Order Number 8907844

A study of the criterion-related validity of North Carolina's Teacher Performance Appraisal Instrument

Riner, Phillip Scott, Ed.D.

The University of North Carolina at Greensboro, 1988

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A STUDY OF THE CRITERION-RELATED VALIDITY OF NORTH CAROLINA'S TEACHER PERFORMANCE APPRAISAL INSTRUMENT

by

Phillip Scott Riner

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

Greensboro
1988

Approved by

John VanHoose
Dissertation Adviser

APPROVAL PAGE

This dissertation has been approved by the following committee of the Graduate School at The University of North Carolina at Greensboro.

Dissertation Adviser John VanHoose

Committee Members William K. Purkey

Shirley L. Hayworth

David B. Strahan

June 21, 1988
Date of Acceptance by Committee

June 21, 1988
Date of Final Oral Examination

RINER, PHILLIP SCOTT, Ed.D. A Study of the Criterion-Related Validity of North Carolina's Teacher Performance Appraisal Instrument. (1988)
Directed by Dr. John Van Hoose. 166 pp.

The purpose of the research was to test the criterion validity of a high inference rating scale the North Carolina Teacher Performance Appraisal Instrument (TPAI). In 1987 North Carolina State Department of Public Instruction mandated annual evaluation of teachers in each school district utilizing the TPAI. The TPAI is composed of eight functions, five based on effective teaching research and three derived by professional consensus found in the literature. The TPAI is administered by principals and other personnel who have received specific training in the use of the instrument. Teachers are ranked on each of the eight functions using a six point scale ranging from Unsatisfactory to Superior.

A sample of 40 teachers and 400 students were used to calculate partial correlation coefficients between each TPAI function rating and student achievement as measured by the California Achievement Test (CAT). Within-class regression was employed to estimate average student gains for each teacher. Stepwise multiple regression was used to select the student variables to be held constant. Student variables used to statistically equate classrooms were grade, IQ, number of absences during the first six months of school and student sex.

The study found only one TPAI function (Non-Instructional Duties) to be significantly related ($p < .05$) to total achievement as measured by the CAT. The data revealed a significant positive relationship between each TPAI function rating and estimated student achievement in math with coefficients ranging from .36 to .48 ($p < .05$). There was no significant relationship between TPAI rating and estimated student achievement in reading.

Five of the 64 correlations between student variables and TPAI function rating scores were significant ($p < .05$). It was concluded that there was a relationship between TPAI rating and class composition although the implication of the relationship is unclear. The data revealed no significant relationship between teacher variables and TPAI rating.

ACKNOWLEDGMENTS

The writer wishes to express appreciation to John Van Hoose, chairman of the dissertation, for guidance, encouragement and insight throughout the writing of this research project. A special thanks to William Purkey for his encouragement and gentle reassurance. The writer would also like to thank David Strahan for his pensive questions and help in obtaining scarce information. To Shirley Haworth a devoted thanks is sent for her guidance and encouragement over the past five years.

A note of appreciation is sent to the teachers and administrators of the participating school district. Without their assistance and willingness to share their evaluation process this dissertation would not be possible.

To my wife Patsy I would like to send a message of deep appreciation for listening to the manuscript and supplying ideas for improvement. Her devoted support and encouragement was constant. To Ezra and Miriam, whose bedtime hugs served as an inspiration in the late nights of of calculating and rewriting, the author would like to send his appreciation for their restless patience. It is for the good of all children that educators struggle with unyielding problems and pursue elusive knowledge.

Phillip S. Riner

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

During the course of the 1980's, North Carolina's Department of Public Instruction (NCSDPI) responded to a series of legislative mandates concerning the improvement of instruction in North Carolina's public schools. The process of teacher education and teacher evaluation was subject to particular scrutiny. The state legislature, through a series of actions, mandated a major revision of teacher evaluation practices. An increased attention to initial licensure and tenure policies accompanied this reexamination of current practices in teacher evaluation. Also, the possibility of establishing a promotion and salary system based on teacher evaluation was placed under study.

The vehicle selected to satisfy this mandate was an evaluative rating scale developed by NCSDPI specifically for use in North Carolina schools. It was to be administered by school principals and other trained personnel. The state legislature charged NCSDPI to develop this instrument based on an extensive examination of the empirical research on teacher effectiveness which had accumulated since 1960. It was hoped that by employing

empirical research, rather than professional consensus, an instrument would be developed that covered the essential aspects of teaching in a generic omnibus format that could be objectively administered and defended. The resulting instrument based on this literature was a high inference rating scale called the Teacher Performance Appraisal Instrument (TPAI).

The TPAI, however, has not been empirically defended and there is great doubt among educational professionals in North Carolina as to whether the legislative mandate has been successfully completed (Williams, et al., 1987). While the instrument provided copious references to empirical research for each item in the instrument, formal criterion validity studies and a technical manual have yet to be provided by the TPAI developers. This study addressed the question "Does the TPAI rating scale indicate which teachers are most successful in bringing about basic skill gains in students?"

An Historical Perspective

Although teacher rating has been a formal method of teacher evaluation since the turn of the century, these ratings have been based primarily on criteria established through professional or administrative consensus. Criterion validity and reliability for most of these systems have not been established. A bureaucratic

evaluation process for teacher evaluation has emerged in American public schools (Darling-Hammond, 1986). In this bureaucratic model administrators conduct a few classroom observations and report their findings in the form of a general rating. The evaluation instrument employed is usually composed of various standardized criteria using ratings on a three- or five-point scale (Darling-Hammond, 1986). These rating scales are often developed from consensus measures derived from surveys designed to ascertain what those in the profession consider good teaching practice. This has been a popular source of evaluative criteria and is almost assured to have face validity. These professional opinion polls, however, typically reflect what is currently fashionable in teaching rather than what is defensible in practice. The resulting rating criteria are generally thought to reflect the kind of learning environment a teacher creates in the classroom rather than teaching effectiveness in meeting specified goals (Darling-Hammond, 1986).

The subjective judgmental practices of the past have persisted, at least in part, because of a lack of a substantial body of empirical knowledge about teaching. Practices such as being judged by appearance and dress were tolerated by teachers even in the late 1970's (Kowalski, 1978) primarily because no one knew of a better system.

Teachers also had little influence over the direction of evaluation practices.

During the 1970's and 1980's, what was known about teaching increased dramatically as a result of a body of studies known as effective teaching research. Likewise, the demands for efficient and effective education for every child steadily increased. Renewed pressure on educational institutions to provide evidence of their effectiveness required an objective and rational approach to teacher evaluation. North Carolina's Teacher Performance Appraisal Instrument (TPAI) was an attempt to provide a rational, defensible, and fair method of evaluating the quality of instruction provided by North Carolina's teachers.

TPAI Development and Use

Developers of the TPAI were given three guidelines to be followed. First, any practice expected of teachers should be equally applicable to effective teaching regardless of the teacher's grade level or subject area assignment. Second, the practice must be identified as desirable in more than one effective teacher study. Third, the teacher could alter the behavior identified as effective; that is, the teacher could learn to exhibit the practice (NCSDPI, 1986b).

Based on an extensive review of teaching research literature sponsored by the NCSDPI, twenty-eight practices

that appeared to relate directly to classroom teaching were identified. These in turn were clustered under five major functions:

1. Management of Instructional Time (four practices)
2. Management of Student Behavior (five practices)
3. Instructional Presentation (eleven practices)
4. Instructional Monitoring of Student Performance (four practices)
5. Instructional feedback (four practices)

Three additional functions were later added to the TPAI for use with tenured teachers consisting of ten practices that were considered job related but not necessarily a part of daily practice. They were:

6. Facilitating Instruction (five practices)
7. Communication Within the Education Environment (two practices)
8. Performing Non-Instructional Duties (three practices)

From this total of thirty-eight practices, the Teacher Performance Appraisal Instrument (TPAI) was established and piloted (NCSDPI, 1986). A copy of the instrument can be found in Appendix A.

Each of the eight functions were to be scored on a one-to-six scale for which a rating of "one" indicated an unsatisfactory demonstration of that cluster of skills and a "six" indicated a superior demonstration. While these

practices were established as independent criteria, a normative meaning was appended to the function definitions thus making it unclear whether teachers were to be rated by an external explicitly stated criteria or whether teachers were compared to one another in a normative fashion. The following explanation of an unsatisfactory performance is an example of the dual standards presented to evaluators:

> Performance within this function area is CONSISTENTLY INADEQUATE/UNACCEPTABLE and MOST practices require CONSIDERABLE IMPROVEMENT to fully MEET MINIMUM PERFORMANCE EXPECTATIONS. Teacher requires CLOSE AND FREQUENT SUPERVISION in the performance of ALL RESPONSIBILITIES.
>
> Teacher's performance on this major function area could be characterized as being in the BOTTOM 5% of ALL THE TEACHERS IN NORTH CAROLINA. (NCSDPI, 1985a, Transparency 7.8)

The implication was that five percent of North Carolina teachers were to be found unsatisfactory in each function regardless of the overall quality of teaching found statewide.

Implementing the TPAI Process

To disseminate the new scale and to educate school personnel concerning the TPAI process, an extensive statewide series of interrelated workshops was developed around the instrument and made available to local education agencies. These included:

- Effective Teacher Training (ETT) for teachers, evaluators and other staff (30 hours).

-Teacher Performance Appraisal Training (TPAI) for evaluators and interested other staff (24 hours).

-Professional Development Plans (PDP) for those who assist staff with professional growth needs (6 hours).

-Mentor/Support Team Training (M/STT) for those who assist beginning or initially certified staff (30 hours). (NCSDPI, 1986b, p. 6)

This extensive network of training programs cleared the way for the North Carolina State Board of Education (NCSBE) to mandate annual evaluation of certified teaching personnel (NCSDPI, 1986b). Through this sequence of events, the TPAI became the official method for the evaluation of all teachers in North Carolina's public schools on July 1, 1987.

Purpose of Study

The purpose of this study was to establish and examine criterion-related validity evidence for North Carolina's TPAI against the criterion of effective teaching. In evaluating tests, the most important consideration is validity. Validity refers to "the appropriateness, meaningfulness, and usefulness of the specific inferences made from the test score" (AERA, APA and NCME, 1985, p. 9).

Test validity cannot be established unilaterally. Evidence must be collected which supports the specific inferences to be made of test results. This evidence can be accumulated by a variety of means, including evidence

gathered on similar instruments used in similar situations (AERA et al., 1985). The process of establishing validity is a process of gathering a preponderance of evidence to support a generalization about the appropriateness of a particular pattern of inferences in a given set of circumstances.

Although several criteria could be proposed for evaluating the criterion validity of the TPAI, the most logical choice would be measures of student achievement. The content validity evidence presented by NCSDPI for five of the eight TPAI functions refers to instructional practices. The basis for the inclusion of specific items in these five functions consisted mainly of correlational studies between teacher behavior and various measures of student achievement. Student achievement was the dominant criterion cited for the construct "effective practice" being the criteria 109 times from a total of 190 citations. It was primarily to this evidence that this criterion validity study was directed.

Three of the eight functions concern the maintenance of a professional posture (e.g. maintaining records) and are defended by reasoned argument found in the literature and from commonly expected functions of teachers (e.g. treating all students fairly) (NCSDPI, 1985b). In effect, these last three functions are consensus based. They are used only for tenured personnel and are not used in the

evaluation of initially certified personnel. These three functions were included in the study and compared to achievement criteria. If the validity coefficients for these functions are as strong as the research based functions, then there is evidence that consensus based measures have equal utility in teacher evaluation.

To establish the criterion-related evidence necessary to evaluate the proposed inferences from the TPAI evaluation results, three avenues of investigation were pursued: (1) the relation of TPAI to student achievement, (2) the relation of TPAI to student variables other than achievement and (3) the relation of TPAI to teacher variables.

TPAI and Student Achievement

It is claimed by test developers that the TPAI can function as a valid measure of the teaching skills necessary to bring about changes in pupils' abilities to perform basic academic skills. If this is correct, there would be a statistically significant positive correlation between the teacher's TPAI rating and his or her students' achievement. While a direct measure of pretest-posttest gain is desirable, these gain scores are subject to several methodological problems. For example, it is known that many variables effect student achievement and may account for as much as 80% of the non-instructional variance in

achievement among students. Since students are not randomly assigned, these variables can favor one teacher and disadvantage another when measuring achievement test gains. To correct this problem, the major moderating variables that effect student achievement were controlled. These non-instructional student variables were those which the teacher could not reasonably be expected to affect. Included were family status, IQ, days in attendance, number of parents in the home and the family's economic situation as indicated by participation in the federal lunch assistance program. All gain scores are subject to distortion due to regression (Glass and Hopkins, 1984). A method of estimating student gains utilizing a within-class regression technique developed by Medley, Coker, and Soar (1984) was employed to control for this distortion.

The content validity evidence provided for the TPAI indicated that some teaching functions are most successful in teaching material that is highly structured. Therefore, certain TPAI items may be better predictors of student achievement in some subjects than others. To make the research design sensitive to these possibilities, a concurrent validity coefficient for each function on the TPAI was constructed using an achievement measure for math, reading, and a total composite of basic skills.

TPAI and Student Variables

It is claimed by the TPAI test developers that the TPAI is an omnibus measure of teacher effectiveness which is not affected by class composition or teaching assignment (NCSDPI, 1986b). If the test is applicable to all teaching situations, then all teachers would have had an equal chance for a favorable TPAI rating. To test this assumption simple zero-order correlations between teacher TPAI scores and student variables were calculated. It was hypothesized that these relationships would be equal to zero. Bias would be indicated if a statistically significant relationship between a student variable and the teacher's TPAI rating were found.

TPAI and Teacher Variables

Evaluations should address the teacher's skills in meeting the stated criteria. Evaluations should not discriminate among teachers by race, sex, age, or seniority. It was important to test the TPAI for bias in these areas while establishing validity evidence. If the TPAI is bias-free in regard to these teacher variables, then one would expect simple zero-order correlations between teacher TPAI ratings and teacher variables to be equal to zero. However, it is possible that some teacher variables may have a correlation with TPAI results. In these cases the bias may be justified only if the teacher

variable has a similar significant relationship to student achievement.

Major Hypotheses

The method of study was a statistical analysis of the TPAI function and composite scores of elementary teachers in grades two through six and their relationship to student achievement and student variables. Additionally, a statistical analysis of the TPAI function and composite scores and their relationship to teacher variables was conducted using elementary teachers with assignments in grades one through six and exceptional children's programs. Secondary teachers were omitted from the study due to a lack of criterion measures suited to comparisons across teachers.

Analysis of these data provided evidence to test the following major hypotheses:

Hypotheses Concerning Criterion Validity Coefficients

H_1: There is a significant positive relationship between TPAI total score and estimated student gains of the CAT total score using within-class regression.

H_2: There is a significant positive relationship between each TPAI function score and estimated student gains of the CAT total score using within-class regression.

H_3: There is a significant positive relationship between TPAI total score and estimated student gains on the Math subtest of the CAT using within-class regression.

H_4: There is a significant positive relationship between each TPAI function score and estimated student gains of the Math subtest of the CAT using within-class regression.

H_5: There is a significant positive relationship between TPAI total score and estimated student gains on the Reading subtest of the CAT using within-class regression.

H_6: There is a significant positive relationship between each TPAI function score and estimated student gains of the Reading subtest of the CAT using within-class regression.

Hypotheses Concerning Systematic Bias

H_7: The zero order correlation coefficients between TPAI function scores and the class mean (or ratio) of the student variables of race, sex, IQ, past achievement, age, grade, economic status, attendance, or family structure are equal to zero.

H_8: The zero order correlation coefficients between TPAI function scores and the teacher variables of age, sex, race, highest earned degree, grade taught, years teaching in school, years teaching in system, or total years of teaching experience are equal to zero.

Significance of the Study

Validity studies of evaluation measures are always poignant to test users. A wide range of validity evidence should be presented by test developers. This researcher could not locate any study concerning the criterion validity of the TPAI in the datafiles of the Education Resources Information Center (ERIC). Only two documents were found that mention North Carolina's TPAI; both were documents dealing with the training of evaluators in the use of that instrument and contained limited information on its reliability (NCSDPI, 1986a; NCSDPI, 1986b).

NCSDPI Personnel Relations Division, which supervised the implementation of the TPAI program, was unable to supply a technical manual for the TPAI containing the typical statistics on reliability and validity. The most recent study produced by the Division of Personnel Relations was typical of the limited research concerning the TPAI. This study was a large scale survey of teachers and evaluators ascertaining the attitudes of personnel involved in the evaluation process. The study did not reflect criterion validity issues (Stacey, 1988).

Validity studies on teacher rating scales are uncommon. Lancelot et al. (1935), Reavis and Cooper (1945), Capie (1980b) and Medley and Coker (1987) consistently found low correlations with student achievement and various teacher evaluation rating scales. Medley and Coker (1987) noted the scarcity of criterion validity studies on rating scales which have been used to judge teacher effectiveness:

> Although the question of whether or not these judgments are valid is a natural and important one, it is rarely asked. The validity of principals' judgments of the effectiveness of the teachers they supervise is generally taken for granted (p. 138).

Medley and Coker (1987) were able to cite eight studies examining the validity of principals' judgments since 1935. Each study concluded that there was no appreciable agreement between principals' judgments of teacher's effectiveness and the amount students learn. Seven of the

studies were conducted prior to 1954. The present study explored a major gap in the teacher evaluation literature.

There is an increasing danger for the misuse of teacher rating scales as the use of tests and evaluation measures proliferate. This is particularly true in "high stakes" testing where substantive decisions are made utilizing the test score as a decision-making criteria such as the use of the TPAI as a certification instrument. The experimental use of the TPAI as a tool for defining and granting career ladder promotions further emphasizes the importance of a criterion validity study.

Williams et al. (1987) philosophically note in examining the effect of TPAI use on teaching style, "There's many a slip between the cup and the lip" (p. 26) claiming that current practice falls short of stated goals. In combating the potential for misuse of tests and evaluation measures, as well as governing their construction and guiding their use, the American Education Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) established the Standards for Educational and Psychological Testing (AERA et al., 1985). The Standards provides guidelines for the development and use of tests and measures. Tests for which these standards are designed to apply are broadly defined and "include standardized ability...instruments, diagnostic and

evaluative devices, interest inventories, personality inventories, and projective instruments" (AERA et al., 1985, p. 3).

The Standards consistently reiterate the responsibilities of test developers and test users. Test users are urged to "have a sound technical and professional basis for their actions, much of which can be derived from research done by test developers and publishers" (AERA, et al., 1985, p.3). The test user is encouraged to rely "heavily upon the developer's research documentation that is clearly related to the intended application" (AERA et al., 1985, p. 3). The absence of this supporting document reiterates the need for studies exploring various aspects of the TPAI.

The Standards outlines primary requirements that should be met by all tests before their operational use unless a "sound professional reason is available to show why it is not necessary, or technically feasible, to do so in a particular case" (AERA, et al., p. 2). The following standards are especially applicable to the TPAI:

> Standard 1.1 Evidence of validity should be presented for the major types of inferences for which the use of a test is recommended. A rationale should be provided to support the particular mix of evidence presented for the intended uses. (Primary)
>
> Standard 1.2 If validity for some common interpretation has not been investigated, that fact should be made clear, and potential users should be cautioned about making such interpretations. Statements about validity

should refer to the validity of particular interpretations or of particular types of decisions. (Primary)

Standard 1.11 A report of a criterion-related validation study should provide a description of the sample and the statistical analysis used to determine the degree of predictive accuracy. Basic statistics should include numbers of cases (and the reasons for eliminating any cases), measures of central tendency and variability, relationships, and a description of any marked tendency toward nonnormality of distribution. (Primary)

Standard 1.12 All criterion measures should be described accurately, and the rationale for choosing them as relevant criteria should be made explicit. (Primary).

Standard 1.13 The technical quality of all criteria should be considered carefully. Criteria should be determined independently of predictor test scores. If evidence indicated that a criterion measure is affected to a substantial degree by irrelevant factors, this evidence should be reported. If special steps are taken to reduce the effects of irrelevant factors, these steps should be described in detail. (Primary)

Standard 1.14 When criteria are composed of rater judgments, the degree of knowledge that raters have concerning ratee performance should be reported. If possible, the training and experience of the raters should be described. (Primary) (AERA et al., 1985, pp. 13-16).

These six primary standards concerning validity issues relate directly to the development and use of the TPAI instrument. As of this writing, no published source of this validity information listed as a primary requirement by the Standards could be found by the author through exhaustive data searches and inquiry. The Standards

require that such information be readily available to users of the test.

Considerable evidence was presented by NCSDPI in the North Carolina Performance Appraisal Training Program (1985b) and the North Carolina Effective Teaching Training Program (1985a) to establish the content validity of each TPAI function. Neither document addressed the validity of the actual certification and promotion inferences to be made by users of the TPAI. Content validity evidence is considered by the Standards as inadequate unless the connection between job and test is close and direct. Standard 10.5 addresses the issue of offering content-related validity as evidence supporting test use in employment related decisions:

> When the content-related validation evidence is to stand as support for the use of a test in selection or promotion, a close link between test content and job content should be demonstrated. (Primary) (AERA et al., 1985, p. 61)

The Standards expands the meaning of this standard by providing an explanatory comment:

> For example, if the test content samples job tasks with considerable fidelity (e.g., actual job samples such as machine operator) or, in the judgment of experts, correctly simulates job task content (e.g., certain assessment center exercises), or samples specific knowledge required for successful job performance (e.g., information necessary to exhibit certain skills), then content-related validity can be offered as the principal form of evidence of validity. If the link between the test content and the job content is not singular and direct, additional evidence is required. (AERA et al., 1985, p. 61)

Teaching is an exceedingly complex task and no singular or combined line of direct evidence defining effective teaching has been forthcoming (Berliner, 1984). Clearly, additional evidence of validity is needed for the TPAI. This criterion-related validity study provided basic knowledge to help users appropriately employ the TPAI instrument and address the most critical issues presented by the Standards.

Standardized Achievement as an External Criterion

Standardized achievement tests such as the California Achievement Test (CAT) have been criticized as being invalid criterion as a measure of effective teaching because the test content may not measure what is taught in the classroom (Glass, 1977; Medley, Coker and Soar, 1984). The objection to the evaluation of teachers by these scores has focused on a potential lack of curricular relevance at the classroom level (Medley, Coker and Soar, 1984). The set of circumstances in this study suggested that empirical evidence indicated that TPAI items are related to student achievement as measured by standardized achievement tests.

The researcher's use of the CAT as a criterion in this study was supported by the CAT's use by the NCSDPI as a fundamental tool for student and school assessment. For example, the CAT has been employed as major criteria for

selection of children into programs for the academically gifted. The CAT has been employed as a primary criteria for mandatory summer school for third, sixth, and eighth graders. School and system-wide CAT results have been distributed to newspapers and televisions stations by local educational agencies as part of efforts toward public accountability. It has also been a common practice for schools and school systems to report their test results in public relation brochures. The CAT has been a common criterion by which citizens have judged their schools. If the TPAI were to be found incapable of identifying teachers who are successful in obtaining basic skill gains in students as measured by the CAT, it would be unlikely to assist schools in obtaining the widely held goal of increasing achievement test scores. Likewise, if the TPAI were shown to be a predictor of student achievement, ignoring this evaluative tool in developing and promoting teacher skill would be negligent unless substantive negative consequences could be demonstrated.

The wisdom of the state testing and evaluation policy may be debated. However, the purpose of the study was to establish the concurrent validity of inferences when the TPAI was considered a measure of teacher effectiveness. These inferences must be made within the curricular framework established by the NCSDPI. This framework

included achievement tests as an integral component to school improvement.

It is important to also point out that this study did not propose to evaluate teachers by the achievement gains of their pupils. What was being evaluated was the predictive ability of the TPAI to account for pupil achievement, a claim that was inherent in the instrument design. While it is quite possible that some effective teachers may teach an agenda that obtains basic skills in ways that pencil-and-paper achievement tests such as the CAT do not measure, it is equally possible that some teachers may not give each student an opportunity to learn the basic skills necessary to do well on these tests. School governing bodies in North Carolina have required each teacher to make these opportunities available to each student. The provision of the opportunities has been an integral part of the expected job function. It was well within the purview of these bodies to establish these curricular goals. If basic skills as measured by standardized tests were accepted parts of the curriculum prior to and during the course of the study (and their use indicates they were), then use of standardized tests as a measure of student achievement by this study was justified.

Limitations of Study

This study provided evidence concerning the criterion validity of the TPAI in light of its proposed use as an evaluative instrument to identify those teachers who are most effective in developing basic skills in their students. The study was limited to TPAI use in the elementary school. The criterion for this study was the CAT as commonly administered and utilized in the North Carolina schools systems. There are multiple criteria for validating any instrument. The validation evidence provided by this study cannot be considered definitive. Additional studies are needed to establish the validity of the TPAI in light of the broad inferences to be made on its results. This study, however, provides objective empirical evidence utilizing a design that may be readily replicated. It directly addressed the validity of common inferences made from the TPAI.

Summary

Teacher evaluations are considered important avenues for appropriate school governance and potential school improvement. The public's need for protection from incompetent or ineffective teachers cannot be ignored. Employment of recent research on teaching and teacher evaluation is a judicious course of action, but certain safeguards must be provided teachers. The interplay

between the public's right to effective teachers and the teacher's right not to be subjected to arbitrary or misleading evaluation creates a need for a framework of standards.

The joint Standards for Educational and Psychological Testing of the AERA, APA, and NCME provide prudent and respected professional standards to mitigate the conflict between the public's right to protection and the teachers' rights to a fair, unbiased and valid appraisal of their work. These standards indicate the need for a criterion validity study of any measure for licensure, certification, and promotion. This study provides evidence concerning the criterion validity of North Carolina's TPAI. The TPAI has been used as a licensure measure and has been proposed as a measure to determine teacher promotion.

The method of study was a statistical analysis used to develop a concurrent validity coefficient between the TPAI rating and a criterion of teacher effectiveness. The criterion selected for the study was the California Achievement Test which was mandated by the state of North Carolina and was used by the state to evaluate student academic competence in the basic skills. North Carolina used the CAT as preliminary evidence for mandatory summer school and suggested retention of pupils. Furthermore, the content validity study presented in North Carolina's Performance Appraisal Training Program (1986b) contained

109 references to studies of effective teacher practices using student achievement as the criterion. The majority of these refer to standardized measures of achievement.

The study also evaluated possible bias of TPAI ratings through the correlation study of selected student and teacher variables. While the presence (or absence) of a statistically significant correlation (alpha = .05 with a non-directional hypothesis) is not adequate evidence of unfairness, the presence of bias as indicated by these statistics may negate the test developer's claim for omnibus application among teachers of all assignments and experience levels.

Criterion validity evidence, as indicated by the Standards, is the primary responsibility of the test developer. Considering the lack of that evidence, the current study was an application of applied research that may have far-reaching consequences and enlighten an essential area of study in the evaluation of teachers.

Chapter II considers the historical antecedents leading to North Carolina's current interest in teacher evaluation. Past attempts to develop rating scales that predict student achievement are examined as well as the methodological problems inherent in those attempts. Finally, the effective teaching research used to develop the instrument and provide the content validity evidence is examined and its use as a validating criteria scrutinized.

CHAPTER II

A REVIEW OF THE LITERATURE

The current efforts to evaluate teaching have many precedents. The TPAI is a practical convergence of three of these streams of activity in education. The first stream is the continuing effort to evaluate individual teachers and teaching so that good teaching might be nurtured and rewarded. The second stream is the activity centering around the recurring issue of school direction and improvement. This topic generalized in the 1980's as the accountability movement. The third stream is the ever-increasing body of knowledge referred to as effective teaching research. This knowledge base is primarily composed of correlational studies of observed teacher behavior and student achievement criteria.

These three streams, evaluation processes and purposes (historical); career ladders, merit pay, and accountability (political); and the research on effective teaching (scholarly); each make a unique contribution to the formulation and utilization of the TPAI. To analyze the instrument solely on its scholastic merits or its political issues would surely result in a misunderstanding of the TPAI utilization in the North Carolina schools and distort the validity issues. Therefore, the literature of each

field as it pertains to the story of North Carolina's movement toward the creation and use of the TPAI as its primary evaluative instrument is presented.

An Historical Perspective on Teacher Evaluation

Socrates was executed in 399 B. C. for having corrupted the youth of Athens by his teachings. In 1616 Galileo received a formal warning that his teachings concerning the Copernican theory of planetary motion was contrary to Church teachings. He was imprisoned and, under threats of torture, told to recant his views. He obliged. In 1925 John Scopes, a Tennessee teacher, was placed on trial for violating a state law banning the teaching of the theory of evolution in the Tennessee pubic schools. He was found guilty, fined $100, and had his conviction overturned on a technicality. Teacher evaluation has an historical entangling relationship with censorship and political control of ideas and the education of youth. History records elaborate teacher evaluation procedures dating more than 2000 years ago. Doyle (1983) offered this vignette describing the avenue of remedy for the father who is unhappy with his son's teacher.

> In Antioch in about 350 A. D. any father who felt dissatisfied with the performance of the teacher in whose care he had placed his son had the privilege of examining the boy, or having him examined by competent authority, to determine whether the teacher might have been neglecting

> his duty. If the examination indicated that the teacher had indeed been neglectful, the father could enter a formal complaint against the teacher and have the case tried by a panel of teachers and laymen. Should the trial confirm the teacher's negligence, the father would be permitted to transfer his son--along with his patronage and fees--to another teacher. This evaluation would be an important matter to most teachers because ...[they often] derived the whole of their incomes from these fees. (p. 3).

The issues of governance and control of teaching are not new. They are surrounded by suspicion and justified by necessity.

The modern threads of teacher evaluation in the United States can be gleaned from the literature just decades after the close of the common school movement. Kappa Delta Pi initiated its research publications in 1935 with the publication of The Measurement of Teaching Efficiency (Lancelot et al., 1935) and was subject to a review committee containing, among others, E. L. Thorndike and W. H. Kilpatrick. There exists in the monograph an early recognition of the magnitude of the problem involved in doing teacher evaluations. Thus the editor was prompted to advise the reader:

> The reader who hopes to find here a blueprint giving him a short and easy way to judge the efficacy of teaching will be disillusioned. The more thoughtful reader who is willing to try to understand the all but insurmountable difficulties of the problem will find in these studies relationships worthy of his careful attention, as well as new and stimulating methods of attack. (Walker, 1935, p. ix).

This advice is as applicable to readers of this research endeavor as it was to readers in 1935. The results of teacher evaluation studies are never definitive and almost always illustrative of the difficulty of evaluating teachers.

The Role of Values in Establishing Validation Criteria

Several difficulties had presented themselves to the early researchers that prevented a satisfactory resolution to the task of measuring teacher efficiency. Educators have been unable to agree on "who is a good teacher or what are the concrete manifestations of teaching ability" (Walker, 1935, p. x). The diversity of philosophical approaches forbad the construction of any universal measure of teaching ability. Thus, the problem of value judgments preceding and structuring an ostensibly empirical activity was painfully evident to teaching researchers by 1935. The questions of values in validity studies cannot be ignored. Cronbach examined the role of values in establishing criteria for validation studies.

> When observations at the end of instruction are used to determine how successful some educational activity has been, the interpretation embodies value judgement. If the values are not acceptable, the conclusion is not acceptable. An evaluation battery is a collection of procedures used to decide whether a given educational program is satisfactory, whether the individual student has made satisfactory progress, etc. The conclusion that posttest performance is

> satisfactory (or unsatisfactory) is warranted only if there is a match between the test content and educational aims. Hence the validity of an evaluative conclusion depends on the value question: Did the tests appraise the qualities I consider it most important to teach? That question might elicit a positive answer from one educator and a negative one form another looking at the same tests (p. 459).

The diversity of philosophical approaches among teachers, the varying hierarchy of goals in instruction and differences in the needs of students create for teachers an environment where teachers must act in adaptive and flexible ways. Teacher values as well as institutional values shape not only teacher behavior, but the objectives the he or she may select. In a field where circumstances are fluid and the needs of children are diverse, the independently functioning teacher will rely on his or her values to guide and direct the instructional program offered to the class of children. The dilemma of the evaluator is how to assess the independent activity of teachers and assess the basic instructional goals common to all classrooms.

Assessing Teaching by Teacher Behavior

A second problem continues to perplex current researchers. Because of a lack of agreement in desired pupil outcomes, teaching must be measured, it was thought, directly by assessment of the teacher rather than

indirectly through pupil change. The latter is far more desirable, particularly in light of validity considerations because changes in pupils are precisely the outcomes for which teaching is designed.

Walker (1935) concluded her analysis of teacher evaluation studies by succinctly summarizing the fundamental problem in the validation of teacher effectiveness and teacher rating:

> The lack of an adequate, concrete, objective, universal criterion for teaching ability is thus the primary source of trouble for all who would measure teaching. One typical method of attack used in rating scales is to compile a list of broad traits supposedly desirable for teachers, with respect to which the rater passes judgment on each teacher. This amounts to an arbitrary definition of good teaching, which is subjective and usually vague, but it does not necessarily lead to an identification of it. Only if the traits themselves can be reliably identified can their possessor be identified as a "good teacher" according to the definition laid down in the scale. Even when the scale is made quite specific, relating not to general traits but to concrete procedure, the fundamental difficulty remains, that there is no external and generally accepted criterion against which the scale can be validated to establish the significance of its items. (p. xi).

Correctly identifying and classifying teacher behavior is a complex task. Selecting and defending those behaviors to be used as criteria is even more complex. The inevitable difficulty is that teacher behavior may be designed to bring about a multitude of results. Spotting teacher behaviors thought to be effective in bringing about a particular result is only part of the evaluation task. How

well suited the teacher activity is toward realizing the specific objective in the instructional period is certainly appropriate and probably essential to fair evaluation. Behavior, without knowledge of its intent, is only a partial criteria for evaluating teachers.

Student Variables as Predictors of Teaching Success

A third problem was also becoming evident and was confirmed, if only by controversy, with the Coleman studies (1966). Researchers strongly suspected in 1935 that a child's subject matter achievement is more closely related to his own ability and previous learning than to the instruction he or she receives from a teacher. Likewise, it was noted that pupil success was related to factors other than student ability and quality of instruction. How to attribute student achievement to individual factors remains a monumental difficulty in effective teaching research.

Early Failures of Teacher Evaluation to Enhance Instruction

The three studies in The Measurement of Teaching Efficiency (Lancelot et al., 1935) were unsuccessful in defining successful teaching practices but were uncannily accurate in isolating the major difficulties in teacher

evaluation that have plagued subsequent researchers. But the elusiveness of success did not dampen the move toward measuring teacher contributions to the education of children. By 1945 the evaluation of teachers was widespread and many researchers were concerned with the state of affairs. Conducting a major study concerning the evaluation of teacher merit, Reavis and Cooper (1945) acknowledged the spotty record of teacher evaluation and the increasing necessity to provide fair and accurate assessment.

> The evaluation of merit is a matter of great importance both to officials responsible for the management of the schools and to teachers interested in professional security. Boards of education insist that the merit of teachers be carefully evaluated and not be taken for granted or determined by the snap judgment of professional officers. Likewise, the teachers whose professional careers are at stake object to the perfunctory ratings which are made by administrative officers and which are frequently used in determining salary, promotion, and professional security. All recognize that some evaluation of merit must be made by school officials responsible for the service of teachers. The critical issues are the purpose of the evaluation and the means by which it is made. (p. iii).

A desire to protect the pupil from ineffective and miseducative experiences and be objective and fair to teachers was a sign of a maturing profession. However, this concern had not led to a formulation of the appropriate knowledge to perform this task. This lack of knowledge coupled with an imposing political desire to police teacher ranks <u>created</u> the gloomiest problem

involving teacher evaluation. Reavis and Cooper (1945) documented the building resentment of teachers toward unwise and perfunctory evaluation:

> It is true that many of the means employed in evaluation have proved to be unsatisfactory. The reasons are not difficult to find. Some of the means have been borrowed from civil service and business administration, without having been adapted to the evaluation of teaching. These means of evaluation have been arbitrarily adopted in some cases by administrative officials and imposed upon unwilling subjects who have had no voice in the preparation of the instruments and in the methods of their use. As a result a general antipathy toward the evaluation of teacher merit has developed in many school systems. Furthermore, evaluation has been so unscientifically done and so unwisely used by some school officials that the teachers in these school systems have come to regard evaluation (generally called "rating") as a necessary evil to be endured. Under such conditions the attitude of teachers toward evaluation is naturally unfriendly. Unfortunately improvement in attitude can scarcely be expected until benefits from evaluation are actually experienced by the teachers concerned (pp. iii-iv).

This type of arbitrary summative teacher evaluation cannot boost teacher morale and enhance effective teaching skills. The frustration that often accompanies evaluation existed in 1945 and accompanies the current use of the TPAI ("Fixing", 1987; Keever, 1987; Williams et al., 1987).

Perceived Needs for Teacher Evaluation

The desire to evaluate teachers persists. Lawmakers and educators are concerned about the quality of school experiences provided the youth in schools. There is also widespread perception that public education dictates economic viability and social stability. Teacher evaluation, it was thought, would assure quality educational experiences, spot teacher deficiencies and provide an impetus for remediation.

Harris (1986) provided a list of needs for teacher evaluation. The needs for administrative control and data for decision making dominate the list as they did for Reavis and Cooper (1945). Included in Harris' list is the need for indirect reassurances of educational quality to parents who are now assumed to no longer have close personal contact with the teachers. Harris (1986) pointed out that teachers need evaluation to fulfill their own personal and professional needs: "The classroom teacher or instructor tends to perceive himself with considerable uncertainty and ample distortion, and hence needs reliable feedback from external sources" (p. 2).

Although the argument that teachers are primary beneficiaries of teacher evaluation is ubiquitous, there is little evidence to be found that teachers advocate such policies. Teacher groups have historically expressed

reservations concerning the practice and have generally worked toward limiting their impact and, as a substitute, focused efforts toward promoting teacher growth through education, project participation, and field experiences. In North Carolina charges have been made that teacher groups, by covertly resisting evaluation tied to pay, have caused the objective stringency of one evaluation system that "is so rigid it is irrational" (Keever, 1987, p. 36).

Doyle (1983) also contended that teacher self-improvement and growth is a fundamental purpose of teacher evaluation. However, there has been little empirical evidence presented to support Doyle's conclusion. More often, documents in defense of evaluation programs are presented by researchers that note the cooperation of teachers and enhanced communication between principal and staff (Pigford, 1987; NCSDPI, 1965). What teachers say among themselves, however, may be quite different (Keever, 1987; Williams at al., 1987; "Fixing", 1987) citing a need to "play the game."

Millman (1981), in editing a summary of the state of teacher evaluation for the National Council on Measurement in Education, argued that teacher evaluation is an inevitability. The active questions, he concludes, are "Who should evaluate? For what purpose? Using what means?" (p.12). These unanswered questions are the same issues facing Lancelot (et al., 1935) over fifty years ago.

With thousands of pages of literature published, countless debates conducted, and millions of teacher evaluations performed, the fundamental issues of teacher evaluation have historically persisted, basically unaltered and unilluminated.

Accountability, Merit Pay, and Career Ladders: Historical Antecedents to the TPAI

The current interest in teacher evaluation and its role in accountability, merit pay and career ladders has a continuing history in the United States. Merit pay schemes were advocated in 1898 in St. Paul, Minnesota; in 1904 in Newton, Massachusetts; and in 1913 in Baltimore, Maryland. By 1918 48% of 309 city school districts studied by the National Education Association were using some variation of merit pay (NCSDPI, 1965).

These early scales were based on broad categories of teacher traits thought to be desirable in teaching and, by way of being desirable, effective in bringing about student gains in achievement. Barr was reported to have classified ten categories of all attributes used in teacher rating. Included were instruction, classroom management, professional attitude, choice of subject matter, health, cooperation, personal habits, discipline, personal

appearance, and appearance of room (NCSDPI, 1952; NCSDPI, 1965).

Interest in merit pay declined in the 1930's and merit pay schemes were abandoned. Reasons cited for the decline were:

> -economic conditions of the early thirties
> -failure of merit programs to accomplish their avowed purposes
> -difficulty of judging the exact amount of pupil improvement attributable to any one teacher in view of a rapidly expanding curriculum
> -recognition of the development of many good methods of teaching
> -awareness that the school is only one of many educational influences in the community (NCSDPI, 1965, p. 2).

In viewing the historic trends in retrospect, the interest in merit pay schemes based on teacher ratings seem to accompany economic and political strife and disillusionment. The muddled political state and frustration of World War I accompanied the dramatic rise in teacher merit pay schemes reported in 1918. Likewise, the Second World War brought a resurgence of interest in the 1940's. In the 1980's a decline in the world influence of American business and a perceived breakdown of social values has rekindled a critical examination of American education and an interest in the merit pay issue, now reborn as career ladder plans.

This historical correlation lends credence to the view that Americans look to their schools for the substance and realization of their dreams. As a minimum, it can be

argued that interest in merit pay based on ratings of teacher performance and the accompanying rating scales used to discriminate among teachers has been politically motivated by a perceived dissatisfaction with the current status of the schools. There is evident no management nor research precedent to demonstrate that merit pay has supervision properties shown to be worthy of emulation (Darling-Hammond, 1986). These political motivations are important to note because under such an ephemeral a criterion as political necessity, marginal evidence of validity may indicate marked success to decisions makers. North Carolina's history of interest and experimentation with merit pay based on teacher ratings in the 1940's and again in the 1960's have definite political roots. They also have been straightforward examples of reasoned approaches to personnel management.

The Political Basis for Teacher Evaluation

The North Carolina General Assembly first authorized the Commission on Merit Rating of Teachers in 1945 to study the feasibility of establishing teacher pay based on the ability of the individual teacher. The Commission proceeded with its mission by an exhaustive literature review, a study of current practices in merit rating, consultation with major educator organizations and

consultation with A. S. Barr of the University of Wisconsin and W. A. McCall of Teacher's College, Columbia University.

The results of that study were reported in a printed bulletin Report of the Commission on Merit Rating of Teachers (1946). After extensive investigation of current practices in teacher evaluation by rating the Commission concluded it "had been unable to find an instrument for measuring teaching efficiency which can be accepted as valid for determining salaries" (NCSDPI, 1965, p. 11). The Commission further expressed its belief that such an instrument could be constructed.

The McCall Studies

The creation of such an instrument was the charge of the State Education Committee in 1947. Four school districts were invited to participate. William A. McCall, professor of education, Columbia University, was hired to direct the research on this project. The project initiated on a note of candor as Dr. McCall warned the Committee of the possibility "that the science of education had not yet advanced far enough to permit a satisfying study of such a complex matter as the merit of teachers" (NCSDPI, 1965, p. 12).

McCall indicated a straightforward and appropriate design of such an instrument. He proposed

> to measure comprehensively the growth produced in each class by the teacher of that class, to

> weight the elements of the growth according to importance, to secure as a single composite figure for all the growths made by each class, to correct this weighted crude growth for the capacity of the class to grow, for differences in class size if the latter appeared to influence growth, and then to correlate a large number of measures of the teachers' traits with this purified criterion of each teacher's worth as teacher (NCSDPI, 1952, p. 10).

The results of the study again reiterated that the evaluation of teachers utilizing a single rating scale raised grave validity considerations:

> -The simple, inexpensive rating by superiors lacked sufficient validity to justify its adoption.
> -The method of measuring teachers' merit by measuring the growth each teacher produced in his pupils is workable and can be extended to all grades. But the trouble and expense involved make the systematic use of such a method unwise.
> -The findings of this study show that a battery of the measures used in this research could be assembled that would be much more valid than the State's existing system of measuring merit by training and experience; but that the expense and complexity of such a battery make its use prohibitive for all teachers (NCSDPI, 1952, pp. 36-37).

The Committee's preference of measures of student growth was a direct reflection of validity concerns. The Committee also called for multiple criteria to reflect teacher efficiency. The complexity of data collection and its subsequent analysis employing multiple criteria was an unwieldy process in 1947. Today, even with high-speed optical scanning machines and digital computers alleviating much of that difficulty, the expense (and perhaps the human comprehension) of such a system is still a formidable

obstacle. McCall was unable to develop a system of rating that he felt was valid for pay decisions. He was, however, unequivocal in his opinion of rating scales:

> The research failed to find any system of measuring teacher merit which the writer is willing to recommend be adopted as a basis for paying the salaries of all teachers. This study did establish that the existing system is of little value if salaries should be paid on merit, and the system of merit rating by official superiors which the State was considering for adoption is of no value (NCSDPI, 1952, p. 37).

The Committee was concerned about the validity issues of proposed measures of teacher effectiveness. This preoccupation was not shared by later studies funded by the North Carolina legislature. It was the Committee's opinion that rating scales were unlikely to ever yield a valid measure of teacher merit. The Committee did feel the studies made important contributions to improving instruction.

> The most valuable discoveries of this research are the characteristics which differentiate good teachers from poor teachers. This permits us to paint a partial picture of the ideal teacher, thereby making possible guidance of the proper young persons into teaching, selection of candidates for training, diagnosis of deficiencies in trainees, revision of the program of teacher training in college and in service, and guidance in developing additional instruments for measuring progress toward the valid goal of all training... (NCSDPI, 1952, pp. 37-38).

Resolution 80

The next flurry of interest in teacher rating occurred in 1959 when the General Assembly adopted Resolution 80.

This measure prompted still another study of pay plans for school teachers based on ratings of ability (NCSDPI, 1965). Another commission was appointed. This commission narrowed its focus on past studies including the 1947 Merit Pay Study and issued its findings based on history and debate. A summary of this commission's findings is significant in that it mirrors the contemporary state-of-the-art in teacher ratings of efficiency:

> *Though merit rating is no substitute for intelligent professional leadership, it is a complementing factor to preservice preparation, in-service training, an atmosphere conducive to learning, and provision of teaching facilities and materials.
>
> *There is much sentiment throughout the country against merit rating, with much of the criticism centering around three major areas of concern: wide differences in definitions of good teaching; the measuring instrument itself; and evaluators and the merit evaluation process.
>
> *There is significant evidence that differences in teaching ability may be identified, though there is no single validated instrument acceptable to the entire teaching profession.
>
> *Measurable achievement change in students is but one possible factor among many in measuring relative teaching ability or success and far from being an acceptable basis in itself.
>
> *Criteria of superior teaching, acceptable to teachers and school patrons, should be cooperatively developed at the local level.
>
> *Evaluators must be highly skilled in the process of evaluation (NCSDPI, 1965, p. 14).

The commission reiterated to the General Assembly the need for an adequate salary schedule capable of "attracting and holding qualified individuals sufficient to meet the

demands for teachers" (NCSDPI, 1965 p. 14). Further, the commission felt that systematic experimentation with merit pay schemes tied to teacher ratings should be conducted.

North Carolina Teacher Merit Pay Study

As a response to the 1959 Commission Report, the General Assembly authorized the North Carolina Teacher Merit Pay Study in 1961. The study encompassed four years and involved volunteers in three pilot centers. Each district in the study established a local merit study committee. It is unclear why this study relegated to local committees tasks that experts had been unable to complete successfully. Local committees were charged with a wide variety of tasks which included:

1. establishing a statement of philosophy and objectives for the local study
2. devising and adopting techniques for surveying attitudes and morale
3. developing and adopting a set of criteria which can be used as a basis for evaluating teacher performance
4. determining criteria for the selection of observers and final evaluating officials
5. prioritize factors to be recognized in evaluations

6. study the relationship between merit programs and ethics

The resulting projects were much less systematic than the 1947 McCall Study or other studies of this type (e.g. Lancelot, et al., 1935; Reavis and Cooper, 1945; Nelson, Bicknell, and Holland, 1956).

The three local committees were advised to "Feel free to call on State Merit Study officials for help at any time" (NCSDPI, 1965, p.21) thus inferring wide latitude of freedom for individual districts and a loose supervision philosophy from state officials. Predictably the results of the study lacked significant hypotheses to test and therefore yielded uninterpretable results. Data analysis centered around the sex, race, seniority and degree distributions of participants by region and total, the percent who volunteered, dropped out of the program, received merit pay, and the like. These descriptive statistics in no way supported any inferences about the validity or effectiveness of what was done. In totals for all districts in the study, 22% of those receiving the merit pay award during the 1962-1963 school year chose not to participate during the following year (NCSDPI, 1965, p. 58).

A teacher questionnaire was administered to the participants. Typical items in the questionnaire were:

12. There are practical, satisfactory methods of administering a program of merit pay.

21. Evaluators aimed at improving instruction should be independent of salary determination.

22. A program of observations and conferences, with emphasis on helping teachers improve, is of more value than a program aimed at evaluating teachers for merit pay (NCSDPI, 1965, pp. 137-144).

The results were not supportive of teacher evaluation for merit pay. Question 12 resulted in 21% agreeing, 35% undecided, and 44% disagreeing. Likewise the results for Question 21 were 76%, 14%, and 10%. The results for Question 22 were 85%, 11%, and 4%. Clearly, the grass-roots approach to teacher evaluation and merit awards failed to convince participants of its value and validity.

A major portion of the report included the views of the report writers concerning the strengths and weaknesses of the evaluation approaches and the merit pay connection. Merit pay schemes were considered stop gap measures for covering the insufficiencies in the system for generating good teaching. The report cited as a negative finding the opinion "When teaching conditions are excellent, when teachers are well selected, and when an effective in-service program is in operation, merit rating is superfluous" (NCSDPI, 1965 p. 9).

It was further asserted that merit pay schemes based on merit ratings (summative evaluations) caused morale problems. "Merit-rating plans tend to create problems in teacher relationships or morale--problems related to

jealousy, fear, favoritism, tension, undesirable competition, and insecurity" (NCSDPI, 1965, p. 9). Closely associated with this line of criticism was the accusation that merit pay based on evaluation tends to be divisive. "Merit programs tend to develop divisive and competitive attitudes rather than cooperative attitudes among teachers; for this reason, such programs are psychologically disintegrative" (NCSDPI, 1965, p. 9).

The negative opinion expressed in the report also indicated an apprehension of the effects of merit pay on teaching. There was a fear that merit pay and its concomitant antecedent, teacher rating, would discourage creativity and innovation in instruction. "Programs of merit rating tend to discourage creativity in teaching. Instead, a premium is placed on conformity and rigid adherence to stereotyped criteria. Conformity, it is felt, is the enemy of academic freedom" (NCSDPI, 1965, p. 9). The connection between this criticism and critics of the TPAI ("Fixing", 1988; Keever, 1988; Williams et al., 1988) is poignant. The criticisms are virtually identical.

Teachers generally felt that merit pay schemes did not reflect the views and concerns of the teaching profession. Instead summative evaluations tied to pay were considered to be imposed and external.

> Merit pay is generally condemned by teachers as individuals and by their professional organizations throughout the Nation. Conceived and practiced for the most part by nonteaching

> groups, merit-rating plans are felt by teachers in general to be imposed and consequently of no benefit in improving education (NCSDPI, 1965, p. 9).

Teaching is commonly thought to be best when done in a cooperative environment. Since resources are limited candid sharing of suggestions for improvement must be valued and utilized if maximized effects are to be realized. Cooperation among teachers and supervisory personnel are considered integral components of efficient schooling and school policies should promote cohesive bonds among faculty. The merit pay report was also suspect of teacher rating as a device for determining salaries. "The specter of rating tied to salary disturbs the friendly and frank relations which should exist between teachers and their professional cohorts" (NCSDPI, 1965, p. 9).

Perhaps the most damaging criticism from the study was the accusation that teaching is too complex to evaluate objectively.

> Teaching is an art as well as a science and is too complex to be evaluated objectively. Thus far, it has been impossible to measure teacher competence accurately because of the human qualities in evaluators. Excellence in teaching resists measurement (NCSDPI, 1965, p. 9).

Examining the instruments available for the purpose of evaluation, the writers expressed the opinion that

> Merit rating ultimately depends on subjective judgments. No valid or reliable instrument has yet been developed for measuring teacher effectiveness or the total growth of students, which involves acceptance of responsibility, growth in values, ability to think, development

> of understanding, the instilling of proper attitudes and moral standards, understanding of self, and other intangibles (NCSDPI, 1965, p. 9).

The TPAI instrument does not address the broad concerns of effective and meritorious teaching that occupied the Merit Pay Study committee authoring the 1965 report to the General Assembly. However, these broad yardsticks are as potent a criteria for effective teaching to TPAI developers and users as they were in 1965.

Understandably, the major finding of this study was that "A uniform, statewide program of merit pay is not feasible nor practicable at this time" (NCSDPI, 1965, p. 113). The evidence to support this conclusion was primarily that of a failure to find acceptance among professional ranks. Nowhere among the major findings and recommendations are issues of validity and reliability mentioned in conjunction with empirical data.

The Scholarly Thrust in Teacher Evaluation: The Move Toward Empirical Research

In 1978 the General Assembly initiated a new interest in teacher evaluation. The first of these investigations addressed the qualifications of initially certified personnel (ICP). The Quality Assurance Plan (QAP), initiated in 1978, changed the selection process, education, and support of beginning teachers. The goals of this program were to enhance the training of teachers and

provide a support network during the novice teacher's first three years of teaching. Periodic evaluations of these personnel were considered essential for formative development and an important summative criterion for decisions concerning contract renewal and tenure. By 1980 the Performance Appraisal System instituted annual evaluation of all teachers state-wide using criteria and standards adopted by the State Board of Education. These standards were an initial formulation of the TPAI.

The North Carolina Teacher Differentiation/ Differential Pay Study collected input and reaction from school personnel. This study concluded there was strong sentiment "that teachers should be paid according to their level of effectiveness and responsibility as well as their experience and educational background" (Holdzkom and Kuligowski, 1987, pp. 3-4). By 1985 legislative reform of schooling was at a peak nationwide. North Carolina had made sweeping revisions. Statewide curriculum guides had been developed in North Carolina for all grades and subjects, statewide testing programs had been initiated and mandatory summer school for third, sixth, and eighth grades based on the statewide testing programs had been established. High school competency testing was in place and new certification standards for teachers had been enacted. Several schools of education were put on notice that unless changes were made in their educational

programs, they would no longer be able to issue state certification. In 1985 the General Assembly established pilot career development plans in sixteen local school systems as part of a four year study on career ladders. By 1986 the newly developed TPAI was in use as the evaluation instrument establishing major criteria for both certification and tenure of ICP and differential pay in the sixteen experimental career ladder plans.

By 1987 the TPAI instrument was the required course of evaluation as school districts geared up to implement the new career ladder plans being developed. Teachers were introduced to the plan via a 30 hour workshop entitled Effective Teacher Training. Likewise, potential evaluators had completed an additional 30 hour program called the North Carolina Performance Appraisal Training Program. These programs represented a new twist in the use of teacher rating scales: They were ostensibly based on empirical research and personnel were to receive extensive specialized training in their use. The TPAI was to support the inferences of teacher adequacy and teacher excellence and serve as a vehicle for teacher improvement. To assist that improvement, teachers should be trained in the techniques advocated by the evaluation policy. This was deemed a reasonable approach because it was thought that there was an adequate body of knowledge to establish a uniform core description of effective teaching practice and

that these practices could be reliably measured by rating scales.

Limitations of Effective Teaching Research

In the 1960's, educational researchers began to examine educational effects by a meticulous analysis of what teachers were actually doing in the classroom. The studies also attempted to connect teaching activity to various student outcomes. This research approach immediately met with major methodological difficulty. In examining the overt behavior of teachers, researchers were confronted by the intense complexity and variety of teacher activity and the ends to which that activity was directed. The criterion to be used to assess the effects of teacher behaviors left many unanswered questions. Researchers struggled to maintain objective measures in a profession dominated by subjective outcomes. Attempts to define variables in operational terms led researchers to look for narrowly defined criterion measures. These measures, by nature of the needs for precise low inference items, tended to appear trivial. Adding to the difficulty were doubts about the utility of effective-teaching research findings, the dependence on correlation studies and a lack of experimental designs.

Criteria for Teacher Effectiveness Research

The effective teaching researchers, often known collectively as process-product researchers, asserted several criterion to evaluate teacher behavior. Two criteria developed dominance in the effective teaching research field. The leading criterion was gains in student learning as measured by pencil and paper tests, typically some form of the multiple-choice standardized achievement test. A second criterion utilized widely was the percentage of students' time spent engaged in the designated activity (NCSDPI, 1985b). This has commonly become known as "time-on-task." Time-on-task was considered a valuable criteria by effective teaching researchers because of a relatively strong correlation with achievement test gains.

Another criterion used was measures of student behaviors which calculated the percentage of time students were conforming to the teacher's stated behavioral requirements. Some researchers have taken these findings to the extreme and have advocated teaching what achievement tests measure. Popham (1987) advocated measurement driven instruction methodology where teaching is directed to specific objectives measurable by pencil and paper tests. With this approach to teaching, teacher effectiveness study designs would, of course, be greatly simplified.

Walberg (1974) acknowledged the limitations of using narrowly defined behavioral outcomes as criteria in teacher evaluation studies. Walberg concluded that "qualities of the educational environment that are consistently associated with growth on standardized cognitive and affective outcome measures are valid to some extent" (p. 2). Walberg warned against using simple criteria to the exclusion of more complex and less easily measured variables. Progress in obtaining highly valued outcomes such as growth in creativity and democratic ideals must also be recognized.

Even if agreement over criteria could be reached and standardized tests were to play a prominent and universally accepted role, other difficulties faced the interpretation and use of effective teaching research results.

Stability, Correlation and Causation

Rosenshine (1970, 1973, 1977) indicated by an extensive review of studies that teacher effects were unstable across time, student population, and subject matter. Brophy (1974) summarized the situation:

> These figures obviously suggest that teacher effectiveness in producing student learning gains is not a stable "trait," that a teacher who produces large gains in his students this year is not necessarily going to do the same the next year. Such results, if they accurately reflect the general case, threaten the validity of process-product teacher effectiveness research (p. 34).

While facing validity threats due to a lack of stability, process-product methodology also could not defend the inference that teaching behaviors demonstrating a consistently high correlation with achievement were the cause of the achievement. Process-product methodology primarily searches for correlational relationships between quantified teacher behaviors and quantified student outcomes. These studies are almost exclusively observational (in naturalistic settings) and not experimental. Designs also tended to be atheoretical, assuming instead a post hoc analysis. There are severe limitations to the utility of such results. Glass and Hopkins (1984) noted that the presence of correlation between two variables does not necessarily imply a relationship of causality. Although correlation can be helpful in identifying causal relationships when combined with other methodologies, it is insufficient evidence to support a causal inference alone. Glass and Hopkins cited three reasons for this:

> First, even when one can presume that a causal relationship does exist between two variables being correlated, r_{xy} can tell nothing by itself about whether X causes Y or Y causes X. Second, often variables other than the two under consideration could be responsible for the observed association. Third, the relationships that exist among variables in behavioral and social sciences are almost always too complex to be explained in terms of a single cause (p. 104).

Further, Glass and Hopkins (1984) explain that just as a positive relationship cannot be construed to support

causation, a zero or even negative correlation does not rule out the possibility of a positive causal relationship. The value of correlations is in their ability to predict. This does not mean that teachers who are instructed to emulate a particular behavior that had consistently shown a strong positive relationship with student gains in achievement will bring about any increase in student achievement.

Acknowledging the limitations of correlates, Brophy (1971) maintained that before causal inferences can be established, experimental and quasi-experimental designs must be created and these variables manipulated. Unfortunately, these studies have not been forthcoming in sufficient numbers to warrant causal inferences.

A Summary of Process-Product Findings

Until 1972 fewer than 25 studies had been conducted on any specific aspect of teacher behavior. Since 1976, the literature has shown a growing interest in teaching effectiveness as measured by student achievement (Triosi, 1983). The research since 1974 has yielded a pattern of instructional techniques which have shown consistent links with student achievement gains. Rosenshine (1986) described this pattern as "a systematic method for presenting material in small steps, pausing to check for student understanding, and eliciting active and successful

participation from all students" (p. 60). The findings have been grouped into various patterns of instruction (e.g. Triosi, 1983; Brophy, 1987; Holdzkom, 1987; Rosenshine, 1986) for dissemination and further study. The pattern is that of a fairly traditional teacher (Triosi, 1983).

Generally these summaries have advocated high levels of teacher direction, a whole class approach and teacher demands (or explicit expectations) that students pay attention to instruction. Effective teachers take and exercise responsibility for classroom management and discipline. Instructionally, effective teachers tend to begin with a review of relevant past learning, express an attitude of task orientation, frequently probe for evidence of student understanding, monitor student progress closely and provide corrective feedback.

In managing student behavior, these summary reviews of effective teaching literature generally argue that more effective teachers make clear rules and enforce them, provide student work that allows a high rate of success, are businesslike in their approach to school routines and use direct instruction with the whole class or small groups for basic skill mastery (NCSDPI, 1985a).

These descriptions vary little from the craft knowledge handed down from teacher to teacher over decades. What effective teaching research tends to document best are

the results of such practices. Rosenshine (1986) summarized the method and describes the objectives for which effective teaching practices are most effective:

> This pattern is a systematic method for presenting material in small steps, pausing to check for student understanding, and eliciting active and successful participation from all students.... Specifically, these results are most applicable to the teaching of mathematical procedures and computations, reading decoding, explicit reading procedures such as distinguishing fact from opinion, science facts and concepts, social studies facts and concepts, map skills, grammatical concepts and rules, and foreign language vocabulary and grammar" (p. 60).

These techniques are not appropriate for all instructional objectives. For example, the North Carolina State Department of Public Instruction in the North Carolina Standard Course of Study and the North Carolina Competency-Based Curriculum state that the purposes of the state curriculum are "(1) to help students become responsible, productive citizens and (2) to help students achieve a sense of personal fulfillment" (NCSDPI, 1985c, p. 5).

Holistic philosophical missions such as those advocated by NCSDPI do not fit the effective teaching research pattern of successful teaching endeavors. Rosenshine (1986) summarized the areas of limitations in effective teaching research:

> These findings are less relevant for teaching in areas that are less well structured, that is, where the skills do not follow explicit steps or the concepts are fuzzier and entangled. Thus the results of this research are less relevant for teaching composition, writing of term papers. reading comprehension, analyzing literature or

> historical trends, for the discussion of social issues, or for teaching entangled concepts such as "liberal" or "modernism" (p. 60).

These limitations include the types of educational objectives that are essential to developing informed citizenry and personal fulfillment.

It is the inability to address these analytical higher-order skills that fuel criticism of effective teacher research. The philosophy and rationale of the Standard Course of Study in North Carolina call for esteemed learning goals such as critical thinking, communication skills, positive attitudes towards oneself and one's own culture, a sensitivity to the needs and feelings of others, a willingness to cooperate with others in working toward a common goal, and the ability to understand and cope with a constantly changing society.

Summary

Chapter two has provided an overview of three streams of activity leading to the development of the TPAI. The historical stream reviewed the use of rating scales in teacher evaluation and identified the major problems faced when evaluating instruction. The political stream traced the development of legislative activity mandating several studies in teacher evaluation in hopes of developing fair systems to pay teachers by merit. The scholarly stream

identified the major findings and methodological difficulties of effective teaching research.

Chapter three will outline the method of study in this research endeavor. Methodological problems facing the study are identified as well as the solutions for current purposes. The sample selection methods are stated and data collection techniques described. The procedures for the analysis of the data and derivation of the validity coefficients are explained.

CHAPTER III

METHOD OF STUDY

The method of study in this validation project was a non-obtrusive non-reactive data collection program followed by a statistical analysis designed to ascertain validity coefficients between teacher TPAI function scores and student achievement. In any study directed toward establishing the criterion validity of a teacher evaluation instrument, it is essential that certain methodological problems be dealt with at the outset. Three problems required treatment in the early design phases of this project. They were the selection of the validating criterion, statistically equating classrooms and countering the regression effect. Each will be considered in turn.

Selection of the Validation Criterion

The chief methodological design problem in a criterion-validity study is the selection and defense of the validating criterion. As Cronbach notes (see page 29 this document) the validity criterion is underpinned by an expression of values. However, values can and must be defended by reasoned argument. As stated earlier, student achievement as measured by a pencil-and-paper test is a limited perspective of what is expected of schools in the

educational programs provided students. Student achievement as measured by these tests, however, is a part of the expected school mission. Whatever the tests' limitations and imperfections may be, these tests do have relatively strong validity studies representing the various constructs of learning they purport to measure. More to the point, the public and the agents which it elects in the form of governing boards and lawmakers expect directly measurable learning experiences to be a part of the curriculum. It is the duty of these governing bodies to establish policy to guide and direct schools toward the attainment of society's goals. This, of course, is done within a constitutional framework of fairness and respect for the individual rights of parents, students, and school personnel. If a teacher evaluation system, such as the TPAI, accurately identifies those teachers who are realizing established goals such as direct instruction, inferences regarding effectiveness of teachers can be made.

Controlling Influential Student Variables

A second methodological problem facing this study was the substantial proportion of variance accounted for by student variables. Correlations between individual pupil's intelligence and measures of achievement are generally reported from .40 to .70 (Medley, Coker & Soar, 1984).

However, if correlations are based on class means, as is often the case when calculating various measures of teacher effectiveness, the correlation can go as high as .90 accounting for 80% of the variance in achievement among classes. If these factors were left uncontrolled and a researcher evaluated the effectiveness of the teacher based on these achievement variances, 80% of the variance in achievement could be the result of pupil differences before the teacher had any chance to influence them (Medley, Coker & Soar, 1984).

There are statistical techniques to exercise some measure of control over the differences in classes, yet none are sufficient to match the power of a design based on randomized student assignment. For researchers who must deal with non-randomized pupil assignment, controlling the effects of differences in student variables is essential. Partial correlations can be computed holding other variables constant, that is, neutralizing or "partialing" out the effects of influential variables. However, partial correlations holding influential variables constant cannot be equated to randomized designs. The latter still remain preferable but the difficulty of obtaining research situations where pupil assignment is made at the convenience of the researcher is massive. This study collected data on a variety of student variables for the purposes of statistically equating classroom populations.

Countering the Regression Effect in Calculating Student Achievement

The third methodological problem facing the study was controlling the influence of the regression effect in pretest-posttest correlation. The regression effect was first documented by Francis Galton (1822-1911) in his study of the relationship between heights of fathers and their sons. He noted that fathers who were taller than average tended to have sons who were also taller than average but not as tall as their fathers. The effect was the same with fathers who were shorter than average; their sons were also shorter than average. Surprisingly, their sons tended to be taller than the fathers. Galton labeled this regression toward the mean the law of filial association. In actuality, there is a regression effect when any two variables are not perfectly correlated (Glass and Hopkins, 1984; Medley, Coker and Soar, 1984).

Medley, Coker and Soar (1984) give an illustration of the regression effect. They pose the case of a group of students taking a pretest followed by a similar posttest. The likelihood of any student scoring in the extreme ten per cent above (or below) the mean also scoring in the relatively same position on a readministration of the test is quite small regardless of the intervening treatment.

This can be shown graphically by creating a scatterplot of pretest and posttest results. The extreme 10% of each measure is circled on the graph. Only a small portion of the pretest extreme and posttest extreme overlap. The greatest part of both measures' extremes remain unique to that measure (Medley, Coker and Soar, 1984). Medley, Coker and Soar (1984) present it in a more vernacular fashion:

> The term regression effect comes from the fact that each extreme group tends, on the average, to regress toward the mean from one measurement to another. One of our students, colorfully, characterized this as a "Robin Hood effect," since it steals from the rich and gives to the poor (p. 37).

The consequences of this effect can be dramatic and entirely misleading when the evaluation of teachers is done by pretest and posttest measures of achievement. This is particularly true if students were grouped using the pretest as a criterion. The extremely low scoring pupils will have a natural tendency, due to the regression effect, to improve their score. If a teacher were to be assigned many of these extremely low scoring pupils, the researcher would likely find at the end of the experimental period a very satisfactory improvement. However, if he or she were to be assigned the very high scoring pupils, just the opposite is likely to occur. The effects of student assignment could exert substantial bias due to the regression effect on any evaluation using pretest-posttest gains. Medley, Coker and Soar (1984) note that all

commonly used methods of estimating mean gains in achievement are susceptible to the regression effect unless pupils have been randomly assigned. They propose a solution to this dilemma which appears free of this bias. This process of estimating student gains from within-class regression is the selected treatment for the criterion variables.

Estimating Within-Class Achievement Gains

The method of estimating within-class student gains that was employed in this study is a statistical treatment of pretest and posttest data. First, for each teacher in the sample a regression equation is calculated using the pretest measure as an independent variable. The mean score on the pretest (generally a system-wide mean for that grade) is then used to obtain the predicted score for the dependent variable if this "average" student were to be in this teacher's class. This predicted score for the "average" student assigned to each teacher's class becomes the validating criterion. In correlating the criterion with the TPAI item scores, a partial correlation coefficient is calculated. Medley, Coker and Soar (1984) explain:

> It should be noted that differences between classes in Y' [the predicted score derived from the mean pretest] (like those in any other measures of pupil gains) cannot be attributed

> solely to differences in teacher performance unless pupils have been randomly assigned to classes (within grade and subject). If pupils are not randomly assigned, some portion of these differences may be due to differences in the classes rather than to differences in teachers. When you correlate Y' with scores on the measure of performance you are trying to validate, you will need to calculate partial correlations, holding major contextual factors (such as the average ability of the class) constant (p. 244).

This method was employed in this study in calculating the validity coefficients. The major contextual factors held constant are the within-class means of student variables that are not within the control of the teacher such as grade level, mean IQ and mean of the class absences.

Sample Selection

The study sample was composed of teachers volunteering for the study in five elementary schools in a central North Carolina school district. The classroom teachers had assignments in grades one through six and exceptional education teachers have assignments in learning disabilities, mental retardation, behavioral/emotional handicapped, and gifted and talented. Only classroom teachers in grades two through six were used in obtaining criterion validity coefficients and in calculating correlations with student variables. The total sample of teachers was used in obtaining correlation coefficients between teacher variables and TPAI scores. Each teacher had completed 30 hours of NCSDPI's Effective Teacher

Training and therefore should have a basic understanding of the TPAI and effective teaching research.

Each teacher was evaluated by the principal of the school where the teacher is assigned. Each principal had completed 30 hours of NCSDPI's Effective Teacher Training, 24 hours of NCSDPI's Teacher Performance Appraisal Training, and 30 hours of NCSDPI's Mentor/Support Team Training. Furthermore, each principal had a minimum of three years experience on the job and used the TPAI prior to the study.

Description of the Setting

The system selected to be the research site is typical of many in this region of North Carolina. Total student enrollment hovers around 4000 with the total population inside the attendance district of about 40,000. Teachers receive a small yearly supplement and total per pupil expenditures are about average for the state. School facilities are adequate with no unsuitable or undesirable physical plants although two schools are old and scheduled for replacement. All schools operate under a freedom-of-choice pupil assignment plan although state-supplied bus transportation is provided via the use of attendance zones. The district has no apparent major problems and student ability and achievement appear to be spread equally among schools.

The school district had 100% participation in Effective Teacher Training system-wide and had implemented the TPAI evaluation scheme by obtaining all necessary training from the NCSDPI. As an added measure of support to elementary school principals and teachers, the school district employed two full time teacher evaluators who received all required NCSDPI training. These evaluators served in an advisory role and assisted the principal in data collection. They did not participate in the actual TPAI summative evaluation.

The five elementary schools in the study implemented the statewide testing program in grades three and six. From local resources the system also tested children in grades one, two, four and five. Each child in the system is administered an appropriate form of the California Achievement Test each year. This provided the researcher with an appropriate measure of previous learning for every child.

Research Procedures

The superintendent and school principals were approached, their participation and cooperation requested and the study explained. All offered their support to the effort. They were shown all instruments and explained the safeguards designed to assure data security and protection of anonymity. The researcher visited all five schools and

presented the intent and method of the study to the teachers in school-based teachers' meetings hosted by the principal. In each meeting the researcher presented in oral and written form the purpose and method of study (Appendix A). The researcher explained the safeguards designed to protect the privacy of teacher participants. Each teacher was shown all data collection instruments (Appendix A). The "Letter of Informed Consent" (Appendix A) was read and explained and each teacher given opportunities to ask questions. Teachers were then solicited for participation. A follow-up letter was sent to acknowledge the consent for participation of each volunteer. At the same time, a reinvitation to participate (Appendix A) was sent to each teacher who did not volunteer to participate in the research.

TPAI Administration

All personnel evaluations are governed by state and local school board policy. The local policy governing teacher personnel records in the sample district are quite strict. After the TPAI (or any teacher evaluation) has been completed and signed by the participating parties, it is filed in the teacher's confidential personnel file and is only available to the superintendent and the teacher.

The TPAI evaluations in this study were duplicate administrations of the teacher ratings and not the official

TPAI signed by the teacher. The official copy of the instrument is filed in the teacher's personnel file. A combination of state law and local school board policy prevent access to personnel files except to the teacher and the superintendent. This duplicate administration was done with the volunteer teacher's knowledge.

All TPAI instruments and questionnaires were collected in sealed envelopes and later coded by an unaffiliated party with no knowledge of the school district or volunteer teachers. After coding, all names were removed from all data sheets. This was considered to be the optimal protection of privacy.

Data Collection

The teacher data questionnaire (Appendix A) collected information on teacher age, race, sex, years experience, etc, and was coded by the procedure described above. The student data collection instrument (Appendix A) was completed by the teacher. A direction sheet was attached to the actual data instrument which included an illustrated example of the CAT data as it would be found in the student's cumulative folder. A sample of four teachers were asked to complete the student data sheet prior to the study. All four teachers completed the data sheet without additional instruction and without error.

Missing data on the "Student Data Collection" sheet was reviewed by the researcher. When attempts to find the data failed, the data were coded as missing data. All statistical techniques with bivariate data employed pair-wise deletion of missing data.

Students were selected by a stratified random sampling strategy. Ten students were selected from each class at random using a randomized number generator and a master list of students assigned to each classroom provided by the principal. This was done to reduce the time required by the teacher to approximately 30 minutes. This method yielded data on approximately 45% of the students in the volunteer classrooms.

Data Analysis

The purpose of the study was to assess the criterion validity evidence for intended inferences made from the Teacher Performance Appraisal instrument developed by the North Carolina Department of Public Instruction (NCSDPI). Sources of bias from teacher or student characteristics were also tested. Data were assembled and analyzed utilizing the Stats+ statistical system supplemented by the Advanced Regression Methods package from CSS, both developed by StatsSoft of Tulsa, Oklahoma. This package exceeds all typical benchmarks for precision including the Longley tests for precision in multiple regression which

require the use of double precision calculations and algorithms designed to minimized rounding errors.

The data analysis was conducted in the following manner. First, Exploratory Data Analysis (EDA) was conducted and the means, standard deviations, skewness, kurtosis and the valid number of cases for each variable measured in interval or ratio level of measurement was calculated and reported. Student variables were examined first. Achievement data were analyzed by grade level. Next teacher variables were analyzed and finally the TPAI function scores. The proportions of dichotomously measured data (e.g. sex and race) were reported for both Student variables and teacher variables.

The within-class means of student CAT scores, student IQ, parents in home and days absent was calculated. The proportions of dichotomously measured variables (race, sex and free lunch) were also calculated. These data were used to create a new data matrix where each case was a class mean (or proportion) for the students assigned to a classroom teacher in the sample.

A regression equation was constructed for each teacher with an assignment in grades two through six using the 1987 CAT total (CAT87T) as the independent variable and the 1988 CAT total (CAT88T) as the dependent variable. Separate equations were built for CAT math (CAT88M) and CAT reading (CAT88R). The predicted score for the average student in

that teacher's class (CATt', CATm' and CATr') was computed using the appropriate system-wide mean for the grade level assigned to the teacher. These three vectors of predicted scores (CATt', CATm', and CATr') were the validating criteria and were appended to the new data matrix created above.

The first step in establishing the respective validity coefficients was the construction of a correlation matrix composed student variables as row variables and CATt', CATm' and CATr' vectors as column variables. This matrix revealed the relationships between class composition variables and estimated achievement. The next step was the construction of a stepwise multiple regression equation using student variables to predict the CATt', CATm' and CATr'. The variables retained in the equation were identified as the influential (contextual) variables and were held constant in calculating the validity coefficients. It is interesting to note that it is possible for the variables included in the resulting three equations predicting CATt', CATm' and CATr' to differ.

The student variables retained by the stepwise multiple regression equations were the control variables when the validity coefficients between TPAI function scores and the estimated student achievement for the average student in each class was calculated. The validity coefficients were calculated by forcing the retained

student variables into a multiple regression equation and leaving out the TPAI function scores. The multiple regression equation was reported as well as the partial correlation coefficients calculated for each TPAI function score as if it were to be entered next into the equation. Statistics reported for each TPAI function score were the partial correlation coefficient, the t-value associated with each statistic, the statistical significance of the t-value, and the beta in (standard regression weight for the respective variable if it were entered into the regression equation as an independent variable). The t-value of each function score and the total composite score was the statistic used to test hypotheses one through six.

To test for bias, two correlation matrices were constructed. The first was composed of a comparison between TPAI function scores and the within-class means of student variables. This revealed any bias the TPAI may have by indicating any relationship between class composition and TPAI scores. If the test developer's claim that the TPAI is equally appropriate for all teacher assignments is true, no correlation between TPAI function scores and the various student variables represented by class means would be indicated. If a significant part of TPAI variance is predictable from a student variable, that claim is untenable and a bias is indicated.

The second matrix was composed of a comparison between TPAI function scores and teacher variables. No correlation between TPAI and a teacher variable should be found unless that variable is also correlated with estimated student gains. All correlation matrices indicate the coefficient and the p-value of the correlation. The p-values from the first matrix were used to test hypothesis seven. The p-values from the second matrix were used to test hypothesis eight.

Summary

The method of study in this research project was a non-obtrusive data collection program followed by a statistical analysis designed to ascertain validity coefficients between teacher TPAI function scores and student achievement. The validating criterion selected for this study was the California Achievement Test (CAT) as implemented by North Carolina's statewide testing program and as augmented by local education agency policy. The CAT was selected as the validating criterion measure in this study for three reasons. (1) The CAT is the primary instrument used in the state mandated testing program and is widely considered a barometer of educational well being. (2) Its use is integral in measuring the status of student achievement in basic skills as part of a general assessment of school effectiveness conducted throughout the state.

(3) North Carolina also uses the CAT as preliminary evidence for mandatory summer school.

The study sample was taken from elementary school teachers in a central North Carolina school system. The teacher sample was comprised of teachers in five elementary schools who volunteered to participate in the study. All teachers and principal/evaluators in the sample have completed the state required and recommended instruction for use of the TPAI. All tenured teachers in the sample have been evaluated previously using the TPAI.

The statistical treatment for the validating criterion was a within-class regression technique developed by Medley, Coker, and Soar (1984) which provided an estimate of achievement of the average student for the each teacher in the sample. The study also proposed to evaluate possible bias of TPAI ratings through a correlation study of selected student and teacher variables.

Chapter four reports the results of the study. The sample obtained is discussed. The student variables and their relationship to achievement are examined and the results of an analysis of teacher variables is presented. The TPAI scores are analyzed and basic statistics are presented for each function. Intercorrelations between functions are reported.

Finally, TPAI functions are correlated with estimated within-class scores of an average student utilizing the

equation built for each teacher for the CAT total, CAT math and CAT reading. The validity coefficients are reported. Hypotheses one through six are tested by the coefficients. Correlation between student variables and TPAI scores are calculated and hypothesis seven is tested. Correlation between teacher variables and TPAI scores are calculated and hypothesis eight is tested.

CHAPTER IV

REPORT OF FINDINGS

Data were collected on 53 teachers in the five elementary schools which composed the elementary education program for a central North Carolina school district. Teachers in grades two through six were asked to supply data on ten randomly selected students. Data were collected on 400 students in 40 classrooms. Six teachers that volunteered to participate in the study had assignments in exceptional children programs and seven teachers had assignments in grade one and student data were not collected for these 13 teachers.

Only teachers in grades two through six were used to establish the criterion validity coefficients. For first grade teachers this omission was due to a lack of pretest criterion measures. For teachers of exceptional children the omission was done to avoid the possibility of a student being used twice in the sample, once for the classroom teacher and once for the exceptional children teacher. Also, serious questions of validity could arise in using the selected criterion to measure teaching outcomes intended for many exceptional students.

Teachers in grade one and exceptional children's classes were added to the sample only for tests of possible

bias due to personal traits of teachers such as age, sex, and years in school. The principal from each of the five elementary schools participated by completing a TPAI for each teacher that volunteered for the study.

Volunteering teachers represented 61 per cent of the teaching staff eligible for participation in the study. The participation rate by teacher classification was uniform with 62 per cent of teachers in grades two through six, 58 per cent of teachers in grade one and 60 per cent of teachers with exceptional children's assignments participating in the study. One first grade teacher was not included as a volunteer because her letter of informed consent was received after data collection had begun. All teacher volunteers completed all data collection activities. Missing data on teachers did not exceed two per cent on any variable.

Description of Student Sample

Data were collected on a stratified random sample of 400 students from classrooms of volunteering teachers with assignments in grades two through six. Missing data did not exceed two percent of the cases on any variable except IQ where missing data accounted for 8.5 per cent of the cases. Attempts to locate this data revealed that the bulk of the missing IQ scores were due to students transferring into the school system without having received IQ testing.

In the system under study, IQ testing is done early in the school year for all second and fifth graders. Children transferring after December of their second grade year would not receive testing until December of their fifth grade year. Transferring students generally had complete CAT scores due to the statewide testing requirement and the local system's policy of annual CAT administration.

Data for race, sex, days absent, and number of parents in the home were reported by total sample and by grade level. The sample was distributed almost equally among boys and girls although there was some variation in sex distribution by grade level (see Table 1). The racial

Table 1

Distribution of Student Race, Sex and Absences by Grade

Grade	N	% Total	Race		Sex		Mean
			%White	%Minority	% M	% F	Absences*
2	78	20	60	40	43	57	6.21
3	83	21	72	28	57	43	5.29
4	109	27	72	28	48	52	3.96
5	66	17	73	27	56	44	5.32
6	64	16	70	30	51	49	6.00
Total	400		70	30	51	49	5.23

Note: Percentages may not total 100 due to rounding
*Absences during the first six months of school

distribution was 70 per cent white and 30 per cent minority. Blacks composed 99.5 per cent of the minority

population with only one Asian child being included in the minority classification. The average days absent for children in the sample was 5.23 days. Some differences were evident among classes with grade level averages ranging from 3.96 days absent to 6.21 days absent for the first six months of school.

A student's family structure was thought to be a possible contributor to school achievement and data were collected on the number of parents living in the home. For the purpose of data collection, "parent" was defined as a natural parent or a step parent. A grandparent, aunt or uncle was classified as a guardian other than a parent. If the student was in the custody of a guardian, the numerical coding of zero was given in computing the average number of parents in the home. The result of this analysis revealed an average of 1.61 parents living in the homes of sampled students. The analysis showed that 65.6 per cent of students lived with two parents, 30.3 per cent lived with one parent, and 4.3 percent lived with a guardian other than a parent or step parent (see Table 2). There were two marked deviations from the total average. Grade four has a larger percentage of children living in homes with guardians other than parents than other groups. This group also had more children living in homes with two parents than other groups generally. Grade six had no children reported as living with a guardian other than a parent.

Table 2

Distribution of Student's Family Structure:

Parents per Household by Grade

Grade	N	Mean of Parents in Home*	% Two Parents in Home	% One Parent in Home	% Other Guardian in Home
2	78	1.59	61.5	35.9	2.6
3	83	1.60	61.4	37.4	1.2
4	109	1.60	70.6	18.4	11.0
5	66	1.67	69.5	27.3	3.0
6	64	1.63	62.5	37.5	0.0
Total	400	1.61	65.5	30.25	4.25

Note: Percentages may not add to 100 due to rounding
*Guardian other than parent was coded as zero

A viable measure of the economic condition of the student's family was found in the student's participation in the federal lunch program (see Table 3).

Table 3

Student's Family Economic Status Indicator:

Participation in Federal Lunch Program

	Percentage Within Grade					
Status	2	3	4	5	6	Average Across Grades
Free or Reduced	51	42	29	36	36	39
Paid or Brought	49	58	71	64	64	61

For statistical purposes student status in the lunch program was dichotomously coded as a one for participation and a zero for nonparticipation. Overall 39 percent of the students in the sample received either a free or reduced price lunch. Children in lower grade levels participated in the federal lunch program at a higher percentage than older children.

Table 4 displays the mean IQ, pretest (1987) CAT and posttest (1988) CAT by subtest and total scale score. The mean IQ by grade ranged from a high of 100.03 in grade five to a low of 98.27 in grade six. The CAT pretest means

Table 4

Student IQ, Pretest and Posttest Means by Grade

			Pretest 1987 CAT			Posttest 1988 CAT		
Grade	N	IQ	R	M	T	R	M	T
2	78	97.93	523	551	537	617	639	635
3	83	99.93	615	652	637	660	682	672
4	109	99.77	669	690	681	684	709	693
5	66	100.03	694	717	701	710	739	720
6	64	98.27	712	729	718	731	746	733
T	400	99.22						

Note: R = Reading M = Math T = Total Battery

presented in Table 4 were used as estimates of system-wide averages of achievement for each grade level.

Further analysis of achievement data was conducted and the standard error of the mean, the standard deviation, skewness and kurtosis of each subtest and total scale score by grade computed (see Table 5). The standard error of the

Table 5

Pretest CAT Total Scores: Basic Statistics by Grade

Grade	N	Mean	Standard Error	Standard Deviation	Skewness	Kurtosis
2	74	537	7.49	63.95	.0808	-.6237
3	79	638	5.95	52.55	-.3976	-.4771
4	107	681	4.16	42.79	-.1430	-.2032
5	64	702	4.81	38.20	-.1954	.3932
6	64	718	4.35	35.60	-1.3804	4.6921

Posttest CAT Total Scores: Basic Statistics by Grade

Grade	N	Mean	Standard Error	Standard Deviation	Skewness	Kurtosis
2	77	635	5.67	49.44	-.1750	-.4454
3	83	673	5.34	48.42	-.2048	-.2694
4	108	694	4.02	41.58	-.2845	-.2319
5	66	720	4.25	34.30	-.7740	1.5683
6	64	733	4.06	32.26	-.3033	.2054

mean is used to determine the confidence interval for the estimated mean of the population. For example, doubling the standard error of the mean and adding this sum to the

estimated mean would yield the upper bound of a .95 confidence interval for the mean estimate. Subtracting this same figure would yield the lower bound. The standard deviation is an indicator of the variability found in the sample.

Skewness is an indicator of the degree of asymmetry of a distribution. In a normal distribution the mean and median are expected to be the same point (the middle) along a scale representing the scores on the measure. If a distribution is positively skewed, the mean would be located at a higher value than the median. In appearance, the left tail would appear shorter than the right tail on a graphical representation of the distribution. If the skewness is negative, the mean would be located at a lower value than the median and the right tail will appear shorter. The size of the statistic indicates the degree of deviation from the expected distribution if a normal distribution were to be assumed.

Kurtosis is a somewhat similar concept but indicates how peaked or flat a distribution is when graphed. A positive value for this statistic indicates that the graphed distribution is more peaked (leptokurtic) and has thinner tails than a normal distribution would be expected to have. A negative value for this statistic indicates that the graphed curve will be flatter than expected (platykurtic) and have thicker tails. The magnitude of the

statistic indicates the degree of deviation if a normal distribution were to be assumed.

In the CAT total scores (see Table 5) only one statistic, the pretest for the sixth grade, showed marked skewness. This measure also showed sharp positive kurtosis. Total scores by grade, however, were generally found to be mildly negatively skewed with mild kurtosis.

An analysis of CAT reading subtest scores revealed a similar mild negative skewness and a mild positive kurtosis (see Table 6). The standard deviation of pretest and

Table 6

Pretest CAT Reading Scores: Basic Statistics by Grade

Grade	N	Mean	Standard Error	Standard Deviation	Skewness	Kurtosis
2	74	524	8.72	74.46	.0851	-.6095
3	79	615	7.35	64.91	-.6111	.1244
4	107	669	5.48	56.42	-.8083	1.4653
5	64	695	6.38	50.62	.3497	2.1518
6	64	713	5.03	39.91	-.3177	-.0997

Posttest CAT Reading Scores: Basic Statistics by Grade

Grade	N	Mean	Standard Error	Standard Deviation	Skewness	Kurtosis
2	77	617	7.72	67.77	-.7038	.3871
3	83	660	7.22	65.40	-.4636	.5409
4	108	684	5.08	52.56	-.4851	.5725
5	66	711	4.80	38.73	-.5881	1.0229
6	64	731	4.58	36.32	-.2278	1.0727

posttest reading subtests decreased as grade level increased as did the CAT total score standard deviation. Math scores also displayed a comparable pattern of negative skewness and mild kurtosis (see Table 7). Again, the

Table 7

Pretest CAT Math Scores: Basic Statistics by Grade

Grade	N	Mean	Standard Error	Standard Deviation	Skewness	Kurtosis
2	74	551	7.90	67.56	-.1344	-.0238
3	79	652	6.46	57.08	-.1056	-.7426
4	107	690	4.06	41.81	-.1776	.1000
5	64	717	4.42	35.05	-.5627	1.2277
6	64	729	4.65	36.87	-1.9304	7.5657

Posttest CAT Math Scores: Basic Statistics by Grade

Grade	N	Mean	Standard Error	Standard Deviation	Skewness	Kurtosis
2	77	639	6.78	59.09	.1895	-.8175
3	83	682	5.43	49.17	-.1879	-.0568
4	108	709	3.82	39.49	-.1426	-.2032
5	66	739	3.98	32.11	-.1954	.3932
6	64	733	4.06	32.26	-.3033	.2054

standard deviation decreased as the grade level increased. Summarizing the distribution of the entire set of scores it can be concluded that, with the exception of the pretest math and pretest total scores for the sixth grade, the skewness and kurtosis is mild and not suggestive of any

serious deviation from what might be expected for a data set while assuming normality. The narrowing of the standard deviation with increasing grade level in both pretest and posttest scores is an expected function of both student maturation and an increase in the number of items in the tests.

Description of Teacher Sample

The total teacher sample used to test for bias in TPAI scores attributable to teacher traits was composed of 53 teachers in grades one through six and exceptional children's assignments. The teacher sample used to establish the validity coefficients was composed of 40 classroom teachers in grades two through six. The teachers in this group were veteran teachers having an average experience of 17.70 years teaching service (see Table 8). The average number of years teaching within the school system in the study was 13.23 thus reflecting a mature stable teaching staff with an average age of 42.64 years. The sample of teachers in grades two through four was composed of 92.5 per cent female and 7.5 per cent male teachers and had a racial distribution of 85 per cent white and 15 per cent minority.

Of the total teaching sample, the highest degree held by 68 per cent of the teachers was a bachelor of arts degree whereas 32 per cent of the teachers held a masters

degree. Teachers were almost equally divided between public and private institutions when earning their initial degree with 55 per cent attending public and 45 per cent attending private institutions. The highest degree offered by these institutions was about equally distributed among bachelors, masters and doctorates with 37, 33, and 28 per cent of the teachers attending these institutions respectively.

Table 8

Descriptive Statistics of Teachers in Grade 2-6

Teacher Trait	N = 40 Minimum	Maximum	Mean	Standard Deviation
Years Experience	1	32	17.70	7.70
Years in Current School	1	25	13.23	7.08
Years Teaching Current Grade	0	24	10.63	7.19
Teaching Starting Age	21	38	23.75	4.42
Age	22	58	42.64	9.23

Teacher Variables and Estimated Student Achievement

Teacher variables that have been traditionally thought to influence student learning were correlated with the

estimated student gains for an average student in each teacher's class on the CAT total, math and reading tests (see Table 9). These variables included years experience,

Table 9

Partial Correlations Between Achievement and Teacher Variables Holding Grade Level Constant

Teacher Variable N = 40	Predicted CAT Score for Average Student		
	Total	Math	Reading
Years Experience	-.1060	.0121	.1058
Years in School	-.0073	.2235	.1222
Years in Grade	-.1265	-.0251	.0026
Age Began Teaching	-.1473	.0603	-.0688
Age	-.1689	.0691	.0204
Sex[1]	.0814	-.0446	.0048
Race[2]	-.0356	-.1433	.1975
Highest Degree Held	-.0191	-.0318	.0232
Type Institution[3]	.0016	-.0201	.0466
Institution Level[4]	-.2393	-.1798	-.1930

Note: No correlation was significant at $p < .05$
[1]Sex was coded Male = 0 Female = 1
[2]Race was code White = 0 Minority = 1
[3]Type Institution was coded Public = 0 Private = 1
[4]Institution Level was highest degree offered
BA = 4 MA = 5 PhD = 7

highest academic degree held and age. Because grade level is known to be a heavy influence on scale score achievement, it was held constant across teachers and partial correlations were computed. Also correlated with measures of achievement were teacher variables that were possibly a source of bias such as race and sex. No significant correlations were found.

Correlation Between Pretest and Posttest Achievement Scores

Table 10 displays the correlations between administrations of the CAT and CAT subtests. The correlation between the CAT pretest (CAT87Total) and the CAT posttest (CAT88Total) was .8813 accounting for

Table 10

Correlations* Between CAT Pretest and CAT Posttest

N=387	Posttest		
Pretest	CAT88Total	CAT88Math	CAT88Reading
CAT87Total	.8813	.8439	.8447
CAT87Math	.8256	.8455	.7557
CAT87Reading	.8539	.7820	.8562

*All correlations are significant at the $p < .001$ level

78 per cent of the variance between the students' pretest and posttest total scores. The correlation between administrations of the CAT ranged from .7820 to .8813. Part of this strong relationship can be explained by the extreme similarity in forms of the test. Influential variables which affect student achievement also had an opportunity to play their role in affecting both scores.

Student Variables and Achievement

Student variables that were thought to be influential were correlated with the CAT posttest results. All variables examined were significant predictors of a student's total CAT score. Table 11 reports the the calculated coefficients. Student race, lunch program participation, number of absences during the first six months of school and IQ were all significant predictors of the CAT total, math and reading results. Sex and number of parents at home were significant predictors of CAT total and reading scores but not of math scores. Further analysis showed a high degree of intercorrelation among these variables. If variables were combined to predict test scores, as is done in multiple regression, not all variables would make a significant contribution because they duplicate the contribution made by another variable. To eliminate these variables and to obtain the model with the best fit, stepwise multiple regression was

Table 11

Partial Correlations[1] Between Achievement and Student Variables N = 400

Variable	CAT Total	CAT Math	CAT Reading
Race	.2094***	.2062***	.1761**
Sex[2]	-.1357**	-.0409	-.1392**
Lunch Program[3]	-.2743***	-.2428***	-.2524***
Parents in Home	.1089*	.0743	.1180*
Absences	-.2424***	-.2146***	-.2109***
IQ	.6056***	.5421***	.5460***

[1]The effects of grade level have been held constant
[2]Sex was coded Male = 1 Female = 0
[3]Lunch Program was coded Free or Reduced = 1
Not Participating = 0

*** p < .001 ** p < .01 * p < .05

performed using CAT total, math and reading scores as dependent variables. The variables remaining in the equation were the variables used to equate classrooms. This was done by partialing out the effects of these variables prior to establishing the validity coefficients.

Establishing Control Variables

Three statistical models were constructed to predict CAT total, math and reading scores. Table 12 reports the results of stepwise multiple regression using CAT total posttest as the dependent variable with forward entry (F in = 3.57; F out = 2.50). A student's grade level, IQ, number of absences and sex were selected as predictors for the

model and had an adjusted R^2 of .6152 thus accounting for 62 per cent of the variance in students' CAT total scores.

Table 12

Influential Variables in the Predication of CAT Total Posttest

Forward stepwise regression, number of steps: 4

Dependent variable: CAT88Total
Multiple R: .7871
Multiple R^2: .6195
Adjusted R^2: .6152
Minimum pairwise N: 360
F (4, 355) = 144.5071 p < .0000

Intercept: a = 434.3452

REGRESSION WEIGHTS

Variable	BETA	b	standard error of b	t (355)	significance of t
GRADE	.6052	24.4166	1.3214	18.4765	.0000
IQ	.4631	1.7065	.1218	14.0100	.0000
ABSENCES	-.1031	-.8378	.2680	-3.1150	.0024
SEX*	-.1002	-10.7858	3.5280	-3.0572	.0028

*SEX: male = 1 female = 0

These four variables were held constant when correlating TPAI scores to CAT estimated total score gains. Omitted from the equation were measures of race, lunch program participation and number of parents in the home.

A similar equation was built for math (see Table 13). Three of the four predictors for CAT total were retained. Student sex was omitted from the equation. The adjusted R^2

Table 13

Influential Variables in the Prediction of CAT Math Posttest

Forward stepwise regression, number of steps: 3

Dependent variable: CAT88Math
Multiple R: .7631
Multiple R^2: .5823
Adjusted R^2: .5788
Minimum pairwise N: 360
F (3, 356) = 165.4254 p < .0000

Intercept: a = 441.3879

REGRESSION WEIGHTS

Variable	BETA	b	st. err b	t (355)	significance of t
GRADE	.6270	27.0835	1.4800	18.3005	.0000
IQ	.4052	1.5988	.1364	11.7148	.0000
ABSENCES	-.0895	-.7795	.30115	-2.5883	.0098

for this equation was .5788 thus accounting for 58 per cent of the variance in student's CAT math subtest.

The model for the prediction of the CAT reading subtest accounted for 53 per cent of the variance in CAT reading scores and had an adjusted R^2 of .5335. Table 14 gives the results of this model. The same influential variables found in the prediction of the CAT total score were chosen for the prediction of CAT reading. In both the reading subtest and total composite, being male had a negative correlation with an increase in the predicted test

Table 14

Influential Variables in the Prediction of

CAT Reading Posttest

Forward stepwise regression, number of steps: 4

Dependent variable: CAT88Reading
Multiple R: .7340
Multiple R^2: .5387
Adjusted R^2: .5335
Minimum pairwise N: 360
F (4, 355) = 103.6581 p < .0000

Intercept: a = 385.7096

REGRESSION WEIGHTS

Variable	BETA	b	st. err b	t (355)	significance of t
GRADE	.5599	27.8505	1.7945	15.5204	.0000
IQ	.4358	1.9804	.1654	11.9733	.0000
SEX*	-.1084	-14.3933	4.7907	-3.0044	.0032
ABSENCES	-.0933	-.9355	.3652	-2.5615	.0105

*SEX: male = 1 female = 0

score. The strongest influence of the variable student sex was found in the prediction of reading subtest scores, where being male was equivalent to subtracting 14 points from the predicted female score. In all three tests, an increase in absences indicated a decrease in the predicted test score, generally about one point in the predicted score for each day absent.

Predicted scores for each student were computed using each of the three equations and scatterplots were constructed using the predicted score and the actual CAT

score (see Appendix B). The standardized residuals were plotted by predicted score (see Appendix B). While there was a slightly greater spread in the standardized residual for lower values of the predicted CAT score for total, math and reading, the difference was not judged to be great enough to assert a violation in the assumption of homoscedasticity (equal variance across all values of the dependent variable). The spread, in fact, was not as great as expected since it was noted in the exploratory data analysis that lower grades had greater standard deviations than upper grades. Normality and linearity assumptions were also held to be valid. The results of the regression analysis were considered valid and acceptable for use as indicators of influential variables to be controlled in statistically equated classes.

Results of TPAI Evaluations

Scores for each of the 53 teachers in the sample were compiled and a mean score for each function was calculated (see Table 15). The TPAI function means ranged from a high of 4.66 on Student Behavior and Instructional Presentation to a low of 4.33 on Facilitating Instruction on a one to six scale. The largest difference between means on any two items was very small at .33.

A frequency count was conducted for each item by rating category. Table 15 presents the rating distribution

Table 15

TPAI Item Analysis

TPAI Function	Rating (Per Cent Scoring)						N = 53	
	1	2	3	4	5	6	Mean	Standard Deviation
Instructional Time	0	2	11	36	42	9	4.45	.88
Student Behavior	0	0	23	17	32	28	4.66	1.11
Instructional Presentation	0	0	11	30	40	19	4.66	.91
Instructional Monitoring	0	0	17	30	38	15	4.51	.94
Instructional Feedback	0	0	21	34	36	9	4.34	.91
Facilitating Instruction	0	2	17	30	42	7	4.33	.90
Communicating Within the Educational Environment	0	0	15	26	43	13	4.56	.91
Non-Instructional Duties	0	2	10	33	35	20	4.63	.98

Note: Percentages may not add to 100 due to rounding

Scoring Key:

1 = Unsatisfactory
2 = Below Standard
3 = Satisfactory
4 = Above Standard
5 = Well Above Standard
6 = Superior

by percentage of teachers receiving each rating for individual functions. No teacher received an

unsatisfactory rating on any item. Only three items, Instructional Time, Facilitating Instruction and Non-Instructional duties, had teachers included in the below standard category. In each case, only two per cent of the sample were rated in the below standard category. The largest distribution for a satisfactory rating was found in the Student Behavior function with 23 per cent of the teachers being rated in that category.

A total score was computed for each of the 53 teachers in the sample by summing the rating of all eight functions for each teacher. A correlation matrix was created (see Table 16) comparing all functions and the total score. Correlations between all functions were significant beyond the $p < .001$ level and ranged from a high of .83 to a low of .56. Correlations between function scores and total score were consistently high ranging from .92 to .81 representing total score correlation with Instructional Monitoring and Student Behavior respectively.

Evaluator's Questionnaire

Evaluators were asked to complete a short questionnaire concerning their reactions to the results of the TPAI evaluation completed on each teacher. All principals strongly agreed or agreed when asked if the results of the TPAI estimate accurately reflected the official TPAI for this teacher. No principals were

Table 16

Correlations Between TPAI Function Scores

Function	Function 1	2	3	4	5	6	7	8
					N = 53			
1. Instructional Time								
2. Student Behavior	.71							
3. Instructional Presentation	.78	.61						
4. Instructional Monitoring	.83	.70	.79					
5. Instructional Feedback	.77	.71	.75	.83				
6. Facilitating Instruction	.77	.56	.80	.77	.75			
7. Communicating within the Educational Environment	.60	.64	.63	.70	.71	.66		
8. Non-Instructional Duties	.65	.62	.70	.73	.65	.69	.81	
T. Total	.88	.81	.87	.92	.89	.86	.83	.85

Note: all correlations significant at $p < .001$ with a directional hypothesis

undecided or in a disagree category. When asked if the TPAI score for this teacher accurately assessed this teacher's total effectiveness, principals again overwhelmingly agreed with 49 per cent in the strongly

agree and 51 per cent in the agree category. When principals were asked if data were used other than that collected in formal observations a wide difference was found. Thirty per cent of the responses on teachers strongly agreed that only data from observations were used. Sixty-four per cent of the responses indicated strong disagreement with the statement indicating the use of data other than formal observations. All principals indicated they felt they were competent judges of the teachers under evaluation. Sixty-two per cent strongly agreed and 38 per cent agreed to the statement "I feel I am a competent judge of this teacher's effectiveness." The results were similar when principals were asked to respond to the situation in which tenure or pay scale decisions were made utilizing the TPAI just completed. Sixty-two per cent strongly agreed and 38 per cent agreed that a valid decision would be made if the TPAI just administered were used for pay or tenure decisions. The questionnaire also revealed that nine per cent of the teachers evaluated were probationary. Principals also indicated that they had conducted prior evaluations using the TPAI on 77 per cent of the teachers evaluated.

Validity Coefficients

Regression equations were built for each classroom teacher in grades two through six to predict the posttest

CAT score using the pretest CAT score as the independent variable. A predicted score for the average student assigned to each teacher was computed by inserting the estimated mean CAT pretest score for that grade level into the regression equation computed for each teacher. In this manner, an estimated student achievement gain was calculated for the CAT total, reading and math scores for each teacher.

These estimated scores were then correlated with the teacher's TPAI total and function scores holding the effects of the influential variables for each test constant. The results of these partial correlation coefficients are shown in Table 17. The coefficients for CAT total ranged from .05 to .39 with only the function assessing non-instructional duties having a statistically significant relationship with student CAT total achievement ($p < .05$ level). None of the coefficients between CAT reading and the TPAI were significant. There was a clear and consistent statistically significant relationship between predicted CAT achievement in math and each of the TPAI function scores. These coefficients ranged from .33 to .47 with the TPAI total composite score showing the strongest relationship with a .48 coefficient. Two additional composites were created. The empirically based functions (instructional time, student behavior, instructional presentation, instructional monitoring and

Table 17

Partial Correlations Between TPAI and Estimated Achievement

N = 40	Estimated Student Achievement		
Function	CAT Total[1]	CAT Math[2]	CAT Reading[1]
Instructional Time	.24	.43**	.18
Student Behavior	.25	.36*	.18
Instructional Presentation	.23	.41**	.10
Instructional Monitoring	.31	.41**	.19
Instructional Feedback	.19	.38*	.07
Facilitating Instruction	.05	.36*	.06
Communicating Within the Educational Environment	.20	.36*	.06
Non-Instructional Duties	.39*	.47**	.15
Total Composite	.28	.48**	.17

[1]Holding constant the effects of Grade, IQ, Sex and Student Absences
[2]Holding constant the effects of Grade, IQ and Student Absences

** p < .01 * p < .05 (Directional Hypothesis)

instructional feedback) were used to create a summed composite for each teacher. A similar composite was created for the consensus based functions (facilitating

instruction, communicating within the educational environment and non-instructional duties). The partial correlation coefficients for these two composites with predicted math achievement were .4502 and .4485 respectively.

Although the contribution of sex was not significant in the prediction of math achievement, controlling for this variable had the effect of increasing each of the validity coefficients between math and the TPAI functions by .03 to .06 with the largest increase raising the TPAI Total correlation with math from .47 to .53 thus attaining significance at the .001 level. The results of multiple regression can be found in Appendix B.

Correlations Between TPAI and Student Variables

Data on student variables were averaged by class and the means were correlated with teacher TPAI ratings. Data were collected on student race, sex, grade, federal lunch program participation, number of parents in the home, student absences during the first six months of school and IQ. No pattern of significant relationship was uncovered although significant correlations between a TPAI function and grade, student race, parents in home and student absences were found (see Table 18). Although only significant with Instructional Presentation rating, student

Table 18

Correlations Between TPAI Rating and Class Variables

Class Variable	Function 1	2	3	4	5	6	7	8	T
Grade	-.26	-.29	-.30	-.25	-.27	-.17	-.03	-.20	-.25
Race[1]	-.02	-.33	-.05	-.14	-.01	-.08	-.12	-.24	-.14
Sex[2]	-.15	-.16	-.08	-.20	-.22	-.21	-.26	-.23	-.20
Lunch	.22	.21	.06	.11	.23	.14	.00	.07	.13
Parents	-.36	-.28	-.34	.23	-.01	-.18	-.14	-.24	-.29
Absences	-.16	-.21	-.19	-.17	-.14	-.32	-.26	-.26	-.24
IQ	-.00	-.04	.06	.12	.08	0.7	.14	.11	.08

Note: Underlined coefficients are significant at $p < .05$ level using a non-directional hypothesis

[1]Race coded 0 = Minority 1 = White
[2]Sex coded 0 = Female 1 = Male

Function Key:

1 = Instructional Time
2 = Student Behavior
3 = Instructional Presentation
4 = Instructional Monitoring
5 = Instructional Feedback
6 = Facilitating Instruction
7 = Communicating Within the Educational Environment
8 = Non-Instructional Duties

grade had a consistently negative relationship with teacher rating. The proportion of white students in a class had a consistently negative relationship with TPAI rating although the magnitude of the statistic was significant only with ratings of Student Behavior. All TPAI functions had a negative relationship with the proportion of males in the class although no statistic was significant. The number of parents living with the student had a negative relationship with teacher TPAI rating and was significant with two TPAI functions Instructional Time and Instructional Presentation. Absences also had a consistently negative relationship with TPAI rating although only significant in one instance. The class mean IQ had a minor nonsignificant relationship with all TPAI functions.

Correlations Between TPAI Rating and Teacher Variables

Table 19 shows the comparisons of teacher variables that were thought to be predictors of teacher effectiveness or a source of bias in teacher evaluation instruments with each TPAI function. No significant relationships were found. Two of the variables tested for relationship with TPAI rating, Years Experience and Highest Degree held, are the teacher variables which are traditionally used to determine salary scales. For both variables, there was no significant relationship with principal rating on any of the TPAI functions.

Table 19

Correlations Between TPAI Rating and Teacher Variables

Teacher Variable	TPAI Function 1	2	3	4	5	6	7	8	T
Grade	-.13	-.16	-.19	-.10	-.10	-.18	-.14	-.16	-.19
Years Experience	.11	-.03	-.01	-.01	-.07	.13	-.10	-.08	.02
Years in School	.16	.05	.10	.10	.03	.17	-.01	.08	.12
Years in Grade	.16	.00	.02	.08	.00	.06	-.01	-.02	.07
Age Started Teaching	.09	.16	.11	.03	.01	-.03	-.02	.10	.07
Sex[1]	.14	.02	-.02	-.02	-.06	.11	-.10	-.05	.03
Race[2]	-.12	-.07	-.08	-.13	-.18	-.18	-.24	-.16	-.16
Highest Degree	.01	.14	.08	.10	.14	.06	.20	.09	.14
Institution Type	.04	-.01	-.01	-.07	.05	-.00	.02	-.07	-.03
Institution Level	-.01	-.13	-.01	.01	.04	-.00	.24	.05	-.01

Note: No correlation was significant at $p < .05$ level using a non-directional hypothesis

[1]Sex coded 0 = Male 1 = Female
[2]Race coded 0 = White 1 = Minority

Function Key:

1 = Instructional Time
2 = Student Behavior
3 = Instructional Presentation
4 = Instructional Monitoring
5 = Instructional Feedback
6 = Facilitating Instruction
7 = Communicating Within the Educational Environment
8 = Non-Instructional Duties

Summary

Chapter four described the student and teacher sample used in the study. Student variables were compared to student achievement and influential variables were discerned and later employed to statistically equate sample classrooms. Teacher variables were compared to student achievement and none were found to be significantly related in the prediction of student achievement. The results of TPAI evaluation were described and their relationship to student variables, teacher variables, and student achievement was examined.

Chapter five will test the eight hypotheses presented for study, summarize the study findings and make recommendations for TPAI use and further study based on the current study's findings.

CHAPTER V

SUMMARY, CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS

Summary

The study gathered data on 400 students in 40 classrooms and estimates of student achievement were made for an average student of that grade in each teacher's class. Holding constant the influence of student variables which affect achievement such as grade level and IQ, partial correlation coefficients were calculated measuring the relationship between TPAI ratings and student achievement. The TPAI ratings of these 40 teachers also were tested for correlation with student variables. Additional data were gathered on 13 teachers in grades one and exceptional children programs. Combined, these 53 teachers' scores were examined and tested for relationship with teacher variables that might indicate bias in the employment of the instrument. The purpose of this chapter is to examine these data in light of the hypotheses posed for this study and to make recommendations for use of the study's conclusions and suggestions for further study.

Conclusions

Validity coefficients were calculated by correlating the TPAI ratings of elementary teachers to CAT total, math and reading scores. To eliminate the effects of differences among classes in grade level, IQ, parents in the home and sex distribution, partial correlation coefficients were calculated and the effects of these variables were held constant statistically.

H_1: There is a significant positive relationship between TPAI total score and estimated student gains on the CAT total score using within-class regression.

The hypothesis that the total of TPAI function scores would have a positive correlation with the CAT total score was found not to be tenable and must be rejected. The correlation between the two variables was a modest .28 and was not significant at the $p < .05$ level with a directional hypothesis with 34 degrees of freedom.

H_2: There is a significant positive relationship between each TPAI function score and estimated student gains on the CAT total score using within-class regression.

The hypothesis that each TPAI function score would have a significant positive relationship with CAT total scores was found to be untenable for all but one function of the TPAI. Function eight, which relates to the performance of non-instructional duties, was found to have a significant positive relationship with CAT total scores. A correlation coefficient of .39 was reported for this

function which is significant at the $p < .05$ level with a directional hypothesis and 34 degrees of freedom. All other functions had correlations ranging from .05 to .31 and were found not to be significant.

H_3: There is a significant positive relationship between TPAI total score and estimated student gains on the Math subtest of the CAT using within-class regression.

The hypothesis that the TPAI composite score would have a significant positive relationship to CAT math achievement remains tenable. The correlation between TPAI total score and CAT math score was .48 and was significant at the $p < .01$ level with a directional hypothesis and 35 degrees of freedom. This was the largest relationship found between TPAI ratings and measures of achievement.

H_4: There is a significant positive relationship between each TPAI function score and estimated student gains on the Math subtest of the CAT using within-class regression.

The hypothesis that there would be a significant positive relationship between each TPAI function score and estimated student gains in math was found to be tenable for each TPAI function. All eight of the TPAI functions had a significant relationship with math achievement ($p < .05$ with 35 degrees of freedom and a directional hypothesis) with coefficients ranging from a low of .36 to a high of .47. Four of the functions, Instructional Time, Instructional Presentation, Instructional Monitoring and

Non-Instructional Duties, were significant at the $p < .01$ level.

H_5: There is a significant positive relationship between TPAI total score and estimated student gains on the Reading subtest of the CAT using within-class regression.

The hypothesis that there would be a significant positive relationship between TPAI total score and student achievement in reading was not found to be tenable and was rejected. The correlation between the two variables was .15 and was not significant at the $p < .05$ level using a directional hypothesis with 34 degrees of freedom.

H_6: There is a significant positive relationship between each TPAI function score and estimated student gains on the Reading subtest of the CAT using within-class regression.

The hypothesis that there would be a significant positive relationship between each TPAI function score and student achievement in reading was found not to be tenable and was rejected. The coefficients ranged from a low of .06 to a high of .18 and were not significant at the $p < .05$ level with a directional hypothesis and 34 degrees of freedom.

H_7: The zero order correlation coefficients between TPAI function scores and the class mean (or ratio) of the student variables of race, sex, IQ, past achievement, age, grade, economic status, attendance, or family structure are equal to zero.

The hypothesis that TPAI scores would have no statistically significant relationship with student variables reflecting the composition of a teacher's class

cannot be retained. Five of the 63 coefficients calculated to test this hypothesis were found to be statistically significant at the $p < .05$ level with a non-directional hypothesis and 38 degrees of freedom. The correlation between grade and TPAI function scores was consistently strong but only one reached the $p < .05$ significance level. Race was a significant predictor in only one function and showed a negative but insignificant relationship with all TPAI functions (race was coded White = 1, Minority = 0). Absences also showed consistent negative relationship. Only one coefficient, its intersection with Facilitating Instruction, was found significant. The average number of parents living with the child was significant in its relationship with two functions, Instructional Time and Instructional Presentation. There was no significant relationship between class variables and teacher TPAI rating using the total TPAI composite.

H_8: The zero order correlation coefficients between TPAI function scores and the teacher variables of age, sex, race, highest earned degree, grade taught, years teaching in school, years teaching in system, or total years of teaching experience are equal to zero.

The hypothesis that there is no statistically significant relationship between TPAI ratings and teacher variables was retained and remains tenable. Of the 90 correlation coefficients calculated to test this hypothesis, none were statistically significant at the

$p < .05$ level using a non-directional hypothesis with 51 degrees of freedom. Although grade level had a consistently negative relationship, none of the coefficients approached significance.

Implications

During the process of collecting and analyzing data to establish estimates of student achievement in a particular classroom, it became evident that student variables, not teacher variables, were the dominant predictors of school achievement. Undoubtedly, part of this situation originates from the extensive similarities among classrooms. Classes tend to have the same number of children for similar amounts of time. Textbooks and other materials used in classroom are almost identical among classes of the same grade. Teachers have met minimal standards and economic opportunity typically provides a ceiling on teaching ability. Beginning teachers are expected to teach the same mathematical algorithms, spelling words and the like as their more senior colleagues and, by this study's data analysis, do as well as those senior colleagues. The training received by teachers have been generally standardized by state certification boards and accrediting agencies. Most of school decisions are made by school principals and there are few opportunities for teachers to express differences in instructional

strategies that might make differences approaching the magnitude that can be measured by intellectual growth instruments now in use.

However, the differences among children's backgrounds are dramatic. Children come to school with striking differences in past experiences. Some five year olds come to school reading while others come not yet speaking in sentences. Yet, the children are expected to meet the requirements of a standardized curriculum. It is little wonder that knowledge of student variables can be strong predictors. The difficulty in measuring teachers, then, becomes not an issue of whether it can be done, but what are the expected outcomes of the process.

Since differences among teachers result in such small differences in student achievement, the significance of the activity of teacher evaluation in the enhancement of learning is questionable. A superior rating from a teacher in the sample provided no prediction as to how well that teacher's children might do in reading or total achievement. How a teacher performs her <u>non-instructional</u> duties as measured by the TPAI contributes all the predictive relationship attainable from a complete TPAI evaluation in the prediction of student's math achievement. The implication of the findings in this study suggests that standardized forms of evaluation will not be an avenue to widespread educational improvement because of the small

contribution differences among teachers make to a child's general education. This is not to say teachers do not make a significant contribution. It is unlikely students would be able to make the type of annual achievement gains indicated by the students in the sample unless they receive professional instruction from a teacher. What is poignant is that it appears that ranking teachers by differences on any measure, whether it is seniority, highest degree held or formal evaluation is unlikely to make a significant difference in the prediction of overall achievement of students.

Validity Coefficients

Hypotheses one through six tested the validity of the TPAI in predicting student achievement by CAT total, math and reading scores. Only hypotheses three and four were retained. A clear pattern of strong relationship between TPAI function scores and math achievement was established. Hypotheses six and seven, which tested the TPAI with reading achievement, was rejected and no other significant correlation was found in the coefficients. The TPAI's relationship to total achievement reflected was stronger than the TPAI relationship with reading probably because of the inclusion of the math influence. However, only one function was found to be significantly different from zero. This function dealt with the non-instructional duties of

teachers in carrying out school policies, adhering to school rules, and planning for professional development. It can be argued that the behavioral anchors of this function are among the vaguest and therefore subject to the most individual interpretation of any function. Nevertheless, it must be disconcerting that a function designed to measure non-instructional behavior proved to be one of the best predictors of the effectiveness of instruction as measured by student achievement.

The results of strong validity coefficients for math prediction but not for reading or total achievement reinforce the opinions of current reviews of effective teaching research. These reviews indicate that the methodologies identified by effective teaching research are significant predictors of achievement only in highly structured subjects such as math. In situations where the material to be learned is less structured, higher order thinking processes are required or definite easily defined goals are not appropriate, effective teaching methods are not likely to be predictors of achievement.

The implication of these research summaries, which was verified by the results of the current study, is that the use of instruments like the TPAI are not likely to be a valid predictor of academic achievement gains across all subjects. If the relationships found in this study were also found in studies of middle school teachers, for

example, the use of the TPAI in evaluating a math teacher would be valid to the degree that it offers some predictive validity in forecasting the likely outcomes of instruction. The use of the same instrument with a reading teacher would offer no knowledge of student achievement other than what might be accounted for if the evaluator were to roll a die to obtain the TPAI rating. The TPAI developers' claim that the instrument is equally valid across all teacher assignments simply cannot be held tenable in light of current evidence.

The differences in validity coefficients among total achievement, math and reading were not unexpected. The possibility that math would be more sensitive to measurement by the TPAI was evident in the literature reviewed by the study. The study design was therefore made sensitive to that possibility. Clearly, the evidence supports a contention that at least some types of learning outcomes can be predicted from evaluations of teaching behavior and that evaluators can be trained to utilize an instrument in an educational (as opposed to research) assignment. It appears that evaluators were measuring the behaviors they were trained to spot. These behaviors were identified primarily by correlations with structured subject matter. The TPAI results mirror what might be expected if it were possible to reliably identify these behaviors by a high inference, rather than a low inference

tool. The literature reports few coefficients between rating scales and structured learning outcomes above the .25 mark. The fact that all eight of the items and a total composite showed a significant relationship with math achievement well above the .25 mark was certainly an exciting finding. The strength of these coefficients must be acknowledged. Within the history of evaluative instruments based on rating scales these data are a unique and promising finding.

Inherent in the mixed validity of the TPAI is that the omnibus application of TPAI evaluations for teachers of all subjects and grades is not tenable. There are many learning outcomes desired that fit the model of a highly structured discipline such as decoding in reading and map skills in social studies. Evaluating teachers by the TPAI in their efforts to obtain these goals may have validity and certainly further studies to examine this issue would be worthwhile. However, most learning outcomes do not fit this model of a highly structured discipline.

The major goals outlined in North Carolina's standard course of study are of goals that defy rigid structuring. Using thinking skills to solve problems arising in the course of living and working require high levels of generalization of the skills taught in school. Solutions posed to those problems resist simple analysis and tightly organized structuring. Forcing teachers to conform to a

model of instruction advocated by the TPAI may help instruction in highly structured content areas, but may provide no assistance and could possibly even retard efforts to obtain more generalized goals. The use of the instrument must be tailored to the objectives desired from instruction. To require a newly certified elementary classroom teacher to demonstrate a standard performance of the TPAI may be justified by the study results in that these techniques were shown to be valid for math instruction which is a part of almost every elementary teacher's job. Justifying the use of the TPAI with the initial certification of an elementary reading specialist is not supported by the data. One may certainly speculate whether the TPAI evaluation for a teacher of the academically talented would be as valid as the TPAI evaluation of a remedial math teacher. While the TPAI may be justified as a potential predictor of how a teacher may bring about achievement in math, this study does not support any conclusion concerning the overall effectiveness of the teacher.

Teacher Variables and Potential Bias

No evidence was discovered to support any contention that the TPAI evaluations were biased by teacher race, sex or age. Further, such unlikely sources of bias as years in school were found to be unrelated to TPAI rating. While it

is fair to conclude that the use of the TPAI instrument in the sample was not significantly biased, the study does not conclude that the potential for biased results is not there. For example, there was a consistent, but insignificant, negative relationship between grade taught and the teacher's TPAI results. It was quite possible that a single evaluator was behaving in an idiosyncratic way and was systematically biased in favor of lower grade teachers. The neutral effect of other raters could have diluted the bias of a single rater and therefore the bias for the sample was not significant. In this sample that speculation was examined and found not to be justified. However, TPAI users should be constantly vigilant to this and other possible sources of bias.

Student Variables and Potential Bias

Of all the study's findings, the mixed results of the tests for relationship between TPAI and student variables averaged by class were the most difficult to analyze. Only five of 63 coefficients were found to be significant at the $p < .05$ level (see Table 18, page 106, this document). However, there was no pattern to the significant findings. There did appear to be a consistent but not statistically significant pattern, in the TPAI's relationship to grade, sex distribution of class membership, absences and the number of parents living with the child. The five

significant coefficients and the pattern of influence these student variables have on TPAI ratings demand consideration and a review of the evaluating process.

Interpreting these patterns without further study could result in misleading inferences. For example, the negative relationship of all eight functions to an increase in absenteeism could possibly be explained by suggesting that less effective teachers are less appealing to students and therefore they fail to come to school as regularly as they more satisfied peers. In this case, a low TPAI might be justified. However, the reverse is equally plausible. In this case the teacher receives a lower rating on her evaluation because her students do not respond well to excellent instruction and are more difficult to teach. They are absent more because of their disinterest and a lack of parental support.

With five of the coefficients significant at the $p < .05$ level and a pattern of coefficients approaching significance for the same student variables, declaring the TPAI free of bias due to influence of student variables would be premature. Perhaps principals in the study were trying to equate the effectiveness of the instruction they observed with the perceived difficulty of the task for a given set of students. If this is the case, the implications of such informal adjustments as the halo effect could be damaging both to the validity of the

instrument and to any interrater reliability the instrument might have. Whatever the cause of the five significant coefficients, this researcher cannot conclude that the TPAI is free of bias due to student variables which reflect the composition of the classroom membership. Even though the significant coefficients were small in number and may be due to idiosyncratic judgments of individual evaluators, there is sufficient doubt to reserve the claim that the TPAI is fair to all teachers regardless of the composition of class membership. The issue is clearly undecided.

Recommendations

As a result of the study, three recommendations concerning the continued use and development of the TPAI in North Carolina are advocated. First, the use of the TPAI as a basis for inferences regarding licensure and promotion should be suspended until a technical manual is developed and made available to test users and to educational agencies and professional associations whose membership would have a vital interest in its use. Second, additional validity studies should be conducted which would include validating criteria that would be sensitive to highly structured learning goals other than math. This would help ascertain whether the relationship between TPAI and math is unique or can be generalized to highly structured learning in other areas. Third, the use of the TPAI as a general

measure of teacher effectiveness should be discontinued. The results of TPAI evaluations should be employed as evidence that teachers have learned and can demonstrate a particular pattern of teaching behavior that has been shown to be related to the prediction of math achievement. Alternative teaching patterns should continue to be encouraged.

Creation of a Technical Manual

It became apparent early in the study that there was a great need for the North Carolina Department of Public Instruction to create a technical manual for the TPAI. This is necessary to comply with the primary standards as defined by the Standards for Educational and Psychological Testing (AERA et al., 1985). The use of an instrument for evaluation to be employed on such a grand scale as that intended for the TPAI must be studied in a manner consummate to primary professional standards.

The creation and dissemination of a technical manual will be a positive step in meeting these minimal expectations. This manual must include construct definitions, reliability data based on current use of the instrument, formal content validity studies, additional criterion validity studies, and an assessment of the minimal criteria necessary to qualify evaluators to insure fairness across raters. The use of the instrument as a

basis for decisions regarding licensure and promotion should be suspended until these studies have been completed and the instrument has been shown to meet the minimal criteria required by these inferences.

Additional Validity Studies

It is imperative that additional validity studies be conducted on a continuing basis. These additional studies include formal content validity studies as well as additional criterion validity studies. Formal content studies should include a cross section of professionals encompassing members from the groups the instrument is intended to evaluate.

Additional criterion validity studies should be designed to replicate the current study. Attempts to define highly structured teaching goals appropriate for the grade that are considered imperative and expected of each teacher should be conducted. Criterion measures sensitive to these goals should be employed to test the hypothesis that the TPAI is predictive of achievement in highly structured content areas other than math. Additionally, the TPAI should be compared to process oriented learning goals such as those found in social studies and literature. If a continued non-significant relationship is found, inferences regarding teacher effectiveness in a broad range

of instructional endeavors should not be made and can not be defended.

Since it was shown that empirically based items were no more valid in the prediction of achievement than consensus based items, additional items might be sought for the TPAI to assist in identifying teaching behavior that is successful in predicting achievement in loosely structured content areas. These additional items might result from content validation studies of current items with a sample of new consensus based items included. Whether an enhancement of the current TPAI is or is not sought, additional validity studies must be made.

Cessation of Inferences from the TPAI

Although the study was successful in providing validity evidence for limited use of the TPAI, inferences made from the TPAI should be examined through additional study. Certainly, the current belief that the TPAI is valid as an omnibus measure of teacher effectiveness was in no way supported by the data. Inferences based on TPAI scores regarding the general effectiveness of the teacher can not be held valid until a clear and thorough analysis of the desired functioning required of each teaching position is conducted. This analysis should be designed to ascertain which teaching positions would be expected to successfully utilize the skills advocated by the TPAI in

daily job functioning and what proportion of time those skills should be employed. The TPAI evaluation can carry no more weight than the proportion of time a teacher is expected to spend in highly structured direct teaching.

Concluding Statement

The evaluation of teaching for the improvement of instruction is an important endeavor and deserves careful study and generous resources. The limited knowledge that exists about the value and validity of such endeavors should not be put aside for political expediency. Other avenues exist for the improvement of instruction. Until those avenues are exhausted, instruments with limited validity and reliability should not be employed as criteria for decision making. Teacher rating instruments have not been shown to have positive effects that are worthy of emulation. Any effort toward evaluating teachers must have clearly stated goals and must be evaluated not only by the value of these goals, but of the ability of the instrument to realize these ambitions.

This study examined the criterion validity of the TPAI by the criterion of achievement test gains. This criterion was considered most suited to the results the TPAI sought: basic skill gains. It was shown to have validity in the prediction of math achievement when the effects of grade, IQ and student absences were held constant. It was not

shown to be a significant predictor of total achievement and reading. It was noted that the TPAI ratings were not related to teacher variables nor were measures of student achievement related to teacher variables. The TPAI ratings demonstrated a significant relationship with some student variables but no pattern of bias emerged. The research findings are encouraging because the validity coefficients using math as a criterion are among the best ever found for a rating scale. However, these results are not sufficient to warrant generalized inferences about teacher effectiveness in a broad range of endeavors. It was shown that in the prediction of reading achievement, for example, the relationship of reading to TPAI rating could be explained by chance.

In conclusion, the opinion of one of the writers of the 1965 report to the North Carolina General Assembly (1965) is worthy of reflection:

> Teaching is an art as well as a science and is too complex to be evaluated objectively. Thus far, it has been impossible to measure teacher competence accurately because of the human qualities in evaluators. Excellence in teaching resists measurement (p. 9).

The TPAI represents progress toward the goal of evaluating teachers objectively, but much is yet to be done.

BIBLIOGRAPHY

Aleamoni, L. (1981). Student ratings of instruction. In Millman J. (ed.) (1981). Handbook of teacher evaluation. Beverly Hills: Sage Publications.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, D. C.: American Psychological Association.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. (1985). Becoming a nation of readers: The report of the commission on reading. Washington, D. C.: The National Institute of Education.

Ashton, P. T. & Webb, R. B. (1986). Making a difference: Teachers' sense of efficacy and student achievement. New York: Longman.

Associated Press. (1987, August 14). Teachers will get hearing. The Charlotte Observer.

Baker, E. L. (1988). Can we fairly measure the quality of education? NEA Today, 6(6), 9-14.

Barr, A. S., Torgenson, T. L., Johnson, C. E., Lyon, V. E. & Walvoord, A. C. (1935). The validity of certain instruments employed in the measurement of teaching ability. In Lancelot, W. H., Barr, A. S., Torgenson, T. L., Johnson, C. E.Lyon, V. E., Walvoord, A. C. & Betts, G. L. (1935). The measurement of teaching efficiency. New York: The Macmillan Company.

Beecher, D. E. & Bump, J. W. (1950). The evaluation of teaching in New York State: Standards and procedures recommended by local advisory committees. New York: The State University of New York.

Belgard, M., Rosenshine, B., & Gage, N. (1971). Exploration of the teacher's effectiveness in learning. In Westbury, I. & Bellack, A. eds. Research into classroom processes: Recent developments and next steps. New York: Teachers College Press.

Berliner, D. (1977). Impediments to measuring teacher effectiveness. In Borich, G. (1977). The appraisal of teaching: Concepts and process. Reading, MA: Addison-Wesley.

Berliner, D. C. (1984). The half-full glass: A review of research on teaching. In Hosford, P. L. Using what we know about teaching. (pp. 51-77). Alexandria, VA: Association for Supervision and Curriculum Development.

Berman, P. & McLaughlin, M. W. (1977). Federal programs supporting educational change: Factors affecting implementation and continuation. Santa Monica, CA: Rand Corp.

Blanton, W. E. & Moorman, G. B. (1987) The effective teaching training program--What is it teaching about teaching? North Carolina Education, 18(2), 12-13.

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

Bracey, G. W. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. Phi Delta Kappan, 68(9), 683-686.

Bracey, G. W. (1987). The muddles of measurement-driven instruction. Phi Delta Kappan, 68(9), 688-689.

Brophy, J. E. (1973). Stability and teacher effectiveness. American Educational Research Journal, 10, 245-252.

Brophy, J. E. (1974). Achievement correlates. In Walberg, H. J. (Ed.). Evaluating educational performance: A sourcebook of methods, instruments, and examples. (pp. 33-56). Berkeley, CA: McCutchan.

Brophy, J. E. (1987). Synthesis of research on strategies for motivating students to learn. Educational Leadership. (October) pp. 40-49.

Broudy, H. S. & Palmer, J. R. (1965). Exemplars of teaching method. Chicago: Rand McNally.

Capie, W. (1980). Teacher performance assessment instruments. Atlanta: Georgia Department of Education.

Capie, W. (1980, April). Using pupil achievement to validate ratings of student teacher performance. Paper presented at the 64th Annual Meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 191 916)

Career development: North Carolina not alone in incentive effort. (1987). Education Report, 3(4), 3.

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

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rhinehart, and Winston.

Cronbach L. (1971). Validity. In Educational measurement. R. Thorndike (Editor), Washington: ACE.

Darling-Hammond, L. (1986). A proposal for evaluation in the teaching profession. The Elementary School Journal, (86) 553-569.

Dewey, J. (1916). Democracy and education. New York: A Free Press.

Doyle, K. (1983). Evaluating teaching. Lexington, MA: Lexington Books.

Duke, D. & Stiggins, R. (1986). Teacher evaluation: Five keys to growth. Washington, D.C.: National Education Association.

Educational Research Service. (1978). Evaluating teacher performance. Arlington, VA: ERS.

Empey, D. W. (1984). The greatest risk: Who will teach?. The Elementary School Journal, 85(2), 167-176.

Fixing the career ladder. (1987, November 15). The Charlotte Observer. 2-C.

French-Lazovik, G. (1981). Peer review: Documentary evidence in the evaluation of teaching. In Millman J. (ed.) (1981). Handbook of teacher evaluation. Beverly Hills: Sage Publications.

Genck, F. (1984). School management model: Teacher evaluation and development Vol 3. Chicago: Institute for Public Management.

Glass, G. V. (1977). A review of three methods of determining teacher effectiveness. In Borich, G. (1977). The appraisal of teaching: Concepts and process. Reading, MA: Addison-Wesley.

Glass, G. V. (1974). Teacher effectiveness. In Walberg, H. J. (Ed.). Evaluating educational performance: A sourcebook of methods, instruments, and examples. (pp. 11-32). Berkeley, CA: McCutchan.

Glass, G. V. & Hopkins, K. D. (1984). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Gudridge, B. (1980). AASA Critical issues report: Teacher competency problems and solutions. Sacramento: AASA.

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

Hatry, H. P. & Greiner, J. M. (1985). Issues and case studies in teacher incentive plans. Washington: The Urban Institute P. 23.

Holdzkom, David. (1987). Appraising teacher performance in North Carolina. Educational Leadership. (April) pp. 40-44.

Holdzkom, D. & Kuligowski, B. (1987). Career development: Almost all you ever wanted to know.... Education Report, 3(3), 3-4.

Hosford, P. L. (1984). Using what we know about teaching. Alexandria, VA: Association for Supervision and Curriculum Development.

Karnes, E. L. & Black, D. D. (1986). Teacher evaluation and merit pay: An annotated bibliography. New York: Greenwood Press.

Keever, G. (1987). Career ladder frustration builds. North Carolina Education, 18(2), 36.

Kowalski, J. (1978). ERS report: Evaluating teacher performance. Arlington, VA: ERS.

Lancelot, W. H., Barr, A. S., Torgenson, T. L., Johnson, C. E., Lyon, V. E., Walvoord, A. C. & Betts, G. L.(1935). The measurement of teaching efficiency. New York; The Macmillan Company.

Lewis, J. (1973). Appraising teacher performance. West Nyach, NY: Parker.

Marks, M. (1976). Effective teacher evaluation. NASSP Bulletin v 60 #401.

Medley, D. M. & Coker, H. (1987). How valid are principal's judgments of teacher effectiveness? Phil Delta Kappan, 69(2), 138-140.

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

Millman J. (ed.) (1981). Handbook of teacher evaluation. Beverly Hills: Sage Publications.

Millman J. (1981). Student achievement as a measure of teacher competence. In Millman J. (ed.) (1981). Handbook of teacher evaluation. Beverly Hills: Sage Publications.

Mohan, M. & Hull, R. E. (1975). Teaching effectiveness: Its meaning, assessment, and improvement. Englewood Cliffs, NJ: Educational Technology Publications.

National Education Association, Professional and Organizational Development, Instruction and Professional Development. (1985). School Personnel Evaluation Manual. Washington, D.C.: National Education Association.

National School Public Relations Association. (1974). Evaluating teachers for professional growth: Current trends in school policies and programs. Arlington, VA: Author.

Nelson, K. G., Bicknell, J. E. & Hedlund, P. A. (1956). Development and refinement of measures of teaching effectiveness: First report on the Study to predict effectiveness in secondary school teaching. Albany, NY: University of the State of New York.

North Carolina State Department of Public Instruction. (1963). A progress report to the 1963 General Assembly by the North Carolina experimental program of teacher merit pay. Raleigh, NC: Author.

North Carolina State Department of Public Instruction. (1965). The North Carolina teacher merit pay study: A four-year experimental study in three pilot centers, Gastonia, Martin County, Rowan County; A report to the 1964 General Assembly. Raleigh, NC: Author.

North Carolina State Department of Public Instruction. (1985). North Carolina effective teacher training program. Raleigh, NC: Author.

North Carolina State Department of Public Instruction. (1985). North Carolina performance appraisal training program. Raleigh, NC: Author.

North Carolina State Department of Public Intruction. (1985). North Carolina standard course of study and introduction to the competency-based curriculum. Raleigh, NC: Author.

North Carolina State Department of Public Instruction. (1986). Teacher performance appraisal system training: A report of outcomes. Raleigh, NC: Author. (Eric Document Reproduction Service No. ED 271 452)

North Carolina State Department of Public Instruction. (1986). Teacher performance appraisal system: The standards and processes for use. Raleigh, NC: Author. ERIC Document Reproduction Service No. ED 271 453)

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston.

Peterson, P. L. & Walberg, H. J. (Eds.). (1979). Research on teaching: Concepts, findings, and implications. Berkeley, CA: McCutchan.

Pigford, A. B. (1987). Teacher evaluation: More than a game that principals play. Phi Delta Kappan, 69(2), 143-144.

Popham, W. J. (1971). Designing teacher evaluation systems: A series of suggestions for establishing teacher assessment procedures as required by the Stull Bill (AB 293), 1971 California Legislature. Los Angeles: The Instructional Objective Exchange.

Popham, W. J. (1987). The merits of measurement-driven instruction. Phi Delta Kappan, 68(9), 679-682.

Popham, W. J. (1987). Muddle-minded emotionalism. Phi Delta Kappan, 68(9), 688-689.

Reavis, W. C. & Cooper, D. H. (1945). Evaluation of teacher merit in city school systems. Chicago: The University of Chicago.

Rosenholtz, S. J. (1986). Career ladders and merit pay: Capricious fads or fundamental reforms? The Elementary School Journal, (86) 513-530.

Rosenholtz, S. J. & Smylie, M. A. (1984). Teacher compensation and career ladders. The Elementary School Journal, 85(2), 149-166.

Rosenshine, B. (1970). The stability of teacher effects upon student achievement. Review of Educational Research, 40, 649-642.

Rosenshine, B. (1973). Stability of teacher effectiveness. American Education Research Journal, 10, 245-252.

Rosenshine, B. (1977). The stability of teacher effects upon student achievement. In Borich, G. (1977). The appraisal of teaching: Concepts and process. Reading, MA: Addison-Wesley.

Rosenshine, B. (1986). Synthesis on research on explicit teaching. Educational Leadership. (April), 60-69.

Soar, R. (1977). Teacher assessment problems and possibilities. In Borich, G. (1977). The appraisal of teaching: Concepts and process. Reading, MA: Addison-Wesley.

Spuck, D. W. (1974). Geocode analysis. In Walberg, H. J. (Ed.). Evaluating educational performance: A sourcebook of methods, instruments, and examples. (pp. 339-350). Berkeley, CA: McCutchan.

Stacey, Dennis C. (1988, February). Evaluation of the effectiveness of the North Carolina teacher performance appraisal system (TPAS). Paper presented to the North Carolina Association for Research in Education

Strike, K. & Bull, B. (1981). Fairness and the legal context of teacher evaluation. In Millman J. (ed.) (1981). Handbook of teacher evaluation. Beverly Hills: Sage Publications.

Talmage, H. & Rippey, R. M. (1974). Elementary school cases. In Walberg, H. J. (Ed.). Evaluating educational performance: A sourcebook of methods, instruments, and examples. (pp. 255-276). Berkeley, CA: McCutchan.

Teacher Performance Assessment Instruments. (1985). Atlanta, GA: Georgia Department of Education.

Thomas, M. D. (1979). Performance evaluation of educational personnel. Bloomington, IL: Phi Delta Kappan.

Triosi, Nicholas F. (1983). Effective teaching and student achievement. Reston, VA: NASSP.

Walberg, H. J. (Ed.). (1974). Evaluating educational performance: A sourcebook of methods, instruments, and examples. Berkeley, CA: McCutchan.

Walker, H. M. (1935) Preface. In Lancelot, W. H., Barr, A. S., Torgenson, T. L., Johnson, C. E., Lyon, V. E., Walvoord, A. C. & Betts, G. L. The measurement of teaching efficiency. New York; The Macmillan Company.

Weeks, K., & Cornett, L. (1984). Career ladder clearinghouse. Nashville, TN.: Vanderbilt Institute for Public Policy Studies, Vanderbilt University.

Westbury, I. & Bellack, A. eds. (1971). Research into classroom processes: Recent developments and next steps. New York: Teachers College Press.

Williams, R., Woods, P., Shoaf, T., Little, D., Knight, J. P., Hayes, H., Flynn, P., Doss, C., Clark, M., Chambliss, T. & Bair, B. (1987). "There's many a slip between the cup and the lip.". North Carolina Education, 17(4), 26-27.

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W. & Bernstein, H. T. (1985). Teacher evaluation: A study of effective practices. The Elementary School Journal, 86(1), 61-121.

APPENDIX A

DATA COLLECTION INSTRUMENTS

_____ SCHOOLS

TEACHER PERFORMANCE APPRAISAL INSTRUMENT

INSTRUCTIONS. Located at the end of Instrument

Superior	Well Above Standard	Above Standard	At Standard	Below Standard	Unsatisfactory

Teacher Name _____

School _____

1. Major Function: MANAGEMENT OF INSTRUCTIONAL TIME ☐ ☐ ☐ ☐ ☐ ☐

1.1 Teacher has materials, supplies and equipment ready at the start of the lesson or instructional activity.

1.2 Teacher gets the class started quickly.

1.3 Teacher gets students on task quickly at the beginning of each lesson or instructional activity.

1.4 Teacher maintains a high level of student time-on-task.

Comments _____

_____

2. Major Function: MANAGEMENT OF STUDENT BEHAVIOR ☐ ☐ ☐ ☐ ☐ ☐

2.1 Teacher has established a set of rules and procedures that govern the handling of routine administrative matters.

2.2 Teacher has established a set of rules and procedures that govern student verbal participation and talk during different types of activities—whole-class instruction, small group instructions, etc.

2.3 Teacher has established a set of rules and procedures that govern student movement in the classroom during different types of instructional activities.

2.4 Teacher frequently monitors the behavior of all students during whole-class, small group, and seat work activities and during transitions between instructional activities.

2.5 Teacher stops inappropriate behavior promptly and consistently, yet maintains the dignity of the student.

Comments _____

_____

3. Major Function: INSTRUCTIONAL PRESENTATION ☐ ☐ ☐ ☐ ☐ ☐

3.1 Teacher begins lesson or instructional activity with a review of previous material.

3.2 Teacher introduces the lesson or instructional activity and specifies learning objectives when appropriate.

3.3 Teacher speaks fluently and precisely.

3.4 Teacher presents the lesson or instructional activity using concepts and language understandable to the students.

3.5 Teacher provides relevant examples and demonstrations to illustrate concepts and skills.

3.6 Teacher assigns tasks that students handle with a high rate of success.

3.7 Teacher asks appropriate levels of questions that students handle with a high rate of success.

Superior	Well Above Standard	Above Standard	At Standard	Below Standard	Unsatisfactory

3.8 Teacher conducts lessons or instructional activity at a brisk pace, slowing presentations when necessary for student understanding but avoiding unnecessary slowdowns.

3.9 Teacher makes transitions between lessons and between instructional activities within lessons efficiently and smoothly.

3.10 Teacher makes sure that the assignment is clear.

3.11 Teacher summarizes the main point(s) of the lesson at the end of the lesson or instructional activity.

Comments _____

_____

4. Major Function: INSTRUCTIONAL MONITORING OF STUDENT PERFORMANCE ☐☐☐☐☐☐

4.1 Teacher maintains clear, firm and reasonable work standards and due dates.

4.2 Teacher circulates during classwork to check all students' performance.

4.3 Teacher routinely uses oral, written, and other work products to check student progress.

4.4 Teacher poses questions clearly and one at a time.

Comments _____

_____

5. Major Function: INSTRUCTIONAL FEEDBACK ☐☐☐☐☐☐

5.1 Teacher provides feedback on the correctness or incorrectness of in-class work to encourage student growth.

5.2 Teacher regularly provides prompt feedback on assigned out-of-class work.

5.3 Teacher affirms a correct oral response appropriately, and moves on.

5.4 Teacher provides sustaining feedback after an incorrect response or no response by probing, repeating the question, giving a clue, or allowing more time.

Comments _____

_____

6. Major Function: FACILITATING INSTRUCTION ☐☐☐☐☐☐

6.1 Teacher has an instructional plan which is compatible with the school and systemwide curricular goals.

6.2 Teacher uses diagnostic information obtained from tests and other assessment procedures to develop and revise objectives and/or tasks.

6.3 Teacher maintains accurate records to document student performance.

6.4 Teacher has instructional plan that matches/aligns objectives, learning strategies, assessment and student needs at the appropriate level of difficulty.

6.5 Teacher uses available human and material resources to support the instructional program.

Comments _____

_____

Superior	Well Above Standard	Above Standard	At Standard	Below Standard	Unsatisfactory

7. Major Function: INTERACTING WITHIN THE EDUCATIONAL ENVIRONMENT ☐ ☐ ☐ ☐ ☐ ☐

7.1 Teacher treats all students in a fair and equitable manner.

7.2 Teacher interacts effectively with students, co-workers, parents, and community

Comments _____

_____

8. Major Function: PERFORMING NON-INSTRUCTIONAL DUTIES ☐ ☐ ☐ ☐ ☐ ☐

8.1 Teacher carries out non-instructional duties as assigned and/or as need is perceived.

8.2 Teacher adheres to established laws, policies, rules, and regulations.

8.3 Teacher follows a plan for professional development and demonstrates evidence of growth.

Comments _____

_____

Evaluator's Summary Comments _____

_____

_____

_____

_____

_____

_____

Teacher's Reactions to Evaluation _____

_____

_____

_____

_____

_____

_____

EVALUATOR'S SIGNATURE _____ DATE _____ TEACHER'S SIGNATURE _____ DATE _____

(SIGNATURE INDICATES THAT THE WRITTEN EVALUATION HAS BEEN SEEN AND DISCUSSED)

INSTRUCTIONS

1. Based on the evidence from observation and discussion, the evaluator is to rate the teacher's performance with respect to the 8 major functions of teaching listed below.
2. The evaluator is encouraged to add pertinent comments at the end of each major function.
3. The teacher is provided an opportunity to react to the evaluator's ratings and comments.
4. The evaluator and the teacher must discuss the results of the appraisal and any recommended action pertinent to it.
5. The teacher and the evaluator must sign the instrument in the assigned spaces.
6. The instrument must be filed in the teacher's personnel folder.
7. The rating scale will be as follows:

LEVEL OF PERFORMANCE

6. Superior

Performance within this function area is consistently outstanding. Teaching practices are demonstrated at the highest level of performance. Teacher continuously seeks to expand scope of competenices and constantly undertakes additional, appropriate responsibilities.

5. Well Above Standard

Performance within this function area is frequently outstanding. Some teaching practices are demonstrated at the highest level while others are at a consistently high level. Teacher frequently seeks to expand scope of competencies and often undertakes additional, appropriate responsibilities.

4. Above Standard

Performance within this function area is frequently high. Some teaching practices are demonstrated at a high level while others are at a consistently adequate/acceptable level. Teacher sometimes seeks to expand scope of competencies and occasionally undertakes additional, appropriate responsibilities.

3. At Standard

Performance within this function area is consistently adequate/acceptable. Teaching practices fully meet all performance expectations at an acceptable level. Teacher maintains an adequate scope of competencies and performs additional responsibilities as assigned.

2. Below Standard

Performance within this function area is sometimes inadequate/unacceptable and needs improvement. Teacher requires supervision and assistance to maintain an adequate scope of competencies, and sometimes fails to perform additional responsibilities as assigned.

1. Unsatisfactory

Performance within this function area is consistently inadequate/unacceptable and most practices require considerable improvement to fully meet minimum performance expectations. Teacher requires close and frequent supervision in the performance of all responsibilities.

Initial Letter to Teachers

Dear Teacher,

You are being asked to participate in a research project to examine the criterion validity of the North Carolina Teacher Performance Appraisal Instrument (TPAI). The state has proposed that teachers should be paid by merit, that is, teachers who make greater contributions to children's academic achievement should receive higher salaries. The TPAI is one the factors involved in making the decisions concerning merit pay. The purpose of this research is to examine the contribution the TPAI makes in predicting achievement of students.

If you agree, you will be participating in a carefully designed data collection process which will require no changes in your teaching style or any intrusion by researchers in your classroom. Your participation will be limited to the following activities:

Completing a questionnaire about yourself, your training, and your views about teaching.

Completing a data sheet on a randomly selected number of your students.

Agreeing to allow your principal to fill out a questionnaire about you which will include the eight areas of the TPAI.

After your annual evaluation, completing a questionnaire concerning your experiences in the evaluation process.

The total time you must invest to complete all the above activities should not exceed one hour.

All responses by participants will remain confidential and are not for use by, nor available to, employees of the [redacted] School System. Names of participants will be removed from the data and coded for confidentiality. As a participant in this project, you have certain rights. They are outlined on the attached sheet.

Your participation in this project will be greatly appreciated. The results of the study will enlighten educators and policy makers in the area of teacher evaluation using the TPAI.

Sincerely,

Phil Riner
University of North Carolina at Greensboro

L E T T E R O F I N F O R M E D C O N S E N T

Research Project: Assessing the Validity of the TPAI
Researcher: Phillip Riner

Purpose: The purpose of the research project is to assess the criterion validity of North Carolina's Teacher Performance Appraisal Instrument (TPAI).

Method of study: The method of study is a statistical analysis of the TPAI scores of volunteer teachers and their relationship to student achievement, teacher personological variables, student personological variables, and teacher attitudes and self-assessments.

Participant Involvement: Participants are volunteers who are currently teaching in the research district. Participants are asked to:

-complete a questionnaire about personal history, teacher training, and views about teaching effectiveness

-complete a data sheet on a randomly selected number of the participants students

-agreeing to allow your principal to fill out a questionnaire about you which will include the eight areas of the TPAI

-after the sixth month of school, complete a questionnaire concerning the your experiences in the evaluation process

Protection of Participants: As a participant of a research study you are to be protected from any potential harmful and unpleasant effects.

-You may discontinue participation any time

-You may see any data collected concerning you

-All data collected in the research is confidential and are not for use by, nor available to, employees of Kannapolis City Schools (KSC) nor any other agency other than the researcher

-Names of participants will be removed from the data and coded for research use

-The individual source of any data will confidential and will be made available only to the participant by written request

-Participant involvement is limited to the activities outlined above in this document

-You may decline to respond to any question or request for information contained in any of the questionnaires by leaving that item blank

-You will be provided opportunity to inspect all data collection instruments used in this research prior to consenting to participation

Responsibilities of Participants: Participants will be expected to give honest candid responses to all requests for data (the response may a decline for the specific item request). Participants will be expected to meet the requested due dates for data submission.

Responsibilities of Researcher: The researcher has the responsibility to conduct an ethical and competent research study. To whit, the researcher is expected to...

-honor all the assurances contained in this document

-endeavor to protect subjects from any physical or mental discomfort, harm, or danger resulting from participation in the research and to inform subject if this possibility exists

-provide a summary of the results and any conclusions of the research project

-to protect the privacy of each participant

Statement of Informed Consent:

I hereby agree to participate in the above described study. I have had the purpose and method of the study explained to me. I have been afforded an opportunity to examine all data collection instruments involved in the study.

Furthermore, I agree to allow my principal to complete the "Teacher Performance Appraisal Instrument and Evaluator's Questionnaire" using me as the subject of evaluation. I have been given a copy of this instrument.

I have been given an opportunity to ask questions and be given appropriate answers.

Participant Signature: _____

Date: _____

Teacher:_____ STUDENT DATA COLLECTION

Teacher ID#: _____

STUDENT FIRST LAST	ID #	GRADE	RACE	SEX	FREELUNCH	#PARENTS IN HOME	DAYS ABSENT as of end of sixth month	IQ	CAT 87 READ	CAT 87 MATH	CAT 87 TOTAL (SCALE SCORES)	CAT 88 READ	CAT 88 MATH	CAT 88 TOTAL (SCALES SCORES)

KEY

GRADE: Current grade level
RACE: W = WHITE B = BLACK A = ASIAN O = OTHER
SEX: M = MALE F = FEMALE
FREELUNCH Y = FREE OR REDUCED N = NOT PARTICIPATING IN FEDERAL LUNCH PROGRAM
# OF PARENTS IN HOME: 1 OR 2 INDICATES THE NUMBER OF PARENTS (OR STEP PARENTS) LIVING AT HOME WITH THE CHILD;
X = NOT SURE OR UNKNOWN 0 = FOSTER PARENTS, GRANDPARENTS
DAYS ABSENT (AS OF END OF SIXTH MONTH): THIS IS THE NUMBER OF ABSENCES (EXCUSED AND UNEXCUSED) DURING THE FIRST SIX SCHOOL MONTHS
IQ: THE MOST RECENT QUOTIENT SCORE, NOT THE PERCENTILE:
X = MISSING INFORMATION
IF TEST YIELDS MORE THAN ONE SCORE, INDICATE THE AVERAGE OF THE SCORES
CAT 87 READ = READING SUBTEST SCALE SCORE;
MATH = MATH SUBTEST SCALE SCORE;
TOTAL = TOTAL SCALE SCORE;
SCALE SCORES ARE USUALLY INDICATED BY "SS" AND IS A THREE DIGIT NUMBER
IF IN DOUBT LEAVE THE SCORE BLANK
IF NO ACHIEVEMENT DATA IS AVAILABLE, PLACE AN "N" IN THE FIRST BLOCK

IF CAT 88 SCORES ARE NOT AVAILABLE; WILL BE INSERTED AT A LATER DATE

TEACHER ID#: _____

TEACHER INFORMATION SHEET

_____ 1. How many years of teaching experience do you have in public schools (in any school district)?

_____ 2. How many years have you taught in this school?

_____ 3. How many years have you taught your current grade level?

_____ 4. How old were you when you began teaching?

_____ 5. What is your age?

_____ 6. What is your sex? 1 = male 0 = female

_____ 7. What is your race? 0 = white 1 = black
2 = asian 3 = other

_____ 8. What is your highest earned degree?

4 = bachelors 5 = masters
6 = advanced certificate (sixth year)
7 = doctorate

_____ 9. From what type of institution did you earn your bachelor degree?

0 = private 1 = state supported

_____ 10. What is the highest degree granted from your college/university at the time of your earning your bachelor's degree?

4 = bachelors 5 = masters 7 = doctorate

TEACHER ID#: _____

TEACHER SELF-RATING OF INSTRUCTION

Circle the number which best describes your performance in each area. Use the following code:

1 = Unsatisfactory 2 = Below Standard
3 = Satisfactory 4 = Above Standard
5 = Well Above Standard 6 = Superior

1. INSTRUCTIONAL TIME
   -Materials ready
   -Class started quickly
   -Gets students on task
   -Maintains high time-on-task

   1 2 3 4 5 6

2. STUDENT BEHAVIOR
   -Rules--Administrative Matters
   -Rules--Verbal Partcipation/Talk
   -Rules--Movement
   -Frequently monitors behavior
   -Stops inappropriate behavior

   1 2 3 4 5 6

3. INSTRUCTIONAL PRESENTATION
   -Begins with review
   -Introduces lesson
   -Speaks fluently
   -Lesson understandable
   -Provides relevant examples
   -High rate of success on tasks
   -Appropriate level of questions
   -Brisk pace
   -Efficient, smooth transitions
   -Assignments clear
   -Summarizes main points

   1 2 3 4 5 6

4. INSTRUCTIONAL MONITORING
-Maintains deadlines, standards
-Circulates to check student performance
-Uses oral, written work products to check progress
-Questions clearly and one at a time

1 2 3 4 5 6

5. INSTRUCTIONAL FEEDBACK
-Feedback on in-class work
-Prompt feedback on out-of-class work
-Affirms correct answer quickly
-Sustaining feedback on incorrect answers

1 2 3 4 5 6

6. FACILITATING INSTRUCTION
-Instructional plan compatible with goals
-Diagnostic information to develop tasks
-Maintain accurate records
-Instructional plan for curriculum alignment
-Available resources support program

1 2 3 4 5 6

7. COMMUNICATING WITHIN THE EDUCATIONAL ENVIRONMENT
-Treats all students fairly
-Interacts effectively within school and community

1 2 3 4 5 6

8. NON-INSTRUCTIONAL DUTIES
-Carries out non-instructional duties
-Adheres to laws, policies
-Plan for professional development

1 2 3 4 5 6

TEACHER ID#: _____

TEACHER PERFORMANCE APPRAISAL INSTRUMENT AND EVALUATOR'S QUESTIONNAIRE

Use the standard TPAI summative evaluation techniques to evaluate this teacher. Use the following codes:

1 = Unsatisfactory 2 = Below Standard 3 = At Standard
4 = Above Standard 5 = Well Above Standard 6 = Superior

1.	INSTRUCTIONAL TIME -Materials ready -Class started quickly -Gets students on task -Maintains high time-on-task	1	2	3	4	5	6
2.	STUDENT BEHAVIOR -Rules--Administrative Matters -Rules--Verbal Partcipation/Talk -Rules--Movement -Frequently monitors behavior -Stops inappropriate behavior	1	2	3	4	5	6
3.	INSTRUCTIONAL PRESENTATION -Begins with review -Introduces lesson -Speaks fluently -Lesson understandable -Provides relevant examples -High rate of success on tasks -Appropriate level of questions -Brisk pace -Efficient, smooth transitions -Assignments clear -Summarizes main points	1	2	3	4	5	6
4.	INSTRUCTIONAL MONITORING -Maintains deadlines, standards -Circulates to check student performance -Uses oral, written work products to check progress -Questions clearly and one at a time	1	2	3	4	5	6
5.	INSTRUCTIONAL FEEDBACK -Feedback on in-class work -Prompt feedback on out-of-class work -Affirms correct answer quickly -Sustaining feedback on incorrect	1	2	3	4	5	6

answers

6.	FACILITATING INSTRUCTION -Instructional plan compatible with goals -Diagnostic information to develop tasks -Maintain accurate records -Instructional plan for curriculum alignment -Available resources support program	1	2	3	4	5	6
7.	COMMUNICATING WITHIN THE EDUCATIONAL ENVIRONMENT -Treats all students fairly -Interacts effectively within school and community	1	2	3	4	5	6
8.	NON-INSTRUCTIONAL DUTIES -Carries out non-instructional duties -Adheres to laws, policies -Plan for professional development	1	2	3	4	5	6

EVALUATOR'S QUESTIONNAIRE

Circle the number that best represents your feelings toward the statemement. Use the following code.

1 = strongly agree 2 = agree 3 = undecided
4 = disagree 5 = strongly disagree

1.	The TPAI estimate above accurately reflects the actual TPAI this teacher would earn.	1	2	3	4	5
2.	The TPAI score for this teacher accurately assesses this teacher's total effectiveness in this school.	1	2	3	4	5
3.	If I were to do another TPAI on this teacher next week, the scores would be the same.	1	2	3	4	5
4.	In compiling the TPAI score for this teacher, I used data collected in the evaluation process only and not information gathered in daily contact with the teacher.	1	2	3	4	5
5.	I feel I am a competent judge of this teacher's effectiveness.	1	2	3	4	5
6.	If tenure or pay scale decisions were to be made on the basis of this TPAI score, I would	1	2	3	4	5

feel that a valid decision had been made.

7. I have evaluated this teacher using the TPAI previously. YES NO

8. This teacher is a probationary teacher in the initial certified personnel program. YES NO

Recanvas Letter

Dear Teachers,

With the assistance of Dr. [redacted] and your school principal, I recently visited your school to solicit your participation in a research project designed to explore the criterion validity of the Teacher Performance Appraisal Instrument (TPAI).

The teacher response was gratifying with almost more than sixty per cent of teachers participating. However, because of absences during flu season and conflicting duties during staff meetings not all teachers have been given an opportunity to participate.

I would like to invite you to read the accompanying letter and volunteer to participate in this research.

The time require to participate is minimal, probably less than an hour. All responses to data requests are confidential and the identity of each respondent is protected. It is an excellent opportunity to contribute to the development of knowledge surrounding the TPAI.

If you would be willing to participate in this research please read and sign the enclosed Letter of Informed Consent and return to me via interschool mail.

If you have any questions please call me at 932-5665.

Thank you for your help.

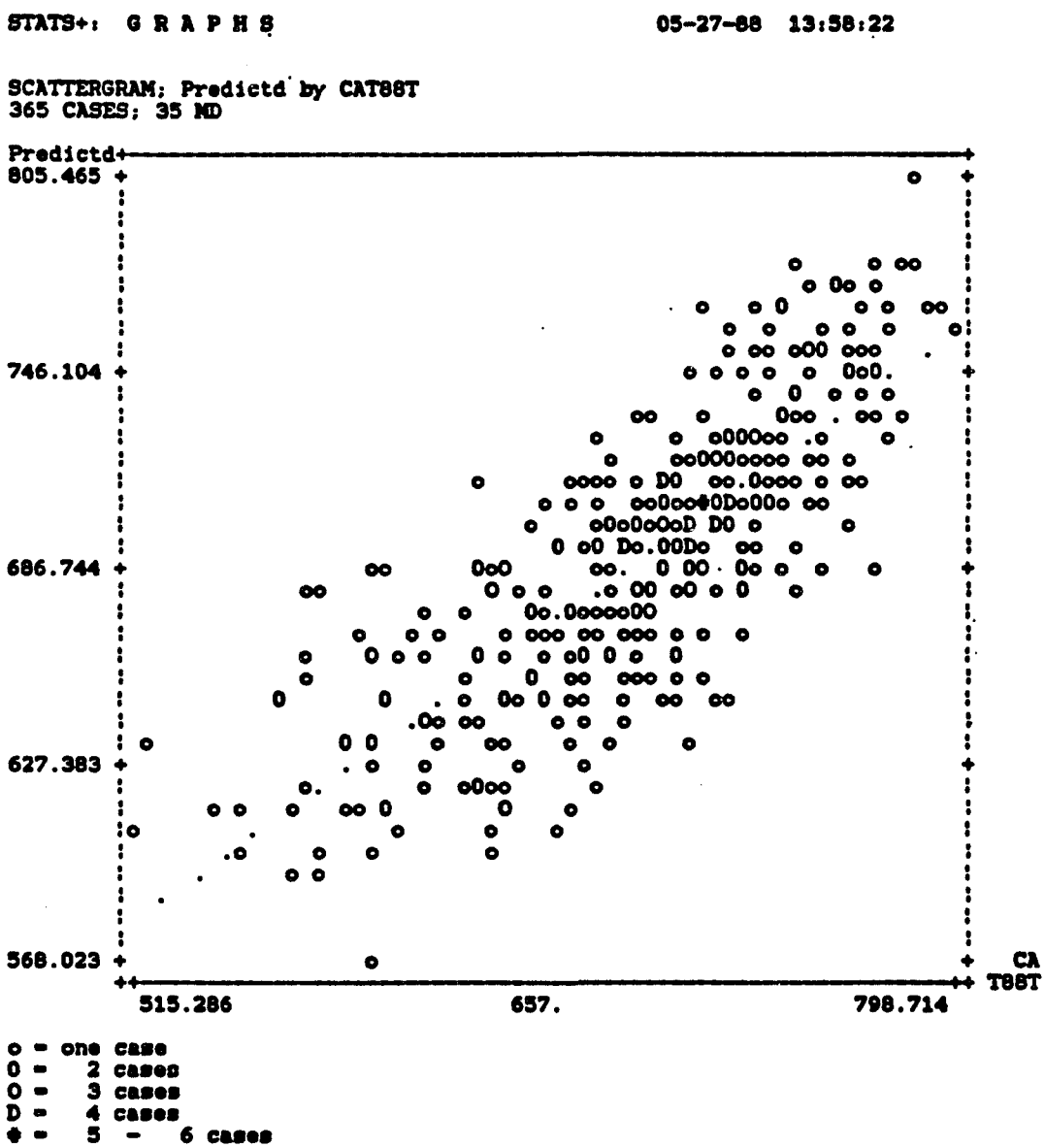
Sincerely,

Phil Riner

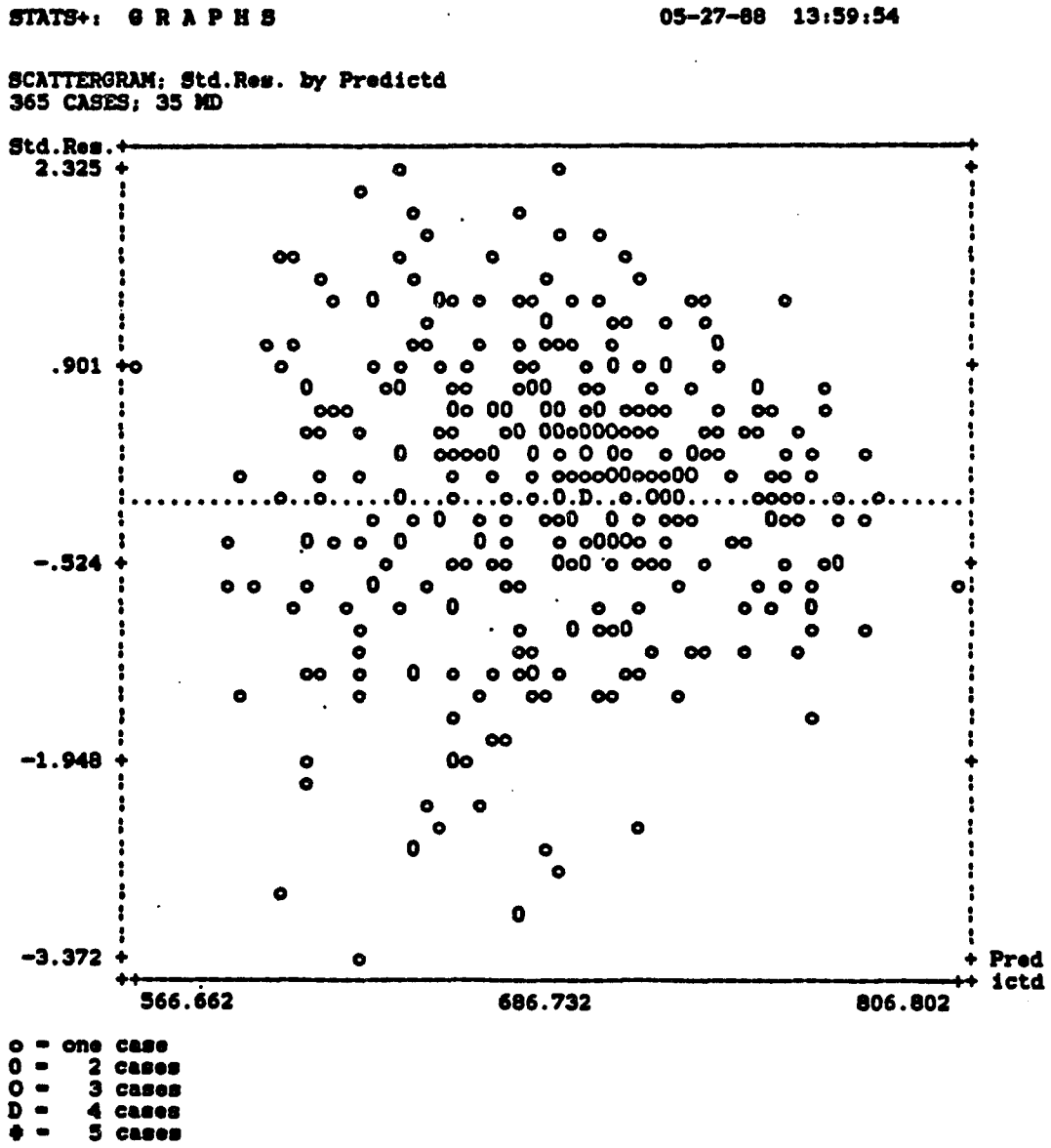
APPENDIX B

ADDITIONAL STATISTICS

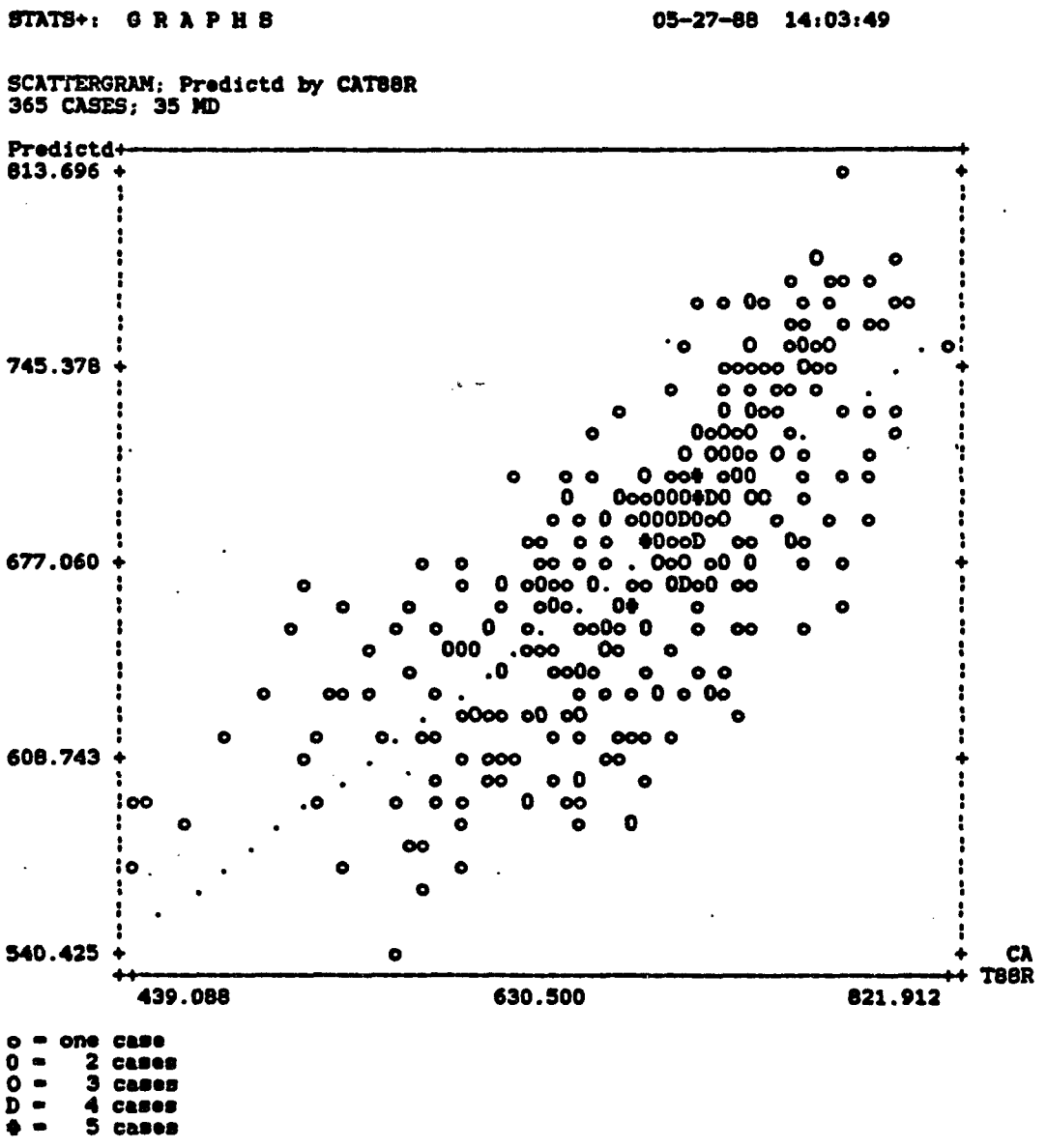
Predicted Scores by CAT Total Posttest



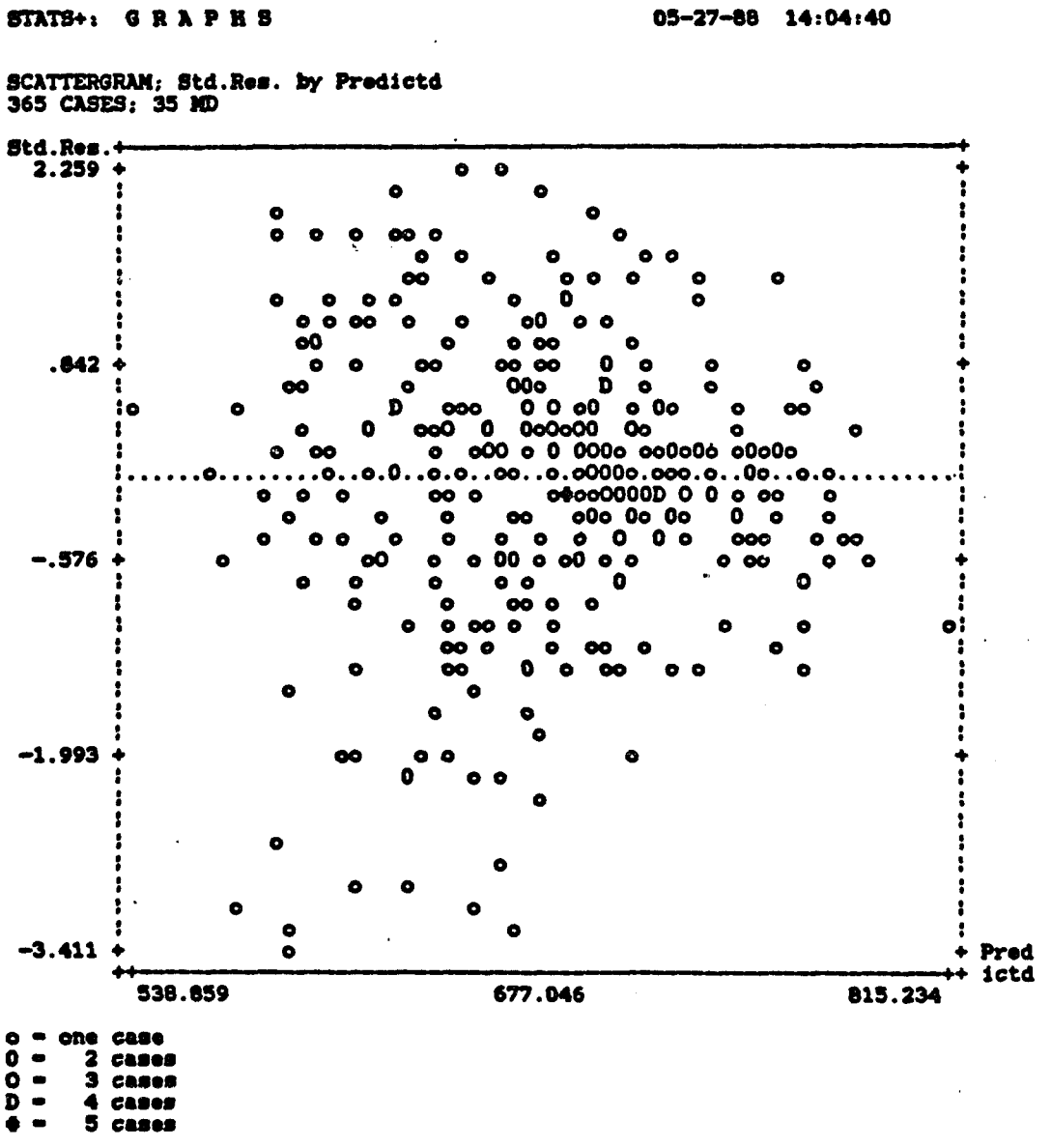
Standardized Residuals by Predicted CAT Total



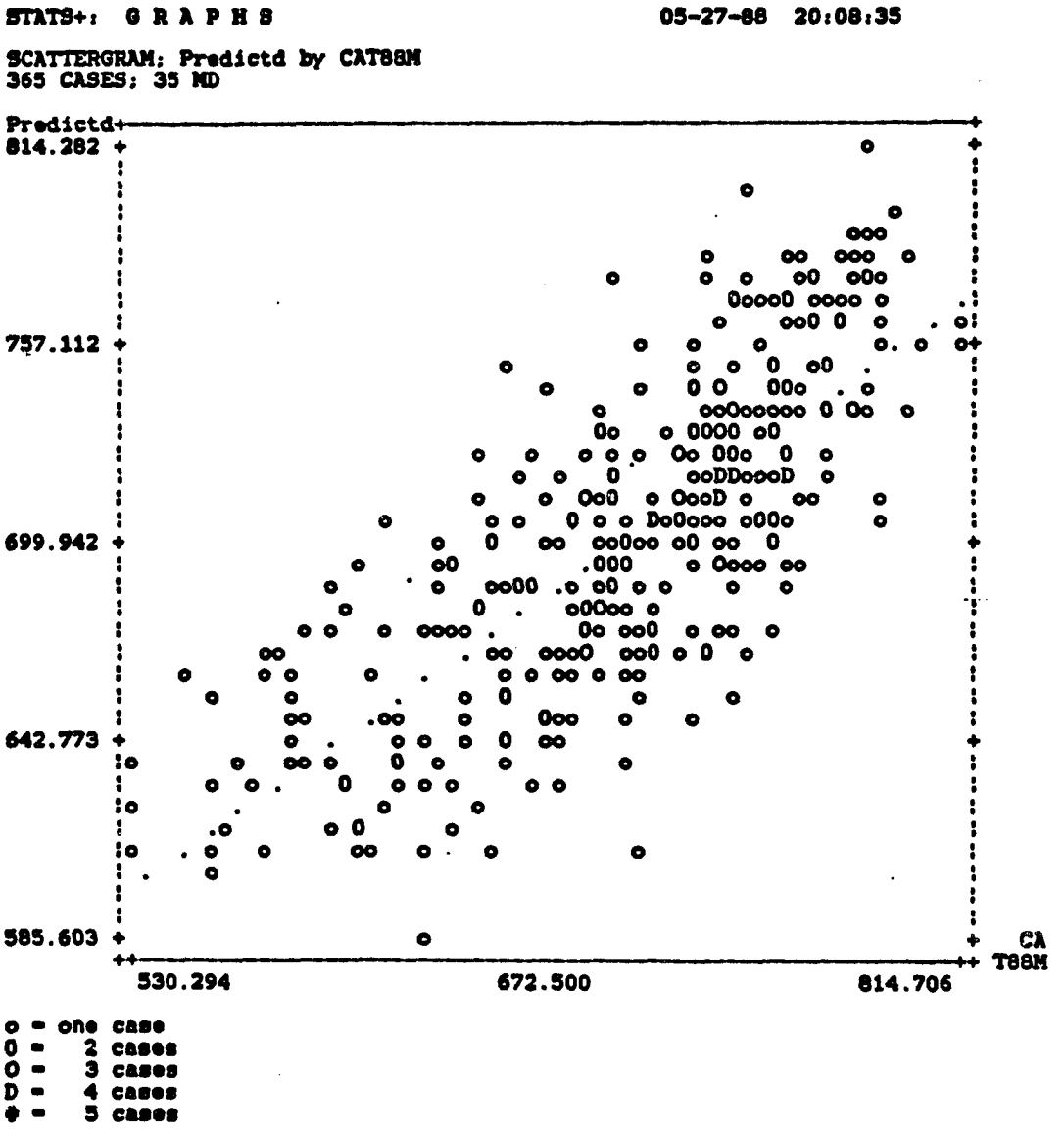
Predicted Scores by CAT Math Posttest



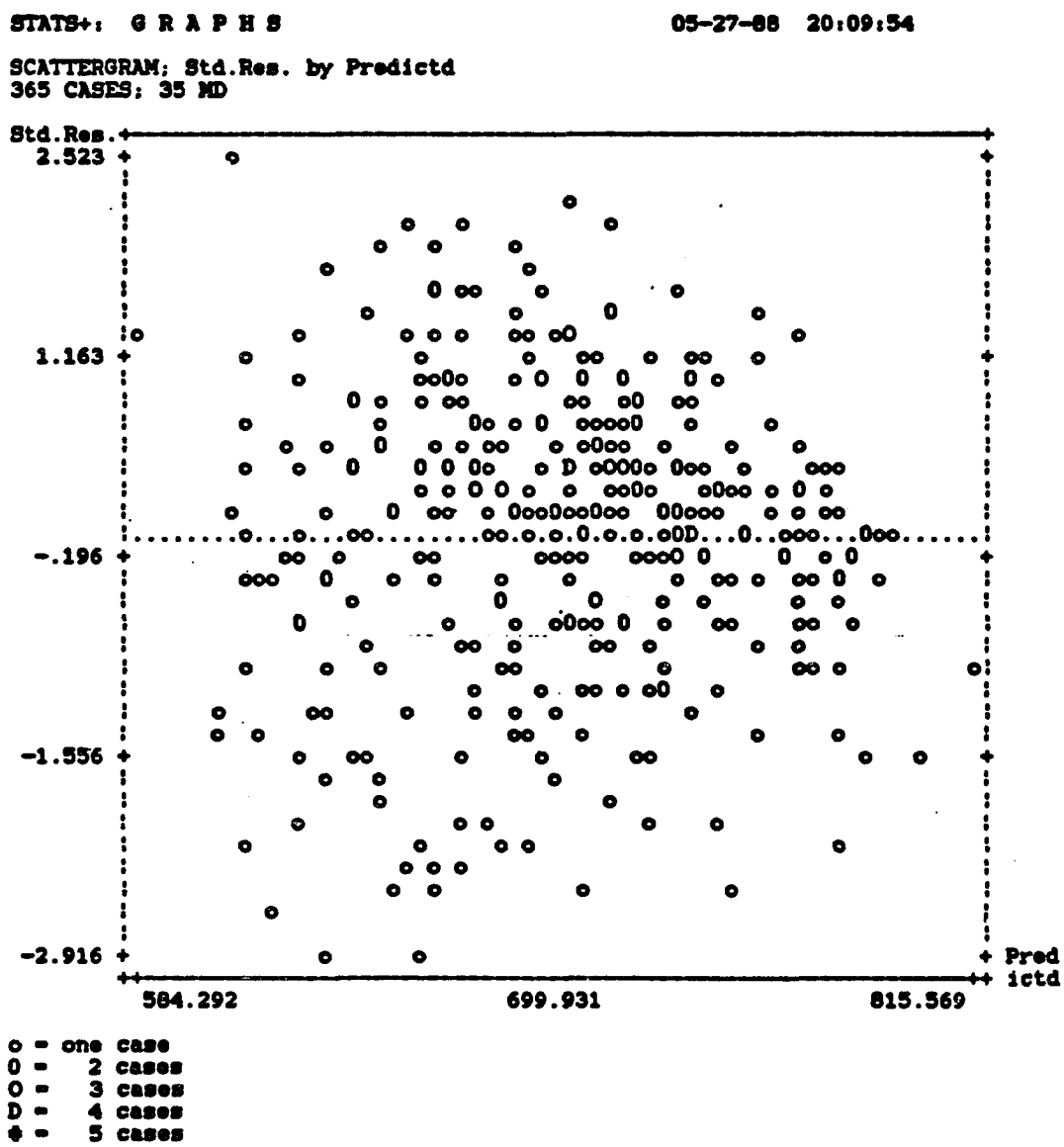
Standardized Residuals by Predicted CAT Math



Predicted Scores by CAT Reading Posttest



Standardized Residuals by Predicted CAT Reading



Multiple Regression Results Predicting CAT Total

CSS/PC: MULTIPLE REGRESSION 06-04-88 09:37:11

Basic Statistics from 53 Elementary Teachers, Spring 1988

Multiple Regression Results:

Variables were entered in one block

Dependent Variable: CATt
Multiple R: .9356159
Multiple R-Square: .8753771
Adjusted R-Square: .8611345
Minimum pairwise N: 40
F (4, 35) = 61.46184 p < .0000
Intercept: 580.617878

css/pc multiple regress. — REGRESSION WEIGHTS

variable	BETA	St. Err. of BETA	B	St. Err. of B	t (35)	Signif. of t	Valid N	Valid N Pairwise
Grade	.89433	.06047	23.10881	1.56261	14.78857	.00000	40	40
Mean_IQ	.03214	.07367	.21680	.49691	.43631	.66855	40	40
Absences	-.19621	.06832	-2.16634	.75428	-2.87206	.00689	40	40
SSex	.05356	.06566	11.82852	14.50116	.81569	.42550	40	40

css/pc multiple regress. — VARIABLES NOT IN THE EQUATION

Variable	Beta in	Partial Cor.	Semipart Cor.	Tolerance	Minimum Tolerance	t	Signif. of t	Valid N	Valid N Pairwise
CATm	.63511	.70275	.24808	.15258	.15258	5.75973	.00002	40	40
CATr	.66432	.66712	.23551	.12567	.12567	5.22170	.00005	40	40
SRace	-.04288	-.11197	-.03953	.84993	.61527	-.65702	.52230	40	40
Freeluch	-.07604	-.19689	-.06950	.83554	.63992	-1.17095	.24819	40	40
Parents	-.04412	-.12413	-.04382	.98659	.65333	-.72942	.47724	40	40
P1	.08969	.23743	.08382	.87329	.64301	1.42521	.15961	40	40
P2	.09574	.24898	.08789	.84289	.64639	1.49898	.13927	40	40
P3	.08856	.23288	.08221	.86174	.65469	1.39631	.16820	40	40
P4	.11594	.30904	.10910	.88545	.65596	1.89475	.06324	40	40
P5	.07123	.18838	.06650	.87175	.65460	1.11848	.27035	40	40
P6	.02077	.05324	.01880	.81881	.63720	.31089	.75201	39	39
P7	.07700	.20368	.07190	.87194	.65197	1.21307	.23140	40	40
P8	.14924	.39070	.13792	.85415	.65151	2.47486	.01738	40	40
TPAItot1	.10786	.28121	.09927	.84716	.64876	1.70869	.09274	39	39
TPAI_1_5	.10636	.27625	.09752	.84065	.65135	1.67602	.09898	40	40
TPAI_6_8	.09755	.25268	.08920	.83619	.64619	1.52276	.13318	39	39

Multiple Regression Results Predicting CAT Math

CSS/PC: MULTIPLE REGRESSION 06-04-88 09:42:08

Basic Statistics from 53 Elementary Teachers, Spring 1988

Multiple Regression Results:

Variables were entered in one block

Dependent Variable: CATm
Multiple R: .9185414
Multiple R-Square: .8437183
Adjusted R-Square: .8306949
Minimum pairwise N: 40
F (3, 36) = 64.78443 p < .0000
Intercept: 478.395979

css/pc multiple regress. — REGRESSION WEIGHTS

variable	BETA	St. Err. of BETA	B	St. Err. of B	t (36)	Signif. of t	Valid N	Valid N Pairwise
Grade	.87223	.06598	27.55760	2.08448	13.22037	.00000	40	40
Mean_IQ	.15319	.07525	1.26337	.62062	2.03564	.04654	40	40
Absences	-.13315	.07527	-1.79757	1.01617	-1.76896	.08185	40	40

css/pc multiple regress. — VARIABLES NOT IN THE EQUATION

Variable	Beta in	Partial Cor.	Semipart Cor.	Tolerance	Minimum Tolerance	t	Signif. of t	Valid N	Valid N Pairwise
CATt	.78640	.70889	.28024	.12699	.12699	5.94604	.00002	40	40
CATr	.54834	.49197	.19449	.12580	.12580	3.34305	.00228	40	40
SRace	.05909	.14178	.05605	.89977	.74412	.84733	.40707	40	40
SSex	.06696	.15393	.06085	.82583	.65602	.92165	.36588	40	40
Freeluch	-.01909	-.04455	-.01761	.85085	.75660	-.26380	.78254	40	40
Parents	-.07505	-.18862	-.07457	.98719	.76177	-1.13629	.26244	40	40
P1	.17881	.42716	.16887	.89193	.73468	2.79495	.00816	40	40
P2	.15554	.36466	.14416	.85895	.72118	2.31689	.02481	40	40
P3	.17660	.41483	.16399	.86234	.73221	2.69722	.01026	40	40
P4	.17019	.40918	.16176	.90338	.75000	2.65300	.01138	40	40
P5	.15952	.38280	.15133	.89997	.75293	2.45142	.01821	40	40
P6	.15361	.35958	.14215	.85643	.68111	2.27981	.02780	39	39
P7	.13442	.32751	.12947	.92775	.72650	2.05068	.04504	40	40
P8	.19774	.47218	.18666	.89107	.72141	3.16895	.00340	40	40
TPAItot1	.20476	.48448	.19153	.87495	.71933	3.27640	.00265	39	39
TPAI_1_5	.19211	.45020	.17797	.85821	.73155	2.98276	.00525	40	40
TPAI_6_8	.18837	.44857	.17733	.88625	.70262	2.96924	.00542	39	39

Multiple Regression Results Predicting CAT Reading

CSS/PC: MULTIPLE REGRESSION

Basic Statis***** from 53 Elementary Teachers, Spring 1988

Multiple Regression Results:

Variables were entered in one block

Dependent Variable: CATr
Multiple R: .9350545
Multiple R-Square: .8743270
Adjusted R-Square: .8599644
Minimum pairwise N: 40
F (4, 35) = 60.87514 p < .0000
Intercept: 531.464783

css/pc multiple regress.	REGRESSION WEIGHTS							
Variable	BETA	St. Err. of BETA	B	St. Err. of B	t (35)	Signif. of t	Valid N	Valid N Pairwise
Grade	.92272	.06073	30.23202	1.98973	15.19400	.00000	40	40
Mean_IQ	.04671	.07398	.39952	.63273	.63142	.53871	40	40
Absences	-.08737	.06860	-1.22321	.96046	-1.27357	.20879	40	40
SSex	-.01226	.06594	-3.43335	18.46488	-.18594	.83172	40	40

css/pc multiple regress.	VARIABLES NOT IN THE EQUATION							
Variable	Beta in	Partial Cor.	Semipart Cor.	Tolerance	Minimum Tolerance	t	Signif. of t	Valid N	Valid N Pairwise
CATt	.66992	.66712	.23650	.12462	.12462	5.22170	.00005	40	40
CATm	.45654	.50304	.17833	.15258	.15258	3.39390	.00206	40	40
SRace	.00892	.02321	.00823	.84993	.61527	.13537	.86260	40	40
Freelnch	-.02591	-.06681	-.02368	.83554	.63992	-.39042	.69933	40	40
Parents	-.02452	-.06871	-.02436	.98659	.65333	-.40160	.69185	40	40
P1	.06719	.17711	.06279	.87329	.64301	1.04932	.30171	40	40
P2	.06887	.17835	.06323	.84289	.64639	1.05691	.29814	40	40
P3	.03736	.09782	.03468	.86174	.65469	.57313	.57686	40	40
P4	.07098	.18840	.06679	.88545	.65596	1.11857	.27[illegible]	40	40
P5	.02637	.06946	.02462	.87175	.65460	.40601	.68889	40	40
P6	.02314	.05907	.02094	.81881	.63720	.34503	.72953	39	39
P7	.07107	.18719	.06636	.87194	.65197	1.11115	.27355	40	40
P8	.05815	.15160	.05374	.85415	.65151	.89429	.38091	40	40
TPAItotl	.06448	.16740	.05934	.84716	.64876	.99008	.33053	39	39
TPAI_1_5	.06314	.16329	.05789	.84065	.65135	.96510	.34322	40	40
TPAI_6_8	.05837	.15056	.05338	.83619	.64619	.88805	.38[illegible]	39	39

Multiple Regression Results Predicting CAT Math Student Sex Forced into Equation

CSS/PC: MULTIPLE REGRESSION 06-04-88 09:44:01

Basic Statistics from 53 Elementary Teachers, Spring 1988

Multiple Regression Results:

Variables were entered in one block

Dependent Variable: CATm
Multiple R: .9205549
Multiple R-Square: .8474213
Adjusted R-Square: .8299838
Minimum pairwise N: 40
F (4, 35) = 48.59748 p < .0000
Intercept: 446.879737

css/pc multiple regress. — REGRESSION WEIGHTS

Variable	BETA	St. Err. of BETA	B	St. Err. of B	t (35)	Signif. of t	Valid N	Valid N Pairwise
Grade	.86272	.06691	27.25713	2.11414	12.89277	.00000	40	40
Mean_IQ	.18172	.08152	1.49869	.67230	2.22921	.03046	40	40
Absences	-.12858	.07559	-1.73579	1.02051	-1.70091	.09419	40	40
SSex	.06696	.07266	18.08224	19.61939	.92165	.36605	40	40

css/pc multiple regress. — VARIABLES NOT IN THE EQUATION

Variable	Beta in	Partial Cor.	Semipart Cor.	Tolerance	Minimum Tolerance	t	Signif. of t	Valid N	Valid N Pairwise
CATt	.77759	.70275	.27450	.12462	.12462	5.75973	.00002	40	40
CATr	.55428	.50304	.19649	.12567	.12567	3.39390	.00206	40	40
SRace	.04657	.16991	.04293	.84993	.61527	.64478	.53012	40	40
Freeluch	-.01043	-.02441	-.00954	.83554	.63992	-.14237	.85837	40	40
Parents	-.07358	-.18711	-.07309	.98659	.65333	-1.11064	.27378	40	40
P1	.19213	.45966	.17955	.87329	.64301	3.01799	.00490	40	40
P2	.16766	.39407	.15393	.84289	.64639	2.50008	.01640	40	40
P3	.17845	.42408	.16565	.85174	.65469	2.73050	.00959	40	40
P4	.18284	.44046	.17205	.88545	.65596	2.86072	.00707	40	40
P5	.17641	.42167	.16471	.87175	.65460	2.71157	.01002	40	40
P6	.17507	.40557	.15842	.81881	.63720	2.58722	.01339	39	39
P7	.15951	.38132	.14895	.87194	.65197	2.40519	.02041	40	40
P8	.21998	.52048	.20331	.85415	.65151	3.55428	.00144	40	40
TPAItotl	.22345	.52652	.20566	.84716	.64876	3.61116	.00127	39	39
TPAI_1_5	.20572	.48287	.18862	.84065	.65135	3.21528	.00310	40	40
TPAI_6_8	.21593	.50549	.19745	.83619	.64619	3.41606	.00196	39	39