaching method.

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
None of the descriptors is evident.	One descriptor is evident.	Two descriptors are evident.	Three descriptors are evident.	Four descriptors are evident.

6. Does the teacher demonstrate ability to work with individuals, small groups, and large groups?

Descriptions for the above.

B. Teaching methods are matched to learners.

C. A smooth transition is made from one method to another.

D. The classroom environment is arranged for the teaching method.

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
None of the descriptors is evident.	One descriptor is evident.	Two descriptors are evident.	Three descriptors are evident.	Four descriptors are evident.

6. Does the teacher demonstrate ability to work with individuals, small groups, and large groups?

Descriptions for the above.

A. The group size for instruction is matched to the objective.

B. The teacher's role is appropriate to the group size being used.

C. The changes from one group size to another are smooth.

D. Different group sizes that are matched to the objectives are used.

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
None of the descriptors is evident.	One descriptor is evident.	Two descriptors are evident.	Three descriptors are evident.	Four descriptors are evident.

7. Does the teacher use procedures which initially get learners involved in lessons?

Descriptions for the above.

A. Helps learners recall past experiences or knowledge.

B. Uses interests of learners as a link to new activities.

C. Stimulates interest in new activities with such things as discrepant events or one of thought-provoking questions.

D. Helps learners understand what they may achieve by participating in the new activities.

Check one fo the following ratings:

1._____	2._____	3._____	4._____	5._____
None of the descriptors is evident.	One descriptor is evident.	Two descriptors are evident.	Three descriptors are evident.	Four descriptors are evident.

8. Does the teacher maintain learner involvement in lessons?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
Most learners(75%) seem to be off task (learners are not attending to the teacher, materials, or other appropriate foci for an activity).	Many learners(50%) are on task but the remainder show only superficial attention to lessons.	Most learners(75%) are on (activity engaged in the lesson).	High on-task(90%) behavior is evident.	All learners are on task.

8. Does the teacher maintain learner involvement in lessons?

Check one of the following ratings:

1.	2.	3.	4.	5.
Most learners(75%) seem to be off task (learners are not attending to the teacher, materials, or other appropriate foci for an activity).	Many learners(50%) are on task but the remainder show only superficial attention to lessons.	Most learners(75%) are on (activity engaged in the lesson).	High on-task(90%) behavior is evident.	All learners are on task.

9. Does the teacher demonstrate knowledge in the subject area?

Check one of the following ratings:

1.	2.	3.	4.	5.
The knowledge demonstrated in the classroom is inaccurate or out-of-date.	Does not demonstrate knowledge of the subject area when it would be appropriate to do so.	Subject area knowledge demonstrated is accurate and related to the subject.	Demonstrates accurate and up-to-date knowledge of the subject to arouse the interest of the learners.	In addition to the items in #4, uses knowledge of the subject area to show learners the importance of the topic in the world outside the classroom.

10. Does the teacher present accurate information about the topic being taught?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
Information presented to learners is clearly inaccurate or the topic is inconsistent with what is currently appropriate for the level taught.	Some information is inaccurate or inconsistent with what is currently considered appropriate.	Information on the topic is accurate. Sources of information and learning materials are timely.	Regular text or other curriculum material is supplemented with recent newspaper, journal, or other information.	A variety of resources are used to make classroom information up to date. The teacher helps learners understand how information (knowledge, values, procedures, etc.) in the topic area has changed over the years.

11. Does the teacher provide learners with information on the purpose and importance of topics?

Check one of the following ratings:

1._____	2._____	3._____	4._____	5._____
Does not designate the purpose or importance of a topic.	Fails to relate specific topics to their purpose or importance in a content area.	The purpose or importance of most topics studied is conveyed to learners.	Topics are taught in context. The teacher shows how topics are but a portion of a larger content area.	In addition to the items in #4, the teacher encourages (or provides opportunities for) learners to either question or relate to specific topics which are important to a content area.

APPENDIX C: SCORING RATIONALE

SCORING RATIONALE FOR INTERMEDIATE LANGUAGE ARTS - PREINSTRUCTIONAL MATERIALS

Indicator	Panel rating	Suggested rating	Rationale
TPM 1	2.3	2	Objectives are listed on the worksheets from the basal reader. Other objectives are listed by group for each day. Some are specific, Level 14, Unit 5, #3 for Thursday. Most are written in broad terms, more like an activity than a specific behavior outcome. The objectives are appropriate for the topic and learners.
2	2.7	3	Procedures are listed in very general terms; specific student activities/objectives are then listed by day and group. Procedures for worksheet objectives are general--"students will do skill sheets." Most procedures seem appropriate.
3	3.3	3	Materials and media are listed. Many were reluctant to rate 4 ("full use") since they didn't know what materials were available.
4	3.3	3	Tests were referred to. The test on "Karken" is planned for Thursday and included in the portfolio. The spelling words for the test were not listed. The worksheets were treated as tests by some. Interviews and questionnaires were not included.
5	3.9	4	Descriptor d was rarely checked. Students did not have opportunities to judge or evaluate within the different activities ("procedures") used.
6	4.0	4	Instruction was designed to meet the different reading levels; grouping facilitated this. The interview infers the use of enrichment activities; specifics are not noted in the plans, so Descriptor d should not be rated.

Indicator	Panel rating	Suggested rating	Rationale
7	3	3	One evidence for Descriptor d could be the use of the filmstrip with the Title I students and the writing and reading of their own endings for the story. Within an objective or activity individuals were not given response options --all wrote a summary or ending, all would spell (did not state that some would orally spell on the test). Multiple activities were available for each day, but it appears that each student does each activity. The combinations of descriptors checked were inconsistent.
8	3.0	3	Within each group it appears that each student does each objective. Descriptor d was rarely chosen. Students were able to go to gifted; slower students were listed as Title I. Within the activities described, special assistance from the teacher was not evident. Descriptors a and c were consistently checked; however, evidence for Descriptor c was stated in the interview but not described in the portfolio.
9	2.7	3	Posttests, end-of-week spelling tests, etc., were listed. Most people counted the skill sheets as progress checks. Individual progress on objectives is not noted.

SCORING RATIONALE FOR CLASSROOM PROCEDURES CRITERIA

Project rating	Suggested rating	
2.8	3.0	Criterion 1: Uses AV and other equipment. Rationale: The chalkboard was used, and a tape player was used but it was too loud: earphones weren't used. There was little evidence of original materials prepared by the teacher. Evidence for assessing mastery of most equipment was limited.
3.4	3.0	Criterion 2: AV equipment provides appropriate practice on the objective. Rationale: Some objectives had only limited practice. Some problems were encountered with matching materials to the skill objectives.
3.5	3/4	Criterion 3: Learning activities in a logical sequence. Rationale: Activities tended to build on each other. Occasionally, however, the activities could have been sequenced better.
4.0	4	Criterion 4: Uses a variety of teaching methods. Rationale: Methods demonstrated included some inquiry, discussion, and lecture (on clipped words).
4.0	3/4	Criterion 5: Methods appropriate for objective and learners. Rationale: Teaching methods seemed to be matched to the specified objective and learners. The learners and furniture were arranged for small group activities; use of tape was distracting.
3.5	3/4	Criterion 6: Ability to work with individuals, small groups, and large groups. Rationale: Group size was matched to the objective. Teacher managed the three groups successfully. He related to the groups. Changes in group size were not shown.

Project rating	Suggested rating	
2.7	3	Criterion 7: Procedures got learners involved in lesson. Rationale: Thought provoking questions were involved in the discussion prior to reading "Fast Sooner Hound". Hangman game stimulated learning of new vocabulary. Teacher mentioned previous study of clipped words; some used this to rate Descriptor. A. Learner interests were not of major importance in presenting the new activities.
3.5	3/4	Criterion 8: Maintain learner involvement in lesson. Rationale: Most students seemed to be concentrating on the assigned tasks. Students talked among themselves; hard to tell if they were discussing their work and lesson.
3.6	3/4	Criterion 9: Knowledge of subject matter demonstrated. Rationale: Knowledge demonstrated was usually accurate. A few instances where things needed an explanation were not followed up by the teacher, e.g., the need to clarify what a summary is. The teacher didn't really use his knowledge to interest each group.
2.8	2/3	Criterion 10: Present accurate information. Rationale: Use of locale and local were confused. Definitions for sooner were not clear. Supplemental materials could have been used to check the definitions. Most of the panel felt the majority of the information was accurate.
2.0	2	Criterion 11: Provides information on purpose and importance of the topic and lesson. Rationale: Other than preparing for the spelling contest, the importance of topic was not demonstrated.

APPENDIX H: RATING FORM FOR PREINSTRUCTIONAL MATERIALS INSTRUMENT

Rating sheet for the Teachers Instructional Plans and Materials Assessment Scale.

Circle the number which is your best answer. A 1 indicates a low rating and a 5 is the highest rating.

	Low				High
1.	1	2	3	4	5
2.	1	2	3	4	5
3.	1	2	3	4	5
4.	1	2	3	4	5
5.	1	2	3	4	5
6.	1	2	3	4	5
7.	1	2	3	4	5
8.	1	2	3	4	5
9.	1	2	3	4	5

If you do not wish your responses to be released for research purposes, please check here. _____

APPENDIX I: RATING FORM FOR CLASSROOM PROCEDURES INSTRUMENT

Rating Sheet for the Classroom Procedures Assessment Scale

Circle the number which is your best answer. A 1 indicates a low rating and a 5 is the highest rating.

	Low				High
1.	1	2	3	4	5
2.	1	2	3	4	5
3.	1	2	3	4	5
4.	1	2	3	4	5
5.	1	2	3	4	5
6.	1	2	3	4	5
7.	1	2	3	4	5
8.	1	2	3	4	5
9.	1	2	3	4	5
10.	1	2	3	4	5
11.	1	2	3	4	5

Please complete the following personal information.

1. Have you had previous training in teacher appraisal? Yes_____ No_____
   If yes, please specify the following: Where trained_____
   Length of training period
   1 day or less_____
   1-3 days_____
   3-5 days_____
   more than 5 days_____
   component of a university course_____

2. What is your educational job description? Superintendent_____
   Principal_____
   Supervisor_____
   Other_____
   (please specify)

If you do not wish your responses to be released for research purposes, please check here._____

This is an Iowa State University field test. For further information you may contact the following:

Mrs. Sally Frudden
University of Northern Iowa
Cedar Falls, Iowa 50613
phone (O) 319-273-6064
phone (H) 319-228-2074

APPENDIX J: TABLES

Table J.1. Distributions of participants by treatment group

	Frequency	Percent	Cumulative percent
Group 1 (Placebo)	139	26.3	26.3
Group 2 (Lesson plans only)	136	25.7	52.1
Group 3 (Lesson plans and high inference)	126	23.8	75.9
Group 4 (Lesson plans and low inference)	127	24.0	100.0
	1[a]	0.2	
	529	100.0	100.0

[a]Missing cases = 1.

Table J.2. Standard deviations by treatment groups on accuracy of teacher performance (ATP) and accuracy of instructional materials (AIM)

	1	2	3	4
ATP 1	.8120	.7907	.7859	.7885
ATP 2	.6794	.5962	.6220	.6610
ATP 3	.9129	.8352	.8618	.9266
ATP 4	.9689	.9151	.6200	.9986
ATP 5	.9913	.9481	.9958	.0499
ATP 6	.9120	.9076	.9084	.9859
ATP 7	.7461	.7120	.6538	.7506
ATP 8	.8994	.8749	.8903	.8708
ATP 9	.8251	.7205	.8167	.9294
ATP 10	.4383	.3338	.5162	.5128
ATP 11	.6657	.6541	.5821	.6532
AIM 1			1.0386	.9737
AIM 2			.7526	.7943
AIM 3			.7310	.7675
AIM 4			.6954	.8704
AIM 5			1.0524	1.1394
AIM 6			1.0068	1.1602
AIM 7			.7692	.8586
AIM 8			.7642	.7990
AIM 9			.7175	.7938

Table J.3. Composite scores for rating teacher performance

Criteria	Composite scores
1	2.16
2	2.82
3	2.71
4	2.95
5	2.79
6	2.95
7	2.83
8	2.59
9	2.99
10	2.67
11	2.22

Table J.4. Composite scores for rating preinstructional materials

Criteria	Composite scores
1	2.91
2	3.02
3	3.15
4	2.40
5	3.02
6	3.10
7	2.56
8	2.44
9	2.29

Table J.5. Distribution of participants by training

	Frequency	Percent	Cumulative percent
1	165	31.2	31.2
2	66	12.5	43.7
3	90	17.0	60.7
4	23	4.3	65.0
5	35	6.6	71.6
6	150	28.4	100.0
Total	529	100.0	

Table J.6. Standard deviation by training on accuracy of teacher performance (ATP) and accuracy of instructional materials (AIM)

	1	2	3	4	5	6
ATP 1	.8145	.8491	.7789	.7674	.7470	.7812
ATP 2	.6529	.5616	.6896	.5898	.5071	.6783
ATP 3	.8725	.8473	.8771	.8592	.8506	.9310
ATP 4	.9526	1.0072	1.0031	1.0426	.9684	1.0003
ATP 5	.9459	1.0247	1.0044	1.0292	.9762	1.0375
ATP 6	.9019	.8493	.9731	1.0108	.9730	1.4440
ATP 7	.9455	.7424	.8222	.6957	1.1429	1.0067
ATP 8	.8464	.8171	.9224	.7603	.7960	.9447
ATP 9	.8345	.8025	.7827	.5165	.9337	.8663
ATP 10	.5318	.2591	.4740	.3875	.4260	.4529
ATP 11	.6295	.5905	.7004	.6503	.5827	.6536
AIM 1	.7508	.8457	.8767	.6503	.8382	.7958
AIM 2	1.0798	1.2265	1.1480	1.2961	1.0396	1.1194
AIM 3	1.1338	1.2832	1.2138	1.2028	1.1754	1.1898
AIM 4	1.0609	1.0665	1.1974	1.1541	1.1099	1.0281
AIM 5	1.5180	1.6101	1.6438	1.6418	1.5568	1.5119
AIM 6	1.5491	1.6195	1.6087	1.6225	1.5448	1.5364
AIM 7	1.0568	1.1162	1.1367	1.1795	1.0784	1.0790
AIM 8	1.0513	1.1299	1.1007	1.1264	1.0331	1.0560
AIM 9	1.0245	1.1023	1.0455	.9261	1.0508	.9024

Table J.7. Distribution of participants by jobs

	Frequency	Percent	Cumulative percent
1	17	3.2	3.2
2	261	49.3	53.0
3	51	9.6	62.7
4	196	37.0	100.0
Missing	4	0.8	100.0
Total	529	100.0	

Table J.8. Standard deviations by job on accuracy of teacher performance (ATP) and accuracy of instructional materials (AIM)

	1	2	3	4
ATP 1	.8090	.7907	.6580	.8142
ATP 2	.6183	.6287	.6739	.6541
ATP 3	1.0308	.8702	.8496	.9092
ATP 4	.9393	.9826	.9766	.9929
ATP 5	.8090	1.0402	1.0059	.9358
ATP 6	1.0388	.9031	.9611	.9394
ATP 7	.8314	.7298	.6934	.7038
ATP 8	1.0388	.8778	.9531	.8582
ATP 9	.7952	.7764	.9091	.8652
ATP 10	.0000	.4438	.4881	.4747
ATP 11	.5145	.6756	.6161	.6366
AIM 1	1.0326	.7688	.5881	.8545
AIM 2	1.0914	1.1457	1.0825	1.1138
AIM 3	1.1311	1.1942	1.1782	1.1836
AIM 4	.9275	1.0892	1.1369	1.0688
AIM 5	1.7945	1.5907	1.3602	1.5270
AIM 6	1.6869	1.5917	1.4692	1.5389
AIM 7	1.0607	1.1074	1.1716	1.0542
AIM 8	1.0290	1.0834	1.0969	1.0555
AIM 9	.9275	1.0032	.9583	1.0055

Table J.9. Distribution of participants by region

	Frequency	Percent	Cumulative percent
1	64	12.1	12.1
2	49	9.3	21.4
3	111	21.0	42.3
4	107	20.2	62.6
5	66	12.5	75.0
6	28	5.3	80.3
7	32	6.0	86.4
8	72	13.6	100.0
Total	529	100.0	

Table J.10. Standard deviations by region on accuracy of instructional materials (AIM) and accuracy of teacher performance (ATP)

	1	2	3	4	5	6	7	8
AIM 1	.7358	.6876	.8520	.8727	.8015	.7445	.7976	.7438
AIM 2	1.2080	.9742	1.2124	1.1285	1.0013	1.3313	1.0758	1.0480
AIM 3	1.2614	1.2664	1.1564	1.1847	1.2137	1.2013	1.2276	1.0845
AIM 4	1.1144	1.2048	1.1313	1.0562	1.0688	1.1841	.9311	.9665
AIM 5	1.5436	1.4410	1.5822	1.6550	1.5369	1.3589	1.5850	1.5428
AIM 6	1.5124	1.5846	1.5942	1.6143	1.6526	1.4525	1.5398	1.4679
AIM 7	1.0521	1.0314	1.1118	1.1198	.9617	1.2615	1.1496	1.1069
AIM 8	1.0671	1.0055	1.0903	1.0888	1.0316	1.1007	1.1622	1.0614
AIM 9	1.0465	.8552	.9953	.9430	1.0610	.8968	1.2728	1.0020
ATP 1	.6234	.6845	.8177	.8545	.7400	.7005	.7620	.7645
ATP 2	.6154	.6974	.6628	.6602	.5866	.6785	.6148	.6050
ATP 3	.9167	.9117	.8394	.8601	.9509	.8219	.7605	.9579
ATP 4	1.0661	1.1138	.9114	.9159	1.0323	.8997	.8175	.9176
ATP 5	.9739	1.0392	.9968	.9349	1.0323	.8751	1.0659	.8863
ATP 6	.9210	1.0289	.9006	.8825	1.0209	.8719	.8542	.9037
ATP 7	.7584	.6611	.6904	.6941	.6768	.8189	.7344	.7609
ATP 8	.8428	.9727	.8776	.8964	.8377	.8122	.9419	.8419
ATP 9	.8751	.8947	.8047	.6826	.9167	.9215	.8322	.7390
ATP 10	.1875	.4150	.3663	.5530	.4966	.2835	.4442	.5810
ATP 11	.6099	.6669	.5982	.6670	.5309	.6372	.6927	.7071

Table J.11. Means, standard deviations, and numbers of the four treatment groups by criteria group on accuracy of teacher performance

	Group I			Group II			Group III			Group IV		
ATP	$\bar{X}_1$	S_1	N_1	$\bar{X}_2$	S_2	N_2	$\bar{X}_3$	S_3	N_3	$\bar{X}_4$	S_4	N_4
1	1.01	.83	139	.93	.79	136	.88	.79	126	1.04	.79	127
2	.70	.70	139	.51	.60	136	.60	.64	126	.65	.66	127
3	.82	.91	139	.61	.84	135	.83	.86	124	.83	.92	127
4	1.06	.97	139	1.15	.91	136	1.20	1.02	126	1.47	1.00	127
5	1.27	.99	139	1.36	.95	136	1.39	1.00	126	1.41	1.05	127
6	.83	.91	139	.81	.91	135	.90	.91	126	1.02	.99	127
7	.96	.75	139	.84	.71	136	.86	.65	126	1.01	.75	127
8	.86	.90	139	.82	.87	136	.87	.89	126	.91	.87	127
9	.58	.83	139	.43	.72	136	.55	.82	125	.72	.93	127
10	.16	.52	139	.09	.41	135	.23	.59	125	.22	.57	127
11	.75	.70	139	.70	.71	136	.60	.58	126	.68	.65	127

Table J.12. Analysis of variance on ATP 1 (use of AV)

Source	df	SS	Mean square	F	F Prob.
Between	3	2.2086	.7362	1.157	.3254
Within	524	333.3036	.6361		
Total	527	225.5120			

	t value	df	t Prob.
Contrast 1 (3-1-1-1)	0.766	524	.444
Contrast 2 (2-1-1)	-0.357	524	.721
Contrast 3 (-1 1)	1.658	524	.098

Table J.13. Analysis of variance on ATP 2 (practice on objectives)

Source	df	SS	Mean square	F	F Prob.
Between	3	2.6660	.8887	2.993	.1002
Within	524	222.5131	.4246		
Total	527	225.1791			

	t value	df	t Prob.
Contrast 1 (3-1-1-1)	0.766	524	.444
Contrast 2 (2-1-1)	-0.357	524	.721
Contrast 3 (-1 1)	1.658	524	.098

Table J.14. Analysis of variance on ATP 3 (logical sequence)

Source	df	SS	Mean square	F	F Prob.
Between	3	4.5257	1.5086	1.926	.1243
Within	521	408.0205	0.7831		
Total	524	412.5461			
			t value	df	t Prob.
Contrast (3-1-1-1)			0.701	521	.484
Contrast (2-1-1)			-2.286	521	.023*
Contrast (-1 1)			-0.071	521	.944

*$p < .05$.

Table J.15. Analysis of variance on ATP 4 (variety of methods)

Source	df	SS	Mean square	F	F Prob.
Between	3	12.5245	4.1748	4.390	.0046
Within	524	498.2881	0.9509		
Total	527	510.8125			
			t value	df	t Prob.
Contrast 1 (3-1-1-1)			-2.232	524	.026*
Contrast 2 (2-1-1)			-1.817	524	.070
Contrast 3 (-1 1)			2.235	524	.026*

*$p < .05$.

Table J.16. Analysis of variance on ATP 5 (appropriate methods)

Source	df	SS	Mean square	F	F Prob.
Between	3	1.5632	.5211	.525	.6651
Within	524	519.7801	.9919		
Total	527	521.3430			
			t value	df	t Prob.
Contrast 1 (3-1-1-1)			-1.173	524	.241
Contrast 2 (2-1-1)			-0.404	524	.686
Contrast 3 (-1 1)			.227	524	.820

Table J.17. Analysis of variance on ATP 6 (group sizes)

Source	df	SS	Mean square	F	F Prob.
Between	3	3.2577	1.0859	1.260	.2874
Within	523	450.7712	.8619		
Total	526	454.0288			
			t value	df	t Prob.
Contrast 1 (3-1-1-1)			-0.852	523	.395
Contrast 2 (2-1-1)			-1.430	523	.153
Contrast 3 (-1 1)			1.019	523	.309

Table J.18. Analysis of variance on ATP 7 (learner involved)

Source	df	SS	Mean square	F	F Prob.
Between	3	2.6513	.8838	1.717	.1624
Within	524	269.6790	.5147		
Total	527	272.3301			
			t-value	df	t Prob.
Contrast 1 (3-1-1-1)			0.888	524	.375
Contrast 2 (2-1-1)			-1.236	524	.217
Contrast 3 (-1 1)			1.671	524	.095

Table J.19. Analysis of variance on ATP 8 (maintains involvement)

Source	df	SS	Mean square	F	F Prob.
Between	3	0.5866	.1955	.250	.8612
Within	524	409.5920	.7817		
Total	527	410.1785			
			t value	df	t Prob.
Contrast 1 (3-1-1-1)			-0.129	524	.897
Contrast 2 (2-1-1)			-0.759	524	.448
Contrast 3 (-1 1)			0.399	524	.690

Table J.20. Analysis of variance on ATP 9 (knowledge of subject)

Source	df	SS	Mean square	F	F Prob.
Between	3	5.5935	1.8645	2.742	.0426
Within	523	355.5789	0.6799		
Total	526	361.1724			
			t value	df	t Prob.
Contrast 1 (3-1-1-1)			0.114	523	.909
Contrast 2 (2-1-1)			-2.326	523	.020*
Contrast 3 (-1 1)			1.660		

*p < .05.

Table J.21. Analysis of variance on ATP 10 (accurate information)

Source	df	SS	Mean square	F	F Prob.
Between	3	1.5089	.5030	1.820	.1424
Within	522	144.2462	.2763		
Total	525	145.7551			
			t value	df	t Prob.
Contrast 1 (3-1-1-1)			-0.330	522	.741
Contrast 2 (2-1-1)			-2.313	522	.021*
Contrast 3 (-1 1)			-0.173	522	.863

*p < .05.

Table J.22. Analysis of variance on ATP 11 (purpose of topic)

Source	df	SS	Mean square	F	F Prob.
Between	9	1.5978	.5326	1.198	.3099
Within	524	232.9448	.4446		
Total	527	234.5426			
			t value	df	t Prob.
Contrast 1 (3-1-1-1)			1.384	524	.167
Contrast 2 (2-1-1)			0.879	524	.380
Contrast 3 (-1 1)			0.977	524	.329

Table J.23. Chi-square analysis on AIM 1 (appropriate objectives) by Tatrain

	1	2	3	4	5	6
0.00	17	13	15	2	5	23
1.00	27	8	20	1	5	22
2.00	10.9	43	47	19	22	99
3.00	12	2	8	1	3	6
	6 out of 24 (25%) have expected frequencies less than 5.					

Table J.24. Chi-square analysis on AIM 2 (procedure for objectives) by Tatrain

	1	2	3	4	5	6
0.00	17	13	13	5	2	18
1.00	34	15	24	3	11	33
2.00	21	6	11	1	4	16
3.00	93	32	42	14	18	83
x^2 = 12.3055			df = 15		Probability = .62	

Table J.25. Chi-square analysis on AIM 3 (specific content) by Tatrain

	1	2	3	4	5	6
0.00	20	16	17	3	5	24
1.00	39	14	22	6	9	32
2.00	13	4	8	0	3	11
3.00	93	32	43	14	18	83
$x^2 = 9.3053$			df = 15		Probability = .86	

Table J.26. Chi-square analysis on AIM 4 (appropriate for mastery) by Tatrain

	1	2	3	4	5	6
0.00	15	6	18	3	4	10
1.00	36	18	13	4	8	38
2.00	21	10	16	2	5	18
3.00	93	32	43	14	18	84
$x^2 = 18.0972$			df = 15		Probability = .26	

Table J.27. Chi-square analysis on AIM 5 (cognitive levels) by Tatrain

	1	2	3	4	5	6
0.00	19	11	17	4	3	17
1.00	26	11	14	2	10	23
2.00	14	10	11	2	3	18
3.00	12	2	4	1	1	8
4.00	94	32	44	14	18	84

9 out of 30 (30%) have expected cell frequencies less than 5.

Table J.28. Chi-square analysis on AIM 6 (learner capabilities) by Tatrain

	1	2	3	4	5	6
0.00	19	10	14	3	5	20
1.00	31	15	19	4	4	19
2.00	11	6	12	2	5	19
3.00	9	3	2	0	3	8
4.00	95	32	43	14	18	84

9 out of 30 (30%) have expected cell frequencies less than 5.

Table J.29. Chi-square analysis on AIM 7 (learning styles) by Tatrain

	1	2	3	4	5	6
0.00	16	8	13	3	4	17
1.00	31	17	19	5	6	26
2.00	24	9	14	1	7	24
3.00	94	32	44	14	18	83

$x^2 = 7.1270$ df = 15 Probability = .95

Table J.30. Chi-square analysis on AIM 8 (rates of learning) by Tatrain

	1	2	3	4	5	6
0.00	13	9	11	3	3	14
1.00	40	15	19	3	7	31
2.00	18	10	15	3	7	22
3.00	94	32	45	14	18	83

Table J.31. Chi-square analysis on AIM 9 (evaluation materials) by Tatrain

	1	2	3	4	5	6
0.00	14	8	9	0	3	3
1.00	30	15	18	7	8	37
2.00	27	11	19	2	6	27
3.00	94	32	44	14	18	83
x^2 = 17.0170			df = 15		Probability = .32	

Table J.32. Chi-square analysis on ATP 1 (use of AV) by Tatrain

	1	2	3	4	5	6
0.00	62	27	6	9	9	47
1.00	55	19	36	10	16	59
2.00	48	22	27	7	10	44
x^2 = 6.3228			df = 10		Probability = .79	

Table J.33. Chi-square analysis on ATP 2 (practice on objectives) by Tatrain

	1	2	3	4	5	6
0.00	75	35	42	11	18	71
1.00	74	29	37	11	17	62
2.00	16	2	11	1	0	17
x^2 = 9.8041			df = 10		Probability = .46	

Table J.34. Chi-square analysis on ATP 3 (logical sequence) by Tatrain

	1	2	3	4	5	6
0.00	97	32	47	17	20	73
1.50	55	30	34	4	13	57
2.50	13	4	7	2	2	19
	x^2 = 11.6480		df = 10		Probability = .31	

Table J.35. Chi-square analysis on ATP 4 (variety of method) by Tatrain

	1	2	3	4	5	6
0.00	51	17	26	6	8	35
1.00	66	21	29	10	11	60
2.00	31	19	24	3	12	31
3.00	17	9	11	4	4	24
	x^2 = 12.9244		df = 15		Probability = .61	

Table J.36. Chi-square analysis on ATP 5 (appropriate methods) by Tatrain

	1	2	3	4	5	6
0.00	39	15	17	7	7	35
1.00	60	18	36	8	12	45
2.00	48	22	20	5	11	43
3.00	18	11	17	3	5	27
	x^2 = 10.2252		df = 15		Probability = .81	

Table J.37. Chi-square analysis on ATP 6 (group sizes) by Tatrain

	1	2	3	4	5	6
0.00	82	33	47	12	21	65
1.50	66	29	29	7	9	61
2.50	17	4	14	4	5	23

x^2 = 10.6055 df = 10 Probability = .39

Table J.38. Chi-square analysis on ATP 7 (involves learners) by Tatrain

	1	2	3	4	5	6
0.00	45	26	35	9	8	37
1.00	84	31	36	12	14	75
2.00	36	9	19	2	13	38

x^2 = 17.8312 df = 10 Probability = .06

Table J.39. Chi-square analysis on ATP 8 (learner involvement) by Tatrain

	1	2	3	4	5	6
0.00	94	29	39	10	15	69
1.50	61	34	39	13	19	59
2.50	10	3	12	0	1	22

x^2 = 22.4815 df = 10 Probability = .01

Table J.40. Chi-square analysis on ATP 9 (knowledge of subject) by Tatrain

	1	2	3	4	5	6
0.00	111	44	59	20	21	96
1.50	43	19	28	3	10	41
2.50	11	3	3	0	4	12

$x^2 = 9.4490$ df = 10 Probability = .49

Table J.41. Chi-square analysis on ATP 10 (accurate information) by Tatrain

	1	2	3	4	5	6
0.00	140	64	80	23	32	134
1.50	24	2	10	0	3	15

$x^2 = 9.9808$ df = 5 Probability = .08

Table J.42. Chi-square analysis on ATP 11 (purpose of topic) by Tatrain

	1	2	3	4	5	6
0.00	70	26	41	7	13	66
1.00	81	36	37	13	20	69
2.00	14	4	12	3	2	15

$x^2 = 7.2922$ df = 10 Probability = .70

Table J.43. Chi-square analysis on AIM 1 (appropriate objectives) by region

	1	2	3	4	5	6	7	8
0.00	7	3	17	22	9	4	5	8
1.00	10	7	17	16	12	7	4	10
2.00	44	34	67	64	41	17	22	50
3.00	3	5	10	5	4	0	1	4

9 out of 32 (28.1%) expected cell frequencies are less than 5.

Table J.44. Chi-square analysis on AIM 2 (procedures for objectives) by region

	1	2	3	4	5	6	7	8
0.00	11	3	21	13	4	7	3	6
1.00	14	9	27	28	16	4	8	14
2.00	5	9	10	11	11	1	5	7
3.00	34	28	53	55	35	16	16	45

$x^2 = 24.9268$ df = 21 Probability = .25

Table J.45. Chi-square analysis on AIM 3 (specific content) by region

	1	2	3	4	5	6	7	8
0.00	13	11	16	17	11	4	6	7
1.00	15	5	31	26	17	7	7	14
2.00	2	5	11	9	3	1	3	5
3.00	34	28	53	55	35	16	16	46

$x^2 = 16.7538$ df = 21 Probability = .73

Table J.46. Chi-square analysis on AIM 4 (appropriate for mastery) by region

	1	2	3	4	5	6	7	8
0.00	9	9	15	8	6	4	1	4
1.00	9	6	28	31	16	6	8	13
2.00	12	6	15	12	9	2	7	9
3.00	34	28	53	56	35	16	16	46

$x^2 = 23.6227$ df = 21 Probability = .31

Table J.47. Chi-square analysis on AIM 5 (cognitive levels) by region

	1	2	3	4	5	6	7	8
0.00	7	5	16	20	7	1	4	11
1.00	13	5	22	16	13	5	7	5
2.00	8	7	11	11	8	5	4	4
3.00	2	4	8	4	2	1	1	6
4.00	34	28	54	5	36	16	16	46

11 out of 40 (27.5%) expected cell frequencies are less than 5.

Table J.48. Chi-square analysis on AIM 6 (learners capabilities) by region

	1	2	3	4	5	6	7	8
0.00	6	7	17	17	11	3	3	7
1.00	13	7	21	19	12	2	8	10
2.00	11	4	12	12	4	5	3	4
3.00	0	3	7	4	3	1	2	5
4.00	34	28	54	55	36	17	16	46

11 out of 40 (27.5%) of expected cell frequencies are less than 5.

Table J.49. Chi-square analysis on AIM 7 (learning styles) by region

	1	2	3	4	5	6	7	8
0.00	6	5	14	13	4	6	4	9
1.00	13	6	25	26	12	3	9	16
2.00	11	10	17	13	14	3	3	8
3.00	34	28	55	55	36	16	16	45
x^2 = 18.2393			df = 21			Probability = .63		

Table J.50. Chi-square analysis on AIM 8 (rates of learning) by region

	1	2	3	4	5	6	7	8
0.00	7	3	13	11	5	3	4	7
1.00	11	11	24	27	15	5	10	12
2.00	12	7	19	14	10	3	2	8
3.00	34	28	55	55	36	17	16	45
x^2 = 11.1817			df = 21			Probability = .96		

Table J.51. Chi-square analysis on AIM 9 (evaluation material) by region

	1	2	3	4	5	6	7	8
0.00	5	1	8	4	7	0	7	5
1.00	16	9	24	27	11	8	7	13
2.00	9	11	24	21	12	4	2	9
3.00	34	28	55	55	36	16	16	45
x^2 = 27.5193			df = 21			Probability = .15		

Table J.52. Chi-square analysis on ATP 1 (use of AV) by region

	1	2	3	4	5	6	7	8
0.00	23	9	25	35	27	11	14	32
1.00	35	26	27	29	27	13	12	26
2.00	6	14	59	43	12	4	6	14
x^2 = 72.2271			df = 14			Probability = .00		

Table J.53. Chi-square analysis on ATP 2 (practice on objectives) by region

	1	2	3	4	5	6	7	8
0.00	33	17	48	48	39	13	15	39
1.00	27	24	51	48	24	12	15	29
2.00	4	8	12	11	3	3	2	4
x^2 = 13.5260			df = 14			Probability = .49		

Table J.54. Chi-square analysis on ATP 3 (logical sequence) by region

	1	2	3	4	5	6	7	8
0.00	28	23	57	68	34	15	15	46
1.50	28	19	47	31	23	12	17	16
2.50	8	5	6	8	9	1	0	10
x^2 = 25.8777			df = 14			Probability = .03		

Table J.55. Chi-square analysis on ATP 4 (variety of methods) by region

	1	2	3	4	5	6	7	8
0.00	12	16	18	37	16	8	11	25
1.00	19	13	40	45	21	12	14	33
2.00	17	11	39	16	18	6	6	7
3.00	16	9	14	9	11	2	1	7

$x^2 = 48.7349$ df = 21 Probability = .00

Table J.56. Chi-square analysis on ATP 5 (appropriate methods) by region

	1	2	3	4	5	6	7	8
0.00	10	9	17	32	16	7	8	21
1.00	20	13	30	40	21	13	11	31
2.00	22	16	39	26	18	6	7	15
3.00	12	11	25	9	11	2	6	5

$x^2 = 34.5551$ df = 21 Probability = .03

Table J.57. Chi-square analysis on ATP 6 (ability to work with individuals and groups) by region

	1	2	3	4	5	6	7	8
0.00	26	25	44	62	31	17	15	40
1.50	29	14	52	36	21	9	15	25
2.50	9	10	14	9	14	2	2	7

$x^2 = 22.7684$ df = 14 Probability = .06

Table J.58. Chi-square analysis on ATP 7 (involve learners) by region

	1	2	3	4	5	6	7	8
0.00	22	10	43	27	11	12	10	25
1.00	27	28	52	56	35	9	15	30
2.00	15	11	16	24	20	7	7	17

x^2 = 20.8576 df = 14 Probability = .11

Table J.59. Chi-square analysis on ATP 8 (maintain learner involvement) by region

	1	2	3	4	5	6	7	8
0.00	25	23	44	61	29	17	16	41
1.50	34	18	55	36	33	10	12	27
2.50	5	8	12	10	4	1	4	4

x^2 = 20.2854 df = 14 Probability = .12

Table J.60. Chi-square analysis on ATP 9 (knowledge of subject) by region

	1	2	3	4	5	6	7	8
0.00	38	21	77	85	38	18	22	52
1.50	21	22	28	19	21	7	8	18
2.50	5	5	6	3	7	3	2	2

6 out of 24 (25%) of the valid cells have expected cell frequency less than 5.

Table J.61. Chi-square analysis on ATP 10 (presents accurate information) by region

	1	2	3	4	5	6	7	8
0.00	63	45	104	89	57	27	29	59
1.50	1	4	7	17	8	1	3	13

$x^2 = 17.6759$ df = 7 Probability = .01

Table J.62. Chi-square analysis on ATP 11 (purpose of topic) by region

	1	2	3	4	5	6	7	8
0.00	30	16	55	39	23	17	14	29
1.00	30	26	50	54	41	9	14	32
2.00	4	7	6	14	2	2	4	11

$x^2 = 23.1926$ df = 14 Probability = .06

Table J.63. Chi-square analysis on AIM 1 (appropriate objectives) by job

	1	2	3	4
0.00	3	35	3	33
1.00	2	41	3	35
2.00	8	174	42	114
3.00	4	11	3	14

4 out of 16 (25%) of the expected cell frequencies are less than 5.

Table J.64. Chi-square analysis on AIM 2 (procedures for objectives) by job

	1	2	3	4
0.00	2	35	6	24
1.00	6	56	6	51
2.00	3	23	6	26
3.00	6	147	33	95
$x^2 = 11.5025$		df = 9	Probability = .24	

Table J.65. Chi-square analysis on AIM 3 (specific content) by job

	1	2	3	4
0.00	2	42	7	33
1.00	6	56	10	48
2.00	2	16	1	20
3.00	7	147	33	95
4 out of 16 (25%) expected cell frequencies are less than 5.				

Table J.66. Chi-square analysis on AIM 4 (appropriate for mastery) by job

	1	2	3	4
0.00	1	28	6	20
1.00	3	54	10	48
2.00	6	31	2	33
3.00	7	148	33	95
$x^2 = 15.9332$		df = 9	Probability = .07	

Table J.67. Chi-square analysis on AIM 5 (cognitive levels) by job

	1	2	3	5
0.00	2	42	7	33
1.00	6	56	10	48
2.00	2	16	1	20
3.00	7	147	33	95

4 out of 16 (25%) expected cell frequencies are less than 5.

Table J.68. Chi-square analysis on AIM 6 (learner capabilities) by job

	1	2	3	4
0.00	5	36	2	27
1.00	2	46	9	28
2.00	0	24	5	28
3.00	3	7	1	17
4.00	7	148	34	96

5 out of 20 (20%) expected cell frequencies are less than 5.

Table J.69. Chi-square analysis on AIM 7 (learning styles) by job

	1	2	3	4
0.00	2	31	7	21
1.00	3	52	9	38
2.00	5	31	1	41
3.00	7	147	34	96

$x^2 = 18.2335$ df = 9 Probability = 0.03*

*$p < .05$.

Table J.70. Chi-square analysis on AIM 8 (rates of learning) by job

	1	2	3	4
0.00	2	27	5	19
1.00	2	55	10	46
2.00	6	31	2	35
3.00	7	148	34	96
$x^2 = 16.6271$		df = 9	Probability = .05	

Table J.71. Chi-square analysis on AIM 9 (evaluation material) by job

	1	2	3	4
0.00	1	20	2	13
1.00	3	48	11	51
2.00	6	46	4	36
3.00	7	147	34	96
4 out of 16 (25%) expected cell frequencies are less than 5.				

Table J.72. Chi-square analysis on ATP 1 (use of AV) by job

	1	2	3	4
0.00	7	87	23	59
1.00	6	98	23	66
2.00	4	76	5	71
$x^2 = 14.3983$		df = 6	Probability = .03*	

*$p < .05$.

Table J.73. Chi-square analysis on ATP 2 (practice on objectives) by job

	1	2	3	4
0.00	11	125	29	85
1.00	5	116	17	91
2.00	1	20	5	20
$x^2 = 6.1738$		df = 6	Probability = .40	

Table J.74. Chi-square analysis on ATP 3 (logical sequence) by job

	1	2	3	4
0.00	8	147	29	99
1.50	5	93	19	75
2.50	3	20	3	21
$x^2 = 4.5741$		df = 6	Probability .60	

Table J.75. Chi-square analysis on ATP 4 (variety of methods) by job

	1	2	3	4
0.00	2	73	17	50
1.00	6	96	18	76
2.00	6	60	11	42
3.00	3	32	5	28
$x^2 = 5.0850$		df = 9	Probability = .83	

Table J.76. Chi-square analysis on ATP 5 (appropriate methods) by job

	1	2	3	4
0.00	1	68	12	37
1.00	4	84	20	70
2.00	9	65	11	64
3.00	3	44	8	25
$x^2 = 14.2377$		df = 9	Probability = .11	

Table J.77. Chi-square analysis on ATP 6 (group sizes) by job

	1	2	3	4
0.00	7	131	30	90
1.50	6	102	14	78
2.50	4	27	7	28
$x^2 = 6.5875$		df = 6	Probability = .36	

Table J.78. Chi-square analysis on ATP 7 (involves learners) by job

	1	2	3	4
0.00	8	80	18	53
1.00	5	121	25	99
2.00	4	60	8	44
$x^2 = 5.3992$		df = 6	Probability = .49	

Table J.79. Chi-square analysis on ATP 8 (learners involvement) by job

	1	2	3	4
0.00	7	132	26	89
1.50	6	107	18	92
2.50	4	22	7	15
$x^2 = 8.3916$		df = 6	Probability = .21	

Table J.80. Chi-square analysis on ATP 9 (knowledge of subject) by job

	1	2	3	4
0.00	13	179	30	128
1.50	3	72	15	52
2.50	1	10	5	16
3 out of 12 (25%) expected cell frequencies are less than 5.				

Table J.81. Chi-square analysis on ATP 10 (accurate information) by job

	1	2	3	4
0.00	17	234	45	174
1.50	0	25	6	22
$x^2 = 2.3878$		df = 3	Probability = .50	

Table J.82. Chi-square analysis on ATP 11 (purpose of topic) by job

	1	2	3	4
0.00	8	110	20	84
1.00	9	119	27	100
2.00	0	32	4	12
$x^2 = 7.5863$		df = 6	Probability = .27	

APPENDIX K: USE OF HUMAN SUBJECTS APPROVAL FOR RESEARCH

The Iowa State University Committee on the Use of Human Subjects in Research reviewed this project and concluded that the rights and welfare of the human subjects were adequately protected, that risks were outweighed by the potential benefits and expected value of the knowledge sought, that confidentiality of data was assured and that informed consent was obtained by appropriate procedures.