THE EFFECT OF TRAINING IN TEST ITEM WRITING ON TEST PERFORMANCE OF JUNIOR HIGH STUDENTS

DISSERTATION

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

For the Degree of

DOCTOR OF PHILOSOPHY

By

Jeanne L. Tunks, B.M.E., M.M.T.

Denton, Texas

May, 1997

Tunks, Jeanne L., The effect of training in test item writing on test performance of junior high students. Doctor of Philosophy (Curriculum and Instruction), May 1997, 137 pp., 25 tables, references, 111 titles.

Students in an inner city junior high school in North Central Texas participated in a study whose purpose was to examine the effect of training in test item construction on their later test performance. Students were randomly placed in experimental and control groups. The experimental group underwent twelve weeks of instruction using the Test Item Construction Method (TICM). In these sessions students learned to develop test items similar to those on which they were tested annually by the state via the Texas Assessment of Academic Skills (TAAS). The TICM aligned with state mandated test specifications.

The data for the study were TAAS diagnostic exams administered by the district in September, 1996 and January, 1997. The January data served as the dependent variable along with the scores from the Test Anxiety Inventory. The September data served as the covariate in the analysis. The independent variable was group membership. A multivariate analysis of covariance (MANCOVA) was applied to the data and no significant differences were observed on any measures. However, the covariates did significantly reduce error variance in a number of dependent variables. A content analysis of student portfolios maintained during treatment yielded information about changes students underwent during treatment regarding their knowledge about testing.

THE EFFECT OF TRAINING IN TEST ITEM WRITING ON TEST PERFORMANCE OF JUNIOR HIGH STUDENTS

DISSERTATION

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

For the Degree of

DOCTOR OF PHILOSOPHY

By

Jeanne L. Tunks, B.M.E., M.M.T.

Denton, Texas

May, 1997

ACKNOWLEDGMENTS

The author wishes to acknowledge and thank her husband, Tom Tunks, whose guidance, support, and love made this project possible. This completed dissertation is dedicated to him.

The author wishes to thank her two sons, Adam and Gordon Tunks, for their understanding and patience as they witnessed their mother as a student.

The author wishes to thank her committee members Dr. James Laney, Dr. Fred Thomas, and Dr. Ronald Wilhelm for their tireless hours of instruction, suggestions for improvement, and willingness to work with the author through the painstaking process of writing a paper of this magnitude.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Introduction

Mandated testing, a fairly new phenomenon in United States educational history, began in the scholarly structuralist era of school reform in the 1960s. In 1959, Russia successfully launched a manned space capsule, unleashing a wave of concern among political leaders in the United States over the control of world power. This matter became the impetus behind the United States government's involvement in national standards in education, particularly science. With the advent of this pressure for increased achievement in science education, numerous symposia and science study committees were formed to ascertain the most efficient manner to change the way American public schools taught children.

This effort brought scholars from across the nation to work as teams to create a set of national standards in teaching and learning. The Woods Hole Symposium, an example of one of the projects, emphasized the need for discovery learning. The conferees concluded that to transfer standards to teachers and subsequently to students, teachers should be supplied with materials that were simple and matched the level of the student (Bruner, 1961). Although there were many other pertinent conclusions reached by the symposium participants, such as the importance of structure, readiness of learning, the act of learning, spiral curriculum, intuitive and analytical thinking, and motives for learning, the simplification of materials and curriculum for teachers and students remains as the

hallmark of the work of this team. Teacher editions of basal texts, published nationally, bear the manifestation of the adoption of simplification (Silver Burdett and Ginn, 1989; Harcourt, 1989; MacMillan, 1989).

Prior to the drive for accountability, a characteristic of the era, testing was done on a local level. Flannagan, in 1939, encouraged the use of criterion-referenced tests in the classroom. These tests generated scores that have some absolute meaning as opposed to group norming (Shaycott, 1979). These tests were structured to match instructional presentations and were administered by the local classroom teacher. Students were judged on their own merit, with their prior achievement as their standard of success. New standards in learning dictated new standards in testing. Consequently, in the 1960s, the nation and subsequently the states began mandating standardized testing on a state and national level.

As of 1992, forty-four of the fifty states in the United States mandated some form of standardized testing (Webster, 1995). The mandates of the 1990s have an attachment that distinguishes them from the mandates of the 1960s, meritocracy. In the past ten years, states and school districts nationwide award merit pay to teachers who can raise standardized test scores in their districts. This most recent occurrence in testing has fostered the use of the term "high stakes testing".

Statement of the Problem

Students and teachers in high-stakes testing environments experience intense pressure to raise scores to ever higher levels at all costs. The practice of preparing students for externally mandated tests results in minimizing the curriculum, reduces thinking expectations of students, and places students at risk for success in the job markets of the future. This study examined the procedures used to prepare students for mandated tests. Specifically, the proposed study sought to determine whether students could benefit, as

measured by achievement scores on the TAAS test, from an alternative preparation strategy; teaching students how to develop TAAS test items.

The alternate test preparation strategy, Test Item Construction Method (TICM), based in constructivist principles and generative learning techniques, fostered greater understanding of testing expectations. The strategy focused on comprehending test items through the construction of such items. This method contrasted with current methodology of confronting items and used testwiseness strategies that promoted short-term thinking for short-term gain. Through a constructivist, generative learning approach, students delved into the particulars of items, their construction, and purpose. The deeper understanding, manifested through test item construction, provided students with a central perspective of items, as opposed to an external perspective, as experienced during testwiseness preparation. The central, or core experience, provided the learner with the tools to approach testing with confidence.

The generative learning model (GLM), from which the TICM formulates, has as its basis, Aristotelian philosophy. Aristotle contended that all humans desire to know and understand the things they encounter. According to Aristotle, purposefulness, the principle cause, the center of understanding, provides the learner with the path for comprehending the aspects or causes that lead to the purpose. These aspects include the form or shape the thing takes, the workmanship required to construct the thing, and the materials necessary. From these four causes, humans define the things of man and nature. These Aristotelian perspectives provide insight into a deeper meaning of things with which humans are confronted. Aristotle's tenets, applied to GLM, provide a position for construction and instruction.

TICM, by extension, embodied the convictions of that philosophy. The purpose of the TICM, guided students in the practice of constructing test items and eventually provided the basis of success in test-taking on like items, established the parameters of the other

aspects of the phenomenon. By virtue of the purpose, the form of the method assumed the quality of constructivism or generation. The efficiency or manner of workmanship dictated several models of instruction that included inquiry, inductive, and mastery. Finally, test items and their component parts, became the materials by which the purpose of test item construction was fulfilled. The establishment of the connectedness of test item construction, provided the learner with needed understanding to approach testing with renewed appreciation, and a sense of positiveness toward testing. The problem under investigation examined whether that renewal transformed into increased achievement on like test items.

The current status of mandated testing places the learner in a precarious position of balancing the pressure of achievement with the need to know and understand. Anders and Richardson (1992) point out that the effects of high-stakes testing result in anxiety about testing and, in general, testing plays a negative role in schools resulting in the confusion by the teachers as to the nature of tests. Smith and Rottenberg (1991) noted that teachers tend to neglect material not tested and that external testing encourages teachers to use instructional methods that resemble tests, worksheets. Because of the negative effects of testing, Smith and Rottenberg encourage the use of a constructivist approach to instruction. Shepard (1991) found when test results are subject to political pressure and media attention, scores can become inflated, thereby giving a false impression of student achievement. Shepard suggested a conversion from old theory of measurement driven instruction to new theories that encourage higher levels of thinking and active construction.

Constructivism, a theory about knowledge and learning (Brooks and Brooks, 1993), draws on a combination areas of study including cognitive psychology, philosophy, and anthropology. The constructivist's perspective of knowledge labels knowledge as "temporary, developmental, socially and culturally mediated, and thus non-objective" (Fostnot, p., vii). Learning is generated or constructed by the learner and therefore self-

regulated. Through the process the student explores and reconciles inner conflicts resulting from concrete experiences, cooperative interaction, and mindful consideration. Brooks and Brooks provide five inherent principles of constructivist pedagogy. These include posing emergent relevant problems, structuring learning to emphasize larger concepts, extolling the merits of student input, adjusting curriculum to accommodate student perceptions, and assessing student learning within the context of teaching.

Wittrock's generative teaching model (1991), designed to enhance learning through generative processing, consists of the components necessary for constructivist learning. The premise of generative processing, based in Aristotelian philosophy, maintains that learners gain understanding through a progressive development from sensation, memory, and past experience which the learner then combines to create meaning in the form of art and eventually science (McKeon, 1947). The teaching component of the model embodies four parts; student knowledge base and preconceptions, motivation, attention, and generation. Actual understanding occurs when the student generates relationships between past experiences, stored knowledge, and new information presented during instruction. Real life situations presented during instruction, with clarity and demonstration, provide the learner with tools needed to generate new models of the phenomenon under consideration.

The generation of new understanding, as related to prior learning, involves transfer. Fogarty, Perkins, and Barell (1992) define transfer as "learning something in one context and applying it in another" (pp. ix). Transferable content bridged across higher order thinking applications result in newly constructed paradigms. Within a content area of study, techniques such as matching, modeling, setting expectations, simulating, and problem based learning (low road means), result in near transfer. However, far transfer across contents occurs through the application of high road transfer techniques such as anticipating applications, generalizing concepts, using analogies, parallel problem solving, and metacognitive reflection. Wittrock (1991), Kourilsky and Wittrock (1992), Harlen

and Osborne (1983), Laney (1990), Wittrock and Alesandrini (1990) applied high road transfer techniques in experimental research in generative learning with the result of higher gains among students engaged in high road transfer.

The method of instruction, TICM, encompassed the procedures employed in the generative learning model. These techniques provided students with opportunities to examine items based on their prior knowledge, with the added sanction and encouragement to redesign and construct the items to suit their level of understanding. The higher level thinking required to complete the particulars of the method heightened the students' awareness and cognizance of test items. This elevated perspective enabled the student to negotiate test items with greater success and less trepidation.

Purpose of the Study

The purpose of the proposed study was to examine the adequacy of an alternative instructional practice in the preparation of students for mandated state tests. A comparison was made between the scores of students who prepared for tests with an alternative method and of those who prepared using the traditional method of preparation. Alternative test preparation practice was defined as practice that existed outside the realm of behavioral theory that promotes traditional drill and practice, testwiseness preparation, and test-taking trials. The alternative practice, employed in this study (TICM), was grounded in active constructivism, as represented in the theory of generative learning (Wittrock, 1975).

Generative learning theory (GLT) employs neuropsychological principles of neural and cognitive processing. Three functional brain systems involving arousal and attention, presented by Luria (1973), provide the neurological basis for GLT. In the first system, processes are influenced by the cortex and by conceptually-driven behavior through the descending reticular activation systems. "The plans and intentions of the learner, which are mediated by the frontal lobes of the cortex, influence the attentional and motivational

processes of the brain, and consequently, the stimuli we attend to and the level of activity we devote to those stimuli and their meaning." (p. 5) Attention and motivation, key components of GLT, are grounded in the selection of stimuli and the management of the process of constructing meaning.

The second and third functions deal with the generation of knowledge. In the second functional brain unit, the learner receives, analyzes, and stores information while coding these through integrated senses. Generative learning results from meaning that is constructed in varied ways by synthesizing stimuli into consistent sequences and patterns reflective of the learner's knowledge and experience. The third function uses the frontal lobes for planning, organizing, and regulating cognition and behavior, essentially functioning as a generative processor. Metacognitive activities reflect higher order processing that generates the construction of meaning and provides integration. These generative processes coordinate attention, arousal, motivation, and learning.

In this processing, according to Wittrock (1992), learners generate meaning and understanding from instruction. The learner draws from his or her own experience to create or construct meaning from teaching and learning. The focus of GLT is on generating relations between prior experiences and new information across concepts. In this process, the learner constructs his/her own understanding and therefore retains information for longer periods of time (Wittrock, 1991; Harlen, 1983; Wittrock and Carter, 1975; Kourilsky and Wittrock, 1992; Wittrock and Alesandrini, 1990).

The generative learning model of instruction finds its roots in constructivism and transfer of learning. According to Brooks (1992), learners construct their own understanding of new information with which they come in contact. Learners generate understanding through the cognitive process of relating prior knowledge and experiences with current challenges. In the process, learners find or develop mental and physical tools to generate understanding of new phenomena.

In extending theory to practice, Wittrock (1977) developed a functional model of instruction that can be applied in educational settings. This model of instruction is based on four cognitive processes that include learning, motivation, knowledge creation, and generation. Essential to the learning process is attention. Learners must be engaged and at a readiness level to learn. The second process, motivation, is based on student interests and attributes. The contention is that students are encouraged to participate in learning if the students' interests and characteristics are considered.

The final two processes involve construction of understanding. In the third process, knowledge creation, learner preconceptions, concepts and beliefs are included in the instructional practice. In this process, students relate their own experiences in the form of tenets. The final process involves the generation of new relations between tenets and new knowledge and skills. Generation takes on the form of high road transfer (Perkins, 1992) in which the learner uses analogies, metaphors, metacognition, and summaries to create meaning from learning. In high road transfer, the learner uses mindful abstraction to establish relationships between prior knowledge and experience and new knowledge and expectations.

In this study, the researcher applied the principles of generative learning theory and the generative model of instruction in a test preparation setting. Students were instructed in the practice of creating test items for state-mandated tests. Students' interests, beliefs, and attributes were considered when constructing the instructional sets. Analogies, metaphors, and summaries were used to promote understanding of what test item writing entails. Students were involved in an evolutionary process over twelve weeks. During this time material created was stored in portfolios for easy reference for students as they generated meaning and understanding about the process of developing test items.

In order to establish a basis for comparison of instructional methods, a preliminary qualitative study was conducted to identify the teachers' current test preparation practices.

Teachers were observed as they prepared students for the TAAS test. Teachers were interviewed regarding their selection of strategies and beliefs about mandated testing. Students in these classes were interviewed regarding their beliefs about the testing procedure. Information from the three sources were triangulated to determine perceptions and conceptions about mandated testing in the site where the study was conducted. Inferences from the triangulation were drawn and meanings extrapolated. These findings served as guidelines in developing instructional lessons for the final study.

Research Hypotheses

Several hypotheses were considered in this study. It was hypothesized that:

1) Students receiving alternative test preparation practice (TICM) would score higher on mandated state tests (TAAS) in the following areas:

a) TAAS total score means

b) TAAS reading subtest score means

c) TAAS reading mastery means

2) Students receiving alternative test preparation practice (TICM) would demonstrate a lower level of anxiety toward testing, as measured by the *Test Anxiety Inventory*, when compared to students who received current practice in test preparation.

Areas of Concern

There were several concerns manifested in mandated testing practices. Among these was the trivialization of instruction that results from teaching to the test. External pressure to increase test scores results in instruction directed toward test items and testing styles (Herman and Golan, 1990, 1993). Shepard (1991) notes that in high-stakes testing situations, instruction is misdirected resulting in a proliferation of tasks that resemble tests. In addition, skills are taught in isolation. The drill and practice type of instruction is

supported by behavior theories of instruction. These theories have been challenged by theories of constructivism and metacognition (Wittrock, 1990; Brooks, 1993)

A second concern was the reduction of the curriculum. External pressure from the media, state departments of education, districts, and the community for higher test scores has forced teachers to rethink their curriculum to accommodate the demand (Madaus, 1989;Mehrens, 1986; Herman, 1993; Neill, 1989). Teachers focus on the parts of the curriculum targeted for testing, thereby disregarding other meaningful parts of the curriculum (Bracy, 1987).

A third concern was for the use of tests for purposes other than diagnosing learner needs, prescribing instructional activities, and measuring student progress in curriculum content (Norman, 1980). Test scores were published in newspapers throughout the United States (Shepard, 1991). In some cities, the schools were ranked by the district and the rankings are published in the newspapers (Dallas Morning News, 1995). The purpose of these publications remained unclear. However, the effects of the print on teachers were generally negative and invasive (Shepard 1991). Another use for test scores was in the real estate industry. Real estate firms publish test results for their sales personnel based on the requests of incoming buyers who want to know the test scores of districts in the area to which they transfer (Martin, 1996). The aberrant use of test scores took their use well beyond the purpose intended.

A final most important concern was for the learner who was expected to comply with the mandates of testing in a high-stakes setting. This matter had numerous consequences that included a reduction of thinking expectations of students, no real gains in learning, minimal knowledge and skill gains, tracking, reduction in problem solving skills, increase in drop-outs, increased grade retentions, increased student anxiety, growing disillusionment about testing, increased use of inappropriate strategies, and decreasing motivation (Madaus, 1989; Shepard, 1990, 1991; Neill, 1989; Paris, 1991). The losses to

students in a high-stakes testing environment extended beyond the classroom or the newspaper and into their futures as contributors to the community.

Basis of Concern

The Secretary's Commission on Achieving Necessary Skills (SCANS Report, 1992) delineates its perception of the required skills for work beyond the year 2000. The report, written with schools as the focus, emphasizes the skills that are being denied students in high-stakes testing environments. There are five areas targeted by the SCANS Report, including resources, interpersonal, information, systems, and technology. Students graduating from schools in America will be expected to identify, organize, plan, and allocate resources of time, money, materials and facilities, and human resources. Interpersonally, they work together with others on teams while teaching others' skills, exercise leadership, and demonstrate an understanding of diversity as exhibited in their ability to work with others different from themselves.

An expansion of interpersonal learning includes systems learning and information gathering. In this competency, students are expected to understand complex social, organizational, and technological systems. They will need to know how to monitor and correct the performance of each of these and improve the design. Students of today were born into the information age and need to be able to access, evaluate, organize, interpret and communicate using the latest technology. To that end, students must be competent in a variety of technologies and be able to select, apply, and adjust as technology changes.

General Information and Background on Testing

Standardized Tests and Criterion Referenced Tests

Standardized tests differ from the criterion-referenced tests of the pre-1960s era in several ways. These tests are designed to compare students being tested to norm groups of students who have taken the test previously. In essence, a testing company, such as the Education Testing Service, develops tests of multiple-choice items that are, by national educational standards, a test of general knowledge and skills. This test is administered to randomly selected groups of students who represent a normal segment of the population. The resulting scores are then standardized using a statistical process that results in a mean (average) for the group and a standard deviation (how far the group varies from the average). These measures of central tendency and variability are then used to compare scores of subsequent test respondents to the normal tendencies of the initial group. The scores are then reported in percentiles that indicate how well the respondent did in relation to the average of the group. Consequently, the standardized tests are considered norm-referenced tests because of the reference of the respondent's score to the original normed group.

Standardized tests comprise primarily four choice multiple-choice items. In these items there is a stem (statement of a problem to solve) and four choices from which to select in answering the problem to solve, one of which is the correct answer. It is the job of the respondent to select the correct answer. The questions on any given norm-referenced test are samples of information and ideas assumed as having been presented in an instructional setting or a part of a real-life experience of the learner. Test constructors assume that learners with excellent general knowledge succeed easily.

Testwiseness

There is, however, an additional challenge to the test taker known as test foils or distracters (Gronlund, 1993). Distracters serve to examine testwiseness. Testwiseness, according to Jones (1981), is "a subject's capacity to utilize the characteristics and formats of the test and or the test-taking situations to receive a high score. Testwiseness is logically independent of the examinee's knowledge of the subject matter for which the items are measured. Testwiseness affects scores on standardized tests and variance, not due to the trait the test seeks to measure." (p.3) Seaton (1992) and Diamond (1972) concur with the point Jones makes regarding independence. Dolly (1986) defines testwiseness as the "cognitive strategies used to maximize chances of selecting the correct option when the test taker is forced to guess due to lack of information concerning the content of the item." (p. 619) Smith (1982) points out further that testwiseness is only useful when there are flaws in the construction of the test.

Preparation for standardized test taking was not a part of the educational fabric of instruction during the post-Sputnik era known as "scholarly structuralism". The materials developed by the various committees and symposia supposedly supplied students with all they needed to succeed in tests on a national level. Unfortunately, an examination of the trends in standardized testing in the 1970s revealed a negative trend. In a historical survey of the state of standardized testing, Norman (1980) delineates the findings of the National Education Association (NEA) and its ensuing recommendations.

NEA intervention

In 1973, NEA noted a decline in scores on standardized tests. Consequently, they proposed resolution 72-74 on standardized testing that encouraged the elimination of standardized testing of intelligence, aptitude, and achievement. A 1977 report, *On Further Examination*, examined the reason for the drop in scores and related the drop to the change

in social climate. The social climate in reference is the effort by the federal government to integrate the schools. In reports that followed in 1977, the tests themselves were questioned by the overseeing committee. Resolution 78-82, issued in 1978, stated that "student testing could serve important education purposes such as diagnosing learner needs, prescribing instructional activities, and measuring student progress in curriculum content." (p. 49) The commission, in testimony before Congress in 1979 opposed tests that are damaging to student self-concept, biased, used for tracking, invalid, unreliable, out of date, used by the media, used to test performance levels of graduation, and used to evaluate teachers. In the Buros reform, NEA recommended that testing be done for measurement, not differentiation.

These recommendations, although blatantly overlooked in the 1980s and 1990s, remained in the minds of reformers in the 80s and 90s as valuable and meaningful (Bracy, 1989). Testing took on an entirely different tenor from the one intended by the initial sculptors of education reform. Mandated testing of the 1960s transformed into an aberration in the 1980s and 90s. Students' standardized test scores have become political and social tools for comparing states, school districts, and neighborhood communities. Corporate mobility, a common phenomenon of the 80s and 90s, sent parents in search of the "ideal" school. In a metroplex in North Texas, eager real estate brokers and agents keep abreast of current standardized test scores and media rankings of area school districts. The information provides a service to adults of school-age children seeking what they perceive as the appropriate school district (Martin, 1996).

High-stakes testing

High-stakes testing, as a phenomenon, grew out of the American taxpayers' need to know that their tax dollars were being spent wisely on public education. Communities around the United States, led by media hype regarding the deterioration of the American

school system, demanded fiscal accountability for education tax dollars spent. Consequently, scores from high-stakes tests provided the necessary data to answer the charge of fiduciary irresponsibility (Dangler, 1994). Smith (1991) defines high-stakes testing as testing which has "serious consequences or trigger actions contributed to selection, merit evaluations, promotion, retention, and take-over decisions." (p.7) One common characteristic of high-stakes testing is media involvement. Published scores provide districts and communities with comparisons (Anders, 1992). In a high-stakes testing situation, teachers bear the burden of responsibility for raising test scores at any cost.

The expense of raising scores extends to numerous areas of the schooling process. Anders (1992), following fifteen months of intense observation, interviews, and surveys, noted that the price for success can result in high anxiety about scores being reported by the media, an overall distrust among the teachers as evaluators, and confusion among teachers about the nature of testing. Smith (1991) found that some of the unintended consequences are the neglect of particular important materials not included in testing, narrowing of the curriculum, and an inordinate amount of time spent preparing students for "the test." Mehrens (1989) observed that, with an increase in judgment of schools and individual teachers based on student scores, some teachers find inappropriate ways to prepare students for the tests.

Test Preparation Practice

Test preparation practice falls into several categories along a continuum of acceptable to unacceptable practice (Mehrens, 1989). Among the acceptable practices are general instruction, specifically aligned with objectives tested, and testwiseness skill development. Marginally acceptable are instruction on objectives developed by commercial groups, instruction based on objectives specifically matched to the test objectives, and

instruction based on specifically matched objectives and formatted identically to the test. Unacceptable practices include practicing on published parallel tests and practice and instruction on the same test. Ligon (1981) defines appropriate preparation activity as "that which contributes to students' performance on the test near their true achievement levels and which contributes to their scores that would require an equal amount of regular classroom instruction." (p.2)

The pressure of producing has compelled some teachers and administrators to cross the line of unacceptability. Canner (1992) documented cases of school personnel stealing copies of the test, giving students answers in advance, allowing students to practice on identical tests, and in some cases practice on the actual test. These extremes in behavior are not widespread but do exist. They are a sign of the circumvention schemes to which school personnel will resort to survive the pressure. Another type of unacceptable activity involves reporting scores. Cannell (1988), in a survey of all fifty states, has found that all states reported that their students were above the 50th percentile on standardized, normed commercial tests. If the tests are truly normed, fifty percent of the students taking the tests would fall below the fiftieth percentile. The term "the Lake Wobegon Effect" grew out of Cannell's findings. The Lake Wobegon Effect refers to a line by Garrison Keillor on the radio show "A Prairie Home Companion", when he describes the children in Lake Wobegon, his hometown, who are "all above average, every one."

The practice of preparing students for mandated tests tends toward the use of ethically acceptable strategies. The most prevalently used method in preparing students for mandated tests is testwiseness preparation. Testwiseness consists primarily of providing students with tools needed to navigate through the wording, structure, physical requirements, and expectations of external tests (Phillips, 1983; Hall, 1990; Palmer, 1984; Ducote, 1982; Ligon, 1981; Seton, 1992). The tools (strategies) most commonly provided the student include looking for clues, familiarity with format, process of elimination, time

planning, physical and emotional preparedness, direction following, knowing the type of item, grammatical relationship between the stem and the answer, carefulness, and knowing the type of item that is being presented (Summers, 1983; Montgomery County Board of Education, 1982; Phillips, 1983; Ligon, 1981; Anderson, 1981; Diamond, 1972;Bangert-Drowns, 1983; Guifford, 1980; Kilian, 1992;George, 1985; Mehrens and Kaminski, 1989). The number of strategies mentioned exceeds twenty-five.

Test Preparation Tools

The proliferation of tools has an inverse relationship to the success of students in raising their scores. According to Cushing (1989), no differences were observed in matched classes of fourth and fifth graders following training in testwiseness. Flippo and Borthwick (1981) found similar results among elementary students trained in additional testwiseness strategies as compared to students who received extra art and library time. Scruggs, et al (1986) reviewed forty-four studies comparing test preparation with achievement and found that training in test-taking skills has limited effects on achievement scores, even though there were some differences noted among some subgroups. In contrast, Seaton (1992) detected in a meta-analysis that all subgroups can benefit from test preparation seminars, although he questioned the value of coaching for a single test.

Coaching for tests differs from instructing for tests. Coaching is defined by George (1985) as "short term intensive training for a test... (p. 24)." Anderson (1981) examines coaching and instruction from the perspective of differences between the two on purpose, techniques, and duration. Coaching is short-term, drill and practice, whose purpose is to improve performance on a particular test. In comparison, instruction is generally long-term and provides alternative explanation of phenomenon and solutions to problems. Its purpose is to develop skills, knowledge, and understanding. Bangert-Drowns (1983) found that coaching raised standardized test scores by .25 standard

deviation, essentially raising a student's score from the 50th to the 60th percentile. Unfortunately, according to Anderson (1981), the retention is generally poor when coaching is used when contrasted to higher levels of retention observed as a result of instruction.

Jones and Ligon (1981) conducted a literature review of testwiseness that resulted in an attribution list of testwiseness. Testwiseness can be described, measured, and taught, and only minimally relates to intelligence. The effects do not last long and are unrelated to gender. However, for some students, testwiseness increases their production power. Dolly and Williams (1986) also trained sophomore college students in testwiseness skills and found testwiseness teachable and learnable. Findings such as these lend support to the test preparation industry.

Teachers, in an attempt to provide a positive means to prepare their students for testwiseness, turn to commercially manufactured materials that purport to increase standardized scores. These materials exist in various forms from paper-and-pencil to computer-interactive preparation. Mehrens and Kaminski (1989) examined four of the twelve commercial test preparation packages currently on the market. These included *Improving Test-Taking Skills:* Riverside Publishing Co., *Scoring High - Subject Center*: Random House, *CAT Learning Materials*: CBT McGraw Hill, and *Scoring High -Test Specific*:Random House. These materials are used in conjunction with nationally advertised and applied standardized tests such as the *Iowa Test of Basic Skills* and the California *Achievement Test.* Although the preparation materials aligned with the tests for the most part, the approach taken was similar to coaching. All included practice sheets (drill and practice) and test taking practice. An analysis of the literature by Mehrens and Kaminski (1989) yielded negligible differences between students trained with commercial test preparation materials and those prepared otherwise. Of the teachers surveyed by Hall (1990), 66% used commercial test preparation materials. The most critical aspect of

testwiseness is that a testwise student can use test characteristics to score higher even with little or no knowledge of the test content (Ducote, 1982).

The effects of testwiseness training extend beyond the purview of the testing site. Herman and Golan (1990) found through a survey of 341 classroom teachers that pressure, especially from affluent parents, has increased to raise scores. Administrators and the media have applied pressure on the lower income group to increase in scores. Edleman (1981) found that testwiseness training influenced teaching methods and that testwiseness preparation increases anxiety in 50% of the students. Students are denied thinking and problem solving skills for the sake of increased scores (Shepard, 1991). Shepard asserts that high-stakes preparation can result in inflated scores that give a false impression of student achievement because an increase in test scores does not necessarily reflect increased learning.

Despite these facts about the questionable effects of mandated testing, $500 million dollars are spent on standardized tests each year on the issuance of 200 million test forms to students in the United States (Prell, 1986). The uses for these tests vary in different school districts and states. Prell noted that test scores are used for grouping of students, placement of transfer students, identification of maladjusted children, vocational and educational goals, assessment of achievement, assessment of teacher effectiveness, and evaluation of programs. Webster et al (1994) define school effectiveness as possessing a sense of mission, strong leadership, high expectations for students and staff, frequent monitoring of student progress, positive orderly climate, sufficient opportunity for learning, and parent and community involvement.

Effectiveness Indices

To assess school effectiveness and subsequently teacher effectiveness, Webster (1995) subdivides the task of assessment into several indicators. First, input indicators comprise school enrollment, socio-economic status, language acquisition, enrollments, ethnicity, and financial resources. Process indicators include what is being taught, the way it is being taught, consensus of school goals, instructional leadership, opportunity to learn, school climate, staff development, and collegiality among teachers. The outcome indicators of school and teacher effectiveness are defined as student academic performance, attendance rates, dropout rates, completion rates, performance in subsequent grade levels of schooling, individual school goals, parent/student satisfaction, percentage of students completing advanced courses, and college attendance.

Webster (1994) and his team of researchers from Dallas Public Schools, in a response to demands from the state and community , have devised a means by which they statistically demonstrate which schools are effective. Using the outcome indicators as criterion variables, the team employs a series of multiple-regression formulas. Input indicators serve as regressors that predict levels of achievement for schools with like populations based on the predictor variables. The predictions are then compared to the actual outcomes of the schools on the various output indicators.

Students' current scores are compared to the predicted scores. Schools whose comparison score (residual gain) is within an acceptable range of the predicted score (on or above the predicted score) are considered effective schools. Schools that fall below the predicted score are targeted for remediation. The targeted schools receive increased amounts of inservice and intervention from the district. To date, no studies and data are available on the effects of remediation in this setting. However, supporting data from other studies of the negative effects of external pressure indicate that this situation is no different

(Anders, 1992; Smith and Rottenberg, 1991;Nolen et al, 1992; Paris et al, 1991 Airasian et al, 1983).

Definitions of Terms

The following terms are pertinent to the understanding of the study.

1. High-stakes testing has serious consequences or trigger actions such as selection, merit evaluation, promotion, retention, and take-over decisions (Smith, 1991)

2. Testwiseness is the ability to respond advantageously to multiple choice items containing extraneous clues and to obtain credit on these items without knowledge of the subject matter (Diamond, 1972); student ability to use characteristics of the test and the test taking situation to an advantage (Seaton, 1992);

3. Generative learning is the process of learners learning to construct relations between prior learning and new information (Wittrock, 1992)

Residual gain scores - A residual gain score is the difference between a respondent's raw score and a predicted score on a the same measure

4. Constructivism is a cognitive process of creating meaning and understanding in learning

5. High-road transfer is mindful abstraction or decontextualization of knowledge or skills for application in another context (Fogarty, Perkins, Barell, 1992)

6. Metacognition is an awareness of one's personal thought processes; thinking about one's own thinking (Burke, 1992)

Assumptions

There are two assumptions associated with the proposal and completion of the study. First, it was assumed that eighth grade students could learn to write test items within six to eight consecutive weeks. Second, it was assumed that the students could motivated by their own experiences to generate test items.

Limitations

This study was limited by several factors. The most obvious limitation was the population in question. Although this study was conducted in a public school, the school in question had such a high Hispanic population that it was somewhat difficult to generalize outside the population. The population under investigation was limited to eighth grade students. Again, generalizability was limited. One advantage of the limitation was that this population was under scrutiny and, in isolating them, the study lends new knowledge to the existing body of knowledge on how eighth grade students prepare and fair on mandated tests.

Another limitation was the implementation of instruction. All instruction was presented by the researcher. This limitation posed threats of bias on the part of the researcher. Due to the limited nature of the proposed presentation format, generalizability was limited. However, students' achievement in test item writing, demonstrated by higher skill in TAAS testing, adds additional knowledge to the existing body of knowledge about preparation for mandated tests.

A final limitation was time. Instruction was limited to 20-25 minutes per session once each week for twelve weeks. The decay in memory across that period of time coupled with the time of day (early morning), the type of period (homeroom), and the infringement on time (announcememnts) could have created some limitations on the success of the project.

CHAPTER II

RELATED LITERATURE

High Stakes Testing and The Effects of High Stakes Testing on Beliefs and Practice

Teachers and students involved in high stakes testing situations are under pressure from external forces to raise scores on mandated tests (Dangler, 1994). Popham (1991) contends that high stakes testing came about in the 1980s and was based on public perception of poor student performance and the questions of whether tax dollars were being spent wisely. Community members, business leaders, state education agencies, district education supervisors, and the media constitute external forces. Constituents of external agencies demands accountability from the schools. Accountability manifests itself in the form of scores on mandated state and district tests.

High stakes testing relates directly to teacher and school accountability and can have far reaching repercussions. Smith (1991) has defined high stakes testing as test results that have "serious consequences or trigger actions such as selection, merit evaluation, promotion, retention, and/or takeover decisions." Popham (1991) states that in the era of educational accountability that has resulted in high stakes testing, teachers should be prepared to teach the test. Shepard (1990) points out that high stakes testing, intensified by pressure from the media and financial incentives, is politically driven and results in distortion in instruction on all levels. In a study of the effects of high stakes as related to the practice of preparation, Shepard (1991) has found that of the teachers surveyed 79% felt pressure to improve test scores and reported that the pressure came from administrators

and the media. Teachers have reported further that there seemed to be an extensive use of test results for external purposes.

Among external purposes cited by Archbald and Porter (1990) were policy purposes of evaluation and accountability. In top-down educational reform the trickle-down effect from external pressure has followed in the steps of reform reports and business pressure on state and national leaders to inform the media. Schools received blame for perceived educational ills. Consequently, schools intensified efforts to increase scores on tests. The greatest loss documented by Archbald and Porter was the lack of measurement or disclosure of what students accomplish in problem solving and conceptually understanding, neither of which are part of reform policies. The authors noted the effect of the shift to prescriptive testing resulted in the erosion of teacher professionalism.

A federal judge, in 1990, declared the state of Kentucky education system unconstitutional. The state was ordered to reform completely its system and educational practices (Bridges, 1996). To that end, numerous levels of reform were initiated, that included state mandated tests consisting of several components. For example, normative and open-ended criterion reference items that require written responses were administered to students at the fourth, eighth, and tenth grades.

Teachers in Kentucky are expected to comply with the new mandates of the state tests by preparing students in a manner that would best enhance student proficiency on the tests. Four levels or thresholds for achievement include novice, apprentice, proficient, and distinguished. Within each of these levels of student achievement, as measured on the KIRIS (Kentucky Instructional Results Information Systems), the state mandated test is held as the constant. Schools are expected to raise student scores by 1% above their threshold each year and 10% above the novice point. If these benchmarks are not met,

sanctions are placed on the school. However, if benchmarks are met or exceeded, monetary rewards are given (Kannapel, et al, 1996).

Various observational research initiatives have been put into place to determine whether teachers are complying with the mandates. These studies reported that 90% of the teachers observed used open-ended items on tests in classes and 70% used the KIRIS assessment of curriculum report (Matthews,1996). In contrast Bridges (1996) found that less than 50% of the elementary teachers observed were using state recommended instructional practices for sciences, social studies, and integrating the arts into the curriculum. In addition, less than 50% were using flexible grouping, a variety of instructional strategies, learning centers, and broad themes. Reidy (1996) reported that fewer than 40% of the students tested, which constituted 99.6% of the student population, were above the novice level in testing, the lowest level measured by the state. In spite of these findings, the state of Kentucky is revered as a bastion in educational reform, testing, and measurement.

The advent of statistical packages (hierarchical linear modeling) that detect differences among students in a given class, from a particular school, within a certain school district has given district and state psychometricians the tools necessary to isolate differences among students. Armed with this knowledge, teachers can then be cited for inadequate preparation of students for mandated tests, labeled as ineffective, and offered remediation for becoming a more effective teacher. Currently forty-four of fifty states require testing in the schools. Seven of the forty-four initiated a value added measurement component for rewarding teachers for raising scores (Webster, 1994, 1995; Sanders, 1996).

Paris, et al (1991) found that high stakes testing served two purposes that included establish performance standards and drive the design for the curriculum. Educational assessment was studied from three perspectives, the psychometric, political, and

psychological impact on students. From the psychometric view, Paris noted that current standardized tests were aligned with outdated learning theories that promote learning as a discrete collection of skills and knowledge and lack instructional validity. Political use of data was generally inappropriately reported and used by real estate companies to compare districts. The tacit acceptance of media reports promoted further financial benefits to commercial test publishers, deepened the reliance of the public in quantitative comparatives, and encouraged the blind faith in the accuracy of data reported (Bracey, 1994 and 1995).

In Paris et al.'s examination of student psychological impact several revealing insights surfaced. Students in elementary, junior high, and high school were administered a 20 item survey regarding their attitudes and beliefs about achievement testing. Results were mixed. Elementary students were highly motivated and overall had a good attitude about testing. In opposition, the junior high and high school students felt victimized by the tests and had an overall negative attitude toward the tests. The attitudes were attributed to growing disillusionment about the tests, decreased motivation, increased inappropriate strategies, school practices, and overall anxiety and fear of failure on the exams. Testing played a major role in the secondary student's perception of self. The authors recommended that stakeholders be surveyed and that emphasis be placed on perspectives that complement dominant views in the development of assessments that are collaborative, authentic, longitudinal, and multi-dimensional (1991).

Dangler (1994) reported on the high stakes situation in Ohio where intervention and remedial strategies were being used to circumvent possible failure of competency tests by graduating seniors. In the study ninth grade students were tested for potential on the competency tests administered to graduating seniors. Those falling below a certain level were given remedial coaching to achieve 100% passage of the mandated test. Thirty administrators were surveyed regarding remedial strategies used. Respondents reported that the five preferred methods included group tutoring, individual tutoring, progress

letters, communication of high expectations, and a shared sense of responsibility. Among these groups, tutoring was used by 96% of the respondents. This study supports Paris' position that tests were used to raise standards and direct the curriculum.

Standards as defined by the American Heritage Dictionary (1996) is "a degree or level of requirement, excellence, or attainment." In order for teachers to raise their practice to the standards of educational reform policies there must be some level of ownership and belief in the paradigm shift from previous practice to current recommended practice.

Anders (1992), during staff development sessions focused on conversations of thirty-nine teachers designed to investigate teachers' beliefs and practices regarding reading comprehension. The sessions were videotaped and then analyzed for patterns and topics. Two major themes emerged that included grading as accountability and teachers' mistrust of their own judgment when evaluating students with 50% of their time spent discussing assessment. The results indicated that teachers were anxious about external tests and believed that testing played a negative role in their schools that lead to confusion among teachers about the nature of the tests. These findings were confirmed by Kannapel (1996) in her observations and interviews with teachers in Kentucky and by Smith (1991) in a fifteen month observation of teachers in Arizona operating under similar conditions.

Sanctions and rewards for meeting benchmark standards netted unexpected outcomes. State agencies, districts, schools, and teachers have resorted to interesting means to ward off the stigma of under achievement. Cannell (1988) surveyed testing agencies all 50 states regarding the status of their students on nationally normed tests. All fifty states reported that their students were above average. By psychometric standards it would seem unlikely that all students could be above average as the nature of a normed reference test would imply that 50% of the students would fall below the central mark. Brandt (1989) concurs that the use of sanctions and rewards consummates in individuals

using extreme means to ward off the humiliation of sanctions and receive the attention of rewards. Unfortunately, the quest for recompense has led to test pollution.

According to Nolen et al (1992) the major pollutants in testing involve preparation methods, variability in testing conditions, and outside factors. These pollutants provide scores that are reflective of increased achievement in test taking, but not necessarily gains in learning (Shepard, 1990, 1991; Nolen, 1992; Popham, 1991). Popham (1991), a proponent of measurement driven instruction, surveyed 172 teachers and found that of the five common practices used in preparing students for tests, only three were educationally defensible and among those three two were professionally ethical. He states further that test preparation practice should commensurately increase student test scores while simultaneously increasing mastery in the content domain tested. Callenbach (1973) administered intensive instruction in test taking techniques to a random group of second graders and found that scores on standardized tests can be raised independent of content knowledge. In spite of this plea for relevance and ethics in preparation, the pressure to succeed has forced teachers, schools, and districts to adjust their educational standards to meet the demands of the testing standards.

Mandated Testing as Related to Instruction

There is some debate among measurement scholars concerning measurement driven instruction (MDI). According to Popham (1987), measurement driven instruction occurs "when a high stakes test of educational achievement, because of important contingencies associated with the students' performance, influences the instructional program that prepares students for the test." He goes on to describe two types of high stakes tests: those that are important to examinees in the case of graduation or promotion and others that reflect instructional quality across the state that is reported by the media. Either of these two types of high-stakes tests influence the nature of instruction.

Popham asserts that measurement driven instruction is the means by which teachers can deal with the challenge of high stakes testing. He reports improvements in basic skills in seven states, including Texas, as a result of the implementation of measurement driven instruction. He reports that five criteria enhance the effectiveness of measurement driven instruction. Popham first recommends that the testing instrument be criterion referenced. Criterion referenced tests, when directly applied to instruction, provide advance knowledge of testing procedures and content and specific targets for instruction, the second criteria. Popham recommends that targets for testing be limited to five to ten areas.

Regarding instruction specifically, Popham suggests three important issues for consideration. First he points out that the content selected for measurement must be defensible both psychologically and philosophically. In addition, the measurement tool must provide instructional illumination. Through this illumination effective sequenced instruction emerges. Finally, he recommends that there be instructional support in the form of staff development and ongoing support in the form of materials, additional information, and support staff who are available to answer questions.

Although Popham paints a picture of success with measurement driven instruction, Bracey (1987) challenges the premise of MDI. Bracey equates measurement driven instruction to mandated testing that he defines as literacy passport. The literacy passport gains the learner passage into the accepted realms of the outside evaluators, but limits them in other ways. He noted that although students can succeed on tests that are directly related to the content, several underlying problems result that have far reaching effects on curriculum and instruction. Among these is the deliberate exclusion of areas that are important but not tested. Bracy delineates several areas of instruction that are affected negatively by the implementation of MDI.

Bracey's overarching thesis relates to minimization. He points out that when MDI is incorporated, fragmentation occurs. Learning broken into small bits leads to a problem

of information bitting. Information bitting involves giving bits of information that are out of context but tested. Bracey notes that considering MDI, extraneous skills become targets of instruction leading to the lack of teaching for transfer. He goes on to add that the emphasis on academic talent fosters test induced inflection. The indifference to other areas of accomplishments compounds the problem for the learner who is academically challenged.

Bracey demonstrates his concern for the total learner in his expose of the tested learner and the life long learner. He questions the mistaken relationship between test scores and personal worth. Bracey ponders the veracity of success in later life and achievement as measured by mandated tests and finds the association limited. Findings from the SCANS report (1992), which examines the needs for the workforce beyond the year 2000, imply that success on mandated tests warrants limited attention. Skills required for employment set expectations for learners who can learn to learn as opposed to learners who imitate and adhere to concrete structures.

Cohen and Hyman (1991) examined alignment of testing and instruction in a study of Missouri's statewide criterion referenced testing. They found that alignment between instruction and testing netted higher test scores. Increase in scores is attributed to several factors. Teachers were given sample items from the Missouri Measurement of Achievement Test (MMAT) and taught to teach models that aligned with the test. In addition, the curriculum had precise outcomes test-related that assumed alignment.

Curriculum theory has purported alignment of instruction with testing as valuable and necessary (Glatthorn, 1994). However, Cohen and Hyman found criticism of the Missouri approach. They observed teachers teaching to the test and matching teaching with items and item types from the test. In many cases the teachers used rote practice to drill the students in the information needed to succeed. These practices brought into question the claim of gains and the interrelated rise in intelligence of the students in Missouri.

McAuliffe (1993) observed differences between instructional practice and test preparation. In an ethnographic study spanning twenty-six weeks of class observations of an eighth grade remedial reading class, McAuliffe noted that practicing for tests liberated the students and granted them the power to ferret out correct answers. Reading improvement was attributed to class discussion of reading items. In addition, the teacher considered student interest and background when approaching instruction.

Links between instruction and testing include the variables of content, curriculum and instructional validity (Airasian and Madaus, 1983). Content is linked to test content and educational objectives. Curricular links connect test content to curriculum content. Instructional validity links test content to actual instruction. To establish links, data analysis of item difficulty and a descriptive overlap between instruction and tests should be explored.

An examination of links across all three areas yielded several observations. Airasian and Madaus noted that total scores and subscores hid unique achievement and that no group item differentiation was detected. Also, in contrast to Cannell's (1988) findings, there was no Lake Wobegon Effect. They recommended strategies for assessing the overlap between instruction and testing that included building new criterion reference tests for each new testing and instructional context, altering tests by adding or subtracting items, constructing detailed taxonomy of the curriculum to mirror the test, and use of judges to determine the extent of the overlap.

Issues pertinent to linking tests to instruction range from definitions to policy. Instruction defined in an operational sense implies that instruction is timely, adequate, and dependent on reinforcement. Providing opportunities for learning is not adequate to the task of linking instruction to testing. The policy issues addressed by Airasian and Madaus stress the importance of using tests to evaluate the instruction of a single group as opposed to making policy. Test publishing, an intervening variable in the creation of links between

instruction and testing, , creates difficulties and exacerbates the problems associated with the relationahip of instruction to testing (Houston, 1994; Bracey, 1995; Stedman, 1994).

Anderson (1981), in a discussion of the differences between instruction and coaching has developed five categories that characterize both. The categories include duration, purpose, techniques, retention, and success. Instruction requires longer periods of time than coaching. The purpose of coaching, to improve performance on a particular exam differs from instruction which develops skills, knowledge, and understanding. While coaching emphasizes drill and practice, instruction provides alternative explanations of phenomenon and solutions to problems. Students in a coaching situation retain less than students in an instructional setting. Overall success varies in a coaching situation and is dependent on student motivation, whereas the instructional setting creates success through structure and motivation.

Instructional Strategies Teachers Use to Prepare Students for Mandated Tests: Appropriate Practice

Whether the approach is coaching or instruction, teachers must consider strategies that will give their students the independence required to succeed. Among the approaches available, the most widely used is testwiseness. Diamond and Evans (1972) define testwiseness as "the ability to respond advantageously to multiple choice items containing extraneous clues and to obtain credit on these items without knowledge of the subject matter." (p.45) Seaton (1992) explains testwiseness as "student ability to use characteristics for the test and test taking situation to an advantage. Testwiseness exists independently of the knowledge a person has about a subject matter." (p. 2) Dolly and Williams (1986) state that testwiseness is " cognitive strategies used to maximize chances of selecting the correct option when the test taker was forced to guess due to lack of

information concerning the content of an item." (p. 620) Essentially, a testwise student can navigate through a multiple choice test using testwise skills and succeed without knowing anything about the subject area being tested.

To maximize scores on multiple-choice tests, Dolly and Williams (1986) experimented with fifty-four sophomore college students. The students were randomly placed in groups of which the treatment group received training in testwiseness strategies. Students were taught to examine the items for length of option, middle value, similarity or opposition of cue, and cues from stem to option. The results of the study indicated that students can be taught testwiseness skills.

Diamond and Evans (1972) administered a testwiseness scale to ninety-five sixth graders to determine the cognitive correlates of testwiseness. Five item faults; association between stem and examination, distraction, correct alternatives, grammar clues, and overlapping distractors were examined. Results showed testwiseness as a trait that is present which correlates at .28. It was noted that there were no significant differences in treatment of test preparation which included training in following directions, proper use of time, use of answer sheet, guessing strategies, deductive reasoning, and cue using. However, there were significant differences in the number of contact hours. More hours in preparation had a greater effect than less hours.

Specific techniques employed by students as testwiseness strategies provide hints for confronting items. Among the methods discussed, seven are commonly recommended and include using context clues, syntax hints, guessing, knowing the logic of the test, know the test rules, time management, and process of elimination (Summers, 1983; Montgomery County Board of Education, 1982; Phillips, 1983; Palmer, 1984; Ligon, 1981). Phillips contends that students need testwiseness strategies to succeed in raising scores. In a high-stakes testing arena, these skills become vital.

The handbook developed by the Montgomery County Board of Education (1982) represents a strong emphasis on testwiseness preparation. Ten additional strategies beyond the common seven include checking answers if too easy, reading carefully, ignore irrelevant material, unusually long or short alternates indicate an incorrect choice, answers with opposites (one correct), examine all choices before selecting, place test in relationship to answer sheet (writing hand), follow directions accurately, watch for double negatives, and question and answers in grammatical agreement. Teachers are encouraged to use the resource in planning test construction and student preparation for additional standardized tests.

The Los Angeles Unified School District directed teachers in assisting students with preparation for tests through a bulletin issued in 1982. The document divided test preparation into two sections. In the first section characteristics of the test were defined. The definition centered on test content, format, answer keys, scoring, and time limits. In the second section students were readied for the test. During this time teachers prepared the students for content, testwiseness, and the emotional impact of external testing.

Additional skills and cues for testing success are poffered by the Austin Independent School District. Ligon (1981) has examined the test preparation recommendations of the district and has found preparation broken into two categories. These include skills and cues. Skills that differ from the seven listed previously involve understanding test terminology, physical readiness, understanding the format, symbols and procedures, knowing how to ask questions about the test, and thinking logically. Differing cues listed the elimination of options that imply each other, the elimination of outrageous options, the selection of more carefully worded items, the selection of items with specific details, and inferring the intent of the test maker. These skills are included in the district's recommendations to teachers as appropriate teaching methods for preparing students for standardized tests.

Robinson and Wronkovich (1993) explored other testwiseness strategies. In their study, eighth grade students, identified as at risk for success in taking and passing standardized tests, participated in a program designed to aid them in test preparation. Remediation programs supervised by adult volunteers and peers provided the training the students needed to prepare for the tests. Alternative to accepted practice students viewed series of one hour videos. Each video focused on one area of the test, covered learning outcomes, and presented several examples. In a final session students practiced the test in a large group. Failure on the test mandated additional remediation and modification in the school calendar.

Flippo and Borthwick (1981) studied pre-service teachers as they studied testwiseness strategies in a test and measurement course and applied their knowledge in an instructional setting with second through sixth graders. Elementary students were assigned to experimental and control groups. Students in the experimental group received training in testwiseness strategies as a part of a social studies unit. The control group received instruction in the same social studies material but were sent to the library or engaged in art projects during testwiseness instruction. There were no significant differences observed between the groups when tested following the treatment on social studies facts presented in a multiple-choice test.

Kher-Durlaghji and Lacina (1992) surveyed seventy-four preservice teachers regarding views of high stakes test strategies. Items on the survey included acceptable and non-acceptable test preparation practices. Pre-test items referred to enhancing student motication, building the curriculum for the test, developing teacher objectives to match the test, provide practice with similar items, use commercial exercises designed to enhance scores, and share actual test items prior to the test. Survey items depicting during-test assistance included giving hints, alteration of response sheets, darken student responses, erase stray marks after test, and request no testing of low ability groups. Results of the

surveys indicated that pre-service teachers were more likely to use sending notes home to parents asking them to teach their children test taking skills (75%), practice alternative forms of the test (60%), and use commercially prepared materials (50%).

Another form of test preparation, coaching, trains students specifically to answer questions for a particular test (George, 1985). This training is intensive, short-term, and generally meaningless in the overall education of the student. Coaching kits converge on the requirements for passing the test and include sample stimulus items mirroring the style of the test, princilples and skills required by the test, sufficient sample items, and point out major principles and skills.

Mehrens and Kaminski (1989) identified twelve different test preparation/coaching packages and reviewed four including *Improving Test-Taking Skills: Riverside Publishing Co., Scoring High: Random House, CAT Learning MAterials: CBT McGraw Hill; Scoring High - Test Specific: Random House.* Teachers reported 66% used some form of pre-packaged materials in test preparation (Hall and Kleine, 1990). The purpose of the materials, to prepare students for nationally sold standardized tests, focused primarily on testwiseness skills. Mehrens and Kaminski noted that two of the packages matched requirements for the tests and two did not. A meta-analysis of studies that examined the effectiveness of the use of the materials revealed no differences among students exposed to the materials and those who were not exposed to the materials. Relating preparation to specific items on tests results in exploitation and invalidation of the test. These findings were supported by a study conducted by the Chicago Public Schools Evaluation Bureau (1987).

Cushing (1989) conducted an experiment on matched groups of fourth and fifth graders to determine the effectiveness of test preparation on test taking using the *Riverside* materials. The students in the treatment group received nine to twelve hours of training and the control group received none. No significant differences were seen between the groups

on residual gain scores. Several within group differences observed. The lack of effectiveness of training did not warrant the loss of content knowledge.

Preparing students for tests in isolation of content delivery jeopardizes overall instruction and learning (Mehrens and Kaminski, 1989). With increased pressure to produce higher test scores, teachers are placed in a position of considering inappropriate means to meet the demands of external forces. Appropriate practice embodies general test-taking skills, matching instruction with objectives measured, and the use of varied formats. Marginally acceptable practice includes using commercial preparation, instruction matched to test items, and instruction matched to the format of the test. Teaching directly to the same form or a previous form of the test are considered unacceptable and unethical (Popham, 1991; Mehrens and Kaminski, 1989; Kilian, 1992).

In a meta-analysis of thirty controlled studies of coaching programs Bangert-Drowns, et al (1983) sought to uncover the characteristics of coaching programs, subjects employed, methods used, and publications. Training program components incorporated training in testwiseness, anxiety reduction exercises, actual practice on test items, and direct content teaching. Findings showed that in 83% of the studies coaching had a positive effect. Nine percent of these studies showed significant differences. Instruction in broad skills was the most effective when compared to cramming or concentrated intensive drill. In contrast, Sruggs et al (1986), in a review of forty-four studies that compared test preparation with achievement of elementary students found that no significant differencs were noted in scores of students who had received training in testwiseness when compared to those who had not received training.

Jones and Ligon (1981) in a review of the literature of preparation for standardized testing, found that three variables affect student achievement scores. These are testwiseness, practice tests, and test practice. They noted that a testwise student's score is not a true reflection of the trait being measured by the test and that testwiseness is

independent of the student's knowledge of content. The review suggested that testwiseness 1) can be described, measured, and taught, 2) is only midly related to intelligence, 3) is unrelated to gender, 4) has short term effectiveness, 5) for some students increases output, and 6) has unknown differential effects.

The Effects of Standardized Tests on Preparation Practice: Teacher Reactions

The impact of mandated testing extends beyond the newspaper reports, school board rooms, and administrative offices of state and local school districts. Madaus (1989) reports that the harmful effects of mandated testing range across schools, educational and curricular goals, and local control. Students are tracked, placed according to test scores, forcing teachers to instruct in a rote fashion to slower learners. The curriculum is narrowed to accomodate the skills and knowledge examined on the tests, impedes student progress, inhibits higher order thinking, and focuses primarily on basic skills (Neill and Medina, 1989).

In a survey study conducted by Edleman (1981), teachers responded to questions about their beliefs about the impact of mandated tests. Among the teachers reporting 50% noted that tests cause student anxiety, 38% thought that test scores were used to evaluate them, 90% questioned the relation between the value and costs of the tests, whereas only 13% thought testing was important. Forty percent of the respondents recounted a change in teaching methods to accomodate testing mandates, while twenty percent were asked to change by the district. Teachers preferred that tests be reported as mastery as opposed to precentiles.

Mandated tests effect adjustment in the curriculum producing spurious results at times. Mehrens and Phillips (1986) sought to uncover the relationship between testing and

instruction in two midwest schools. Teachers were divided into experimental and control groups. Teachers in the experimental group adjusted their curriculum and taught to the test. The control group proceeded as usual. A MANCOVA test yielded no significant differences between the groups. Herman and Golan (1990) surveyed 341 teachers on six areas of concern regarding preparation for standardized tests and determined that mandated testing had a substatial influence on curriculum, content, instruction, and learning activities. In spite of the necessary adjustments, teachers concluded that standardized tests do not improve schools nor do they assess student learning. In the effects of testing project in Los Angeles (1989), the researchers found that teachers were spending at least 10% of their available curricular time in test preparation practice.

The surge of test preparation, brought on by pressure to perform at increased levels, strengthens the teachers' resolve to encourage achievement and test success at all cost. Canner (1992) examined the excesses in non-ethical practices of teachers and administrators in a high-stakes testing setting. The findings indicated that teachers stole tests, gave students answers in advance, practiced on the identical test, focused solely on to-be-tested items, limited instruction to answering questions similar to those on the test, and spent a great deal of instructional time on specific objectives of the test. Increased pressure insures the proclivity of test pollution, the total honesty of test score explanation (Haladyne, Nolan, and Haas, 1991).

Teachers in a high-stakes accountability program in Kentucky were observed and interviewed regarding their impressions of a five year program of measurement driven instruction. Teachers reported adjustments in their curriculum to match the test and in some cases were noted to be holding students back a year before the benchmarked year of testing. They increased the number of credits in areas tested and deleted otherwise excellent electives to accommodate test preparation in areas tested. Most teachers believed

that the tests undermined their professionalism and were not motivated by possible rewards (Kannapel, Coe, Aagaard, Moore, 1996).

In a second study conducted in the state of Kentucky, forty-four schools were investigated for their implementation of professional development initiatives that coordinated with school reform in the state. Although 77% had annual professional development plans and 79% had individual professional development plans, only 18% integrated the two. The implication is that individual teachers are not necessarily aligning their personal efforts with the reform. This further demonstrated by the fact that 57% of them did not have clear mission statements.

Although many studies cite the negative impacts of testing on the curriculum, teacher integrity, and student anxiety, Millman (1981) stands in defense of testing. Testing is viewed as a means of protection from incompetence and ineptitude. Lack of public accountability brings into question the veracity of education. Testing provides an efficient use of social resources and educational procedures while adding to the body of knowledge about education. Finally, talent laying dormant reveals itself through testing.

The Uses of Mandated Test Scores: The Impact on Test Preparation Practice

Student groupings for placement, based on test score results, limits academic freedom and potential for students who may otherwise have the capacity to demonstrate academic excellence through other untested means. Other uses of test scores discussed by Prell (1986) include the establishment of vocational and educational goals, identification of maladjusted children, assessment of achievement, assessment of teacher effectiveness, and evaluation of programs. The use of scores to increase administrative laurels dictates that school districts switch to easier tests, eliminate populations tested, increase coaching, and teach more testwiseness.

In the report *Nation at Risk*, United States students showed an overall decline in school aptitude, lower achievement, gifted students falling below expected levels, and general functional illiteracy among disadvantaged students. These striking results prompted educational reform on a grand scale. Testing, percieved as a means of accounting for reforms, is used for purposes other than those intended. Selden (1985) reports that test scores are used to drive the curriculum. Because teachers teach what is tested, the tests fail to serve as measures of effective programs.

The top-down reform effort of the 1980s placed business, state, and national leaders in partnership with the media. This alliance, according to Bracey (1995), resulted in media-mangling of public education. This led to public suspicion of American public schooling. Speakers such as William Raspberry, William Bennet, Newt Gingrich, and Rush Limbaugh target the public schools as responsibile for the nation's ills (Houston, 1994; Bracey, 1995). Accountability of a quantifiable nature among business supporters of education and media associates increased the stakes for testing and higher scores. To that end lower functioning schools were forced to generate effective school plans, placed on probationary accreditation status, and others taken over by state governments. Archbald and Porter (1990) labeled the use of tests for these purposes academic bankruptcy legislation.

To prompt schools to comply with policies of accountability, districts and state agencies offer financial incentive. In the state of Kentucky schools meeting certain levels of academic excellence, as measured by the state mandated test, KIRIS, are rewarded with cash bonuses of $3600. Schools falling below expected levels are given sanctions until they comply. Fifty-four teachers interviewed in Kentucky resoundingly agreed to reject the money for less pressure to raise scores (Kannapel, Coe, Aagaard, and Moore, 1996). Brandt (1989) contends that when testing is rewarded with cash, testing becomes polluted. When stakes are tied to monetary acquisition, teachers will find a way to raise the scores.

The state of Texas and the Dallas Public Schools work in cooperation to foster test score improvement through incentives similar to Kentucky. In the Dallas schools, the teachers are rated for effectiveness based on test scores and attendance (Webster, 1994). Statistical processing through hierarchical linear modeling allows the district to sort through which teachers within specific schools are producing increases in scores at the predicted level required by the district. Ranked school scores are reported in the newspaper, which also announces which schools have been financially rewarded for compliance. Schools engage in extreme measures to increase their rank, thus prompting the district to add another variable to the regression formula. Schools are now considered for financial rewards if all students are tested, 80% of the students have been enrolled for at least five of the six week terms, and have maintained high attendance (Webster et al, 1995).

How Teachers Prepare Students for Mandated Tests Without Jeopardizing Skills Needed for the Workforce of Tomorrow

Webster (1994) points out that although there may be problems with testing, it is here to stay for a long while as a means of measuring school and teacher effectiveness. With that in mind, it is imperative that alternative methods to those currently used be found to prepare students without imperilling higher order thinking. Shepard (1991) contends that current practice in test preparation practice is based on antiquated learning theories of association and behaviorism that promoted bit learning of facts out of context. Newer theories require thinking and active constructivism.

Driver, et al (1994) defines constructivism as a method of learning that engages the student in discovery and inquiry processing. Teachers serve as catalysts for learning and set the stage for group involvement, decision making, and reasoning. Students are

encouraged to derive meaning out of what they gather from experiments conducted and, in essence, generate their own knowledge about a phenomenon.

Wittrock (1991) conducted a study that examined student retention of passages read. Students were divided into three groups of two experimental and one control. All students read a passage. Group 1 students were asked to generate summary sentences for each paragraph. Group 2 developed headings and generated paragraph summaries. Group 3 read the texts as many times as wanted within the time permitted. Groups 1 and 2 showed higher levels of retention when tested on the passage later. Wittrock attributes the higher levels of retention to the reworking of the information in such a manner that it became meaningful for the learner.

Koiurilsky and Wittrock (1992) studied 142 high school seniors labeled low socio-economic status and sought to determine their level of understanding of economics knowledge. Students were divided into experimental and control groups. The experimental group was trained in the techniques associated with generative learning. The control group received training in cooperative learning groups. Students in the experimental group learned more about economics, was less misinformed, and more confident with economics information.

The experimental works of Wittrock and associates provide the rationale for alternative methods of instruction in all areas of learning. Their discoveries suggest that for learning to be genuine, long-term, and meaningful, understanding must be generated by the learner. Generative learning requires a higher order of thought or processing that imcorporates decision making and judgement. As purported by Fogarty, et al (1992), the use of higher order thinking results in transfer of learning. "Basically, education that does not achieve considerable transfer is not worth much." (Fogarty, et al, 1992; p. x)

CHAPTER III

METHOD

Background Study

In an effort to more fully understand the setting in which the project was conducted, a background study, conducted in the spring of 1996 provided the needed information. This initial study ascertained practices that teachers employed to prepare eighth grade students for the Texas Assessment of Academic Skills (TAAS). In the background study, which employed ethnographic methodology, teachers were observed for two weeks prior to the administration of the annual TAAS tests. The researcher spent one full school day in each of four classrooms and observed teachers and students in their classroom environment in the short period before the tests. During the visits the researcher documented interactions between students and teachers, instruction and subsequent student response to instruction, and patterns of movement from class to class as well as the environs of the school and classrooms as they related to TAAS testing. The results of the background study are extensive and presented in totality in a full report (Tunks, 1996) A summary report is found in Appendix A.

The pervasive atmosphere in the school, fear of failure, emanated from students, faculty, and staff. In summary, the teachers and students in the school dealt with the fear of failure through various means. All of the means led to the end of preparation for the TAAS test. Each person coped with the multiple pressures of high stakes testing. The consequences of failure included embarrassment to the school in the media, teacher tiering by the district, summer school attendance for the students, and upset parents with their

children returning to eighth grade. The lack of middle ground on the issue brought into question the policy of using a single measure to determine the fate of the school, teachers, and students.

The study provided the researcher with valuable, applicable information. The researcher scrutinized the practice of teachers which served as guidance for developing methodology for the study. Student interviews provided the researcher with insights into the perceptions of eighth graders as test designers. The finding from the background study served as a basis for structuring the implementation of the current study.

Subjects/Setting/Context

Subjects/Setting

The teachers and students under examination in the study were members of a junior high school, consisting of seventh and eighth grades, in a large metropolex in north central Texas. The enrollment at the school as of December, 1996 totaled 2,158 students of whom 6.6% were Anglo, 85.9% Hispanic, 6.6% African American, .10% Asian, and .80% American Indian. The percentage of students on free lunch was 73%. Thirty-two percent of the students were served by limited English language programs. There were 94 bilingual and 476 English as second language students in the school. Thirty-two percent of the students were ranked below the 50th percentile on grade equivalency measures. Fifty three percent were ranked above the 40th percentile and thirty-six above the 30th percentile on standardized test scores.

The school was located in an inner-city predominately Hispanic area. This school was a neighborhood school to the majority of the students. However, the school included an arts magnate program and students throughout the district could participate in the arts program. These students applied and received admission into the school based on their interest in the arts. The arts areas included classical ballet, ballet folklorico, band,

orchestra, piano, chorus, visual art, and drama. Students from the neighborhood and the city at large were eligible to apply for entry into the arts academy that was enveloped in the school. Students in the arts exploratory programs attended classes with the students from the neighborhood as opposed to being isolated and instructed separately. Three hundred twenty-four of the students attending the school were bussed in specifically for the arts program.

The neighborhood surrounding the school experienced a fair amount of crime and gang activity. School officials noticed that gang haircuts and clothing styles were becoming more obvious among some of the students in the school (Kimm, 1993). To curb gang pressure and possible violence, the school centered committee, the governing board of the school, attempted to have all students wear uniforms. During the school year 1994-95, approximately 5% of the students complied. Parent organizations and school officials approached the state and the district and received clearance for all students to be either in uniform or attend another school. The population in the school changed from 2,500 to 1,800 within two months. The remaining 1,800 students regularly wore uniforms. During the tenure of the study, the population increased to 2158 students, all of whom wore uniforms daily.

The overall atmosphere reflected an attitude of learning and striving for excellence in education. Posters in the classrooms and the hallways provided encouraging messages to students to do well in classes and on tests. In 1994-95, students convened in the auditorium for "TAAS pep rallies", rallies presented by the building administration to encourage excellence in testing. During rallies, keynote speakers, selected for their inspirational qualities, spoke to the students about succeeding on the TAAS test. Cheers about the caliber of students in the school, as compared to the junior high students down the boulevard, promoted a sense of pride and competition based on the assumption that junior high students could be motivated to excel in testing achievement beyond their

counterparts in another school. In other situations, students who received the highest possible score on the TAAS writing test became eligible to dine at one of the neighborhood's finest restaurants with appreciative teachers. The bilingual principal encouraged all students to do well on tests.

Context

Students, particularly eighth graders, and their teachers were under considerable pressure to produce high scores on the state mandated tests, the Texas Assessment of Academic Skills (TAAS). In the state of Texas, fourth and eighth grade students' scores on state mandated tests were used as pivotal measurement points to assess school and district compliance with state curriculum behests. The state issued a recommended curriculum that embodied certain essential elements. These elements were transformed into specific measurable objectives. Specific objectives became the TAAS objectives.

TAAS objectives were those objectives that the state specifically tested each year in grades three through twelve. The TAAS objectives, issued to all districts, whose responsibility it was to interpret these for the teachers, became the cornerstone for district curriculum planning. The school district in which the junior high operated issued an interpretation in the form of a written curriculum, *Frameworks* (1992). The *Frameworks* document delineated the TAAS objectives into manageable outcomes. Each teacher was responsible for adopting these objectives and incorporating them into his or her taught curriculum.

Due to the overwhelmingly low scores on the TAAS across the state (Webster, 1994), the Texas Education Agency initiated renewed efforts to encourage teachers to use the TAAS objectives as part of overall instruction throughout the school year. Encouragement came in the form of financial rewards for schools that met or exceeded measurement standards. Schools that raised scores, attendance, and graduation rates were

financially rewarded by the state. Those failing to raise scores were targeted by the state for remediation and possible takeover. Within the district, schools and teachers were financially rewarded as an institution and as individuals when gain scores were at or above expected levels (Webster, 1995). Additional incentives encouraged schools to decrease disparity between minority and Anglo students' test scores.

Multiple efforts on the state and district levels were made to raise scores and decrease disparity. The Texas Education Agency (TEA) published measurement specifications for each test at each grade level. Since 1995, teachers have had access to previous tests and were encouraged to use the tests as a guide in preparing students. Teachers were required by the district to profile each student in each area of the TAAS to determine areas of strength and needs. They were then expected to adjust instruction to address student needs. Teachers from targeted schools were trained in incorporation of TAAS objectives into overall instruction. Observations of teachers in targeted schools responded positively to the profiling and adjustment (McNeely, 1996).

Instruments

Test Anxiety Inventory

The *Test Anxiety Inventory (TAI)* was designed to provide users with a clear measure of test anxiety focusing primarily on worry and emotionality. The twenty item self-report instrument assesses how respondents rate themselves on their level of anxiety during tests. A four point frequency scale that includes the levels of: 1) almost never, 2) sometimes, 3) often, and 4) almost always. Testers read statements indicative of worry or emotionality and respond according to their personal perceptions of themselves as test takers.

The TAI, constructed to provide an easy-to-use instrument, compared favorably with the Saranson's Test Anxiety Scale, a more complex measure. The TAI, when

compared to the STAS, concurrent validity correlations for males was .82 and .83 for females. In addition, alpha coefficients for the TAI, .92 on the entire measure and .88 on the TAI/W and TAI/E were .88 and .90 respectively. Test-retest reliability stability coefficients across two weeks and one month were .80 and higher for both high school and college students. The coefficient dropped for the high school students after six months to .62.

Although the instrument was calibrated with students older than eighth graders, it was believed that the measure served the purpose in the proposed study due the simplicity of the questions and response options. Speilberger (1980) confirms that the TAI had been successfully administered to junior high students. The items are open-ended and easily understandable for students only one year younger than the group for whom the test was calibrated. Another consideration was the number of tests these students encountered in their young years. It seems logical that the students would be able to respond appropriately and successfully to the expectations of the instruments. According to Buros Mental Measurement Yearbook the instrument was rated highly on reliability measures. However, the two reviewers of the instrument found some discrepancies with the concurrent validity documentation provided by the author of the instrument. In spite of these concerns, the instrument was recommended for use as a measure of anxiety or attitude toward testing.

Texas Assessment of Academic Skills

The Texas Assessment of Academic Skills (TAAS), a criterion referenced test, has been designed to measure achievement in the attainment of educational objectives outlined in the Essential Elements. The essential elements constitute the state recommended curriculum and encompassed all content areas at all grade levels. The TAAS tests, primarily consist of multiple-choice items in the areas of math, language arts, science, and social studies, are administered annually across the state in all public schools in Texas.

Students in third grade through twelfth are required to take the tests annually. Exceptions are made for special education, non-English speaking, and special needs students.

Validity and reliability figures, provided by the Texas Education Agency imply overall stability. According to the test designers, the test meets content validity requirements by aligning with state objectives. However, no predictive validity information is publicly available on the test. The instrument was tested for internal consistency using the Kuder-Richardson Formula which yielded coefficients ranging from .77 to .93 across all content areas tested and all grades levels tested. These figures indicate that the tests are stable, according to the TEA.

The test was not listed among tests in print, nor was it reviewed in the Buros Mental Measurement Yearbook. This lack of external scrutiny brings into question the figures provided by the TEA. This limitation does not preclude the use of the measure, as it was the measure by which the state judges achievement and eventual passage from eighth grade to ninth, graduation, and sanctions and rewards for schools. Consequently, the measure was used as one of the criterion variables in the study.

The language arts reading test, of concern to this project, consists of several parts. These parts include word meaning (4), supporting ideas (10), summarization (6), relationships and outcomes (6), inferences and generalizations (14), and point of view, propaganda, fact and non-fact (6). The numbers following each component indicates the number of items on the TAAS that are presented within that area of examination. There were forty items on the eighth grade test. School and student performance was rated on mastery, 75% accuracy, in each item category within each subject area tested.

The diagnostic TAAS, administered at the beginning of each school year, included twenty items in language arts reading, which correspond in direct ratio of .50 to the items on the actual TAAS. This version of the TAAS provided teachers with pertinent information from which they adjusted their curriculum, course of study, and instruction to

accommodate the needs of learners in each class. These adjustments aimed toward mastery of the TAAS in all areas of examination, for all students. For the this study, the diagnostic TAAS was used. The actual TAAS will be administered in the late spring and will be beyond the time frame of the study.

Design

The study was designed using an experimental posttest only control group design. Students were, by design of school scheduling, randomly assigned to homeroom classes. One team was selected by the principal to serve as the subject pool for the study. The team consisted of five classes of eighth graders who were randomly assigned to the team. Of these five classes, four of the classes were randomly selected to participate in the study. Due to Friday school-wide volleyball games, the fifth class of students did not participate in the study. To select the four classes from the team that participated in the study, Monday through Friday, was written on slips of paper and one was left blank. Each of the five teachers secretly selected one piece of paper, thereby eliminating one intact class from the selection pool. The four classes of students became the subjects for the study. From the pool, students were randomly placed into experimental and control groups.

Based on a mean class size was twenty-five, twelve students from each class were randomly selected to participate in the study as experimental subjects. The remaining students were included in the control group. Students were selected using a random number table to which the last two digits of each student identification were compared. Teachers were apprised of the selection of students. To accommodate the concerns of the teachers over the absence of the treatment in the lives of the control group, although they did not consider the non-selected group as control, the researcher agreed to return in the spring to give the remaining students the treatment, TICM training.

During the experimental portion of the study, the researcher met with each group of twelve students during the same time of the day on each of four week days across a twelve week period. Students in the experimental group were extracted from the class and received twenty minutes of instruction on test item development. Students in the control group remained in their classes. Following instruction, all students, from both the experimental and control groups, were administered the Test Anxiety Inventory followed by the diagnostic TAAS test.

This design accounts for all sources of internal validity due to the random nature of the group assignment. With regard to external validity issues, the only concern might have been the Hawthorn Effect. Teachers, from whose classes the students were selected, could have chosen to apply the same techniques on other days to assure that all students succeeded on the post measure. The competitive nature of testing in high stakes testing situations exacerbates that possibility. The offer to meet the control group students through the spring of 1997 allayed the teachers' concerns and no attempt was made by teachers to duplicate the treatment among the control group students.

Procedure

Current practice for preparing students for the TAAS tests in the eighth grade promotes a confrontive approach to items on the test. Students drill in techniques of testwiseness, accumulating rules for successful testing. While this practice may lead to higher test scores, the question of the quality of learning and understanding remain unanswered. To move students from a plane of basic knowledge about test items to a higher level of understanding, an instructional program, based on the principles of the generative learning theory, was implemented with a group of forty-eight eighth grade students.

Wittrock (1991), in a discussion of the generative learning model of instruction suggested that teachers include four components when instructing during the use of this model. These include 1) establishing students' knowledge, perceptions, and preconceptions, 2) establish motivation for learning, 3) guiding students in attending to the processes of constructing meanings about the subject matter, and 4) combining students' perceptions, models of learning, and learning styles so as to guide students in the generation of the relationship between their prior experience and the new material being presented. Finally, in the method, students direct their thinking through metacognitive processing as they reflect on the cognitive and affective aspects of generating new constructs. These strategies were applied to the final outcome of students generating TAAS reading test items from content with which they were familiar.

The unit of study (TICM), designed to accomplish the outcome of creating test items, was divided into twelve instructional sessions. Each session was designed to accommodate the various aspects of the generative learning model. Throughout the process, the students incremented toward the objective through the use of inquiry, cooperative learning, and mastery learning models of instruction. Assessments at the end of each session provided the learner with documentation of his or her incremental accomplishment toward the overriding objective of test item development. These assessments were part of a portfolio that each student maintained throughout the project, to which they referred for guidance as they progressed toward the outcome.

Sessions were divided into sections that corresponded to the generative learning model component parts. Session I and II established student knowledge and perception of TAAS reading items and provided the initial motivation for participation in the instructional part of the project. These two goals were accomplished through open-ended discussions with the students about their perceptions of TAAS items, which they examined in detail. Students responded to these discussions verbally and in writing. Students were

encouraged to assist the researcher in the project based on the motivating factor of their possible contribution to the improvement of TAAS items, as seen from the perspective of an eighth grade test taker. Discussion regarding their personal betterment, resulting from participation, was also suggested.

During session II, students were given items from TAAS reading tests which they analyzed. Students were divided into three cooperative learning groups. While using inquiry, within cooperative learning structures, two groups of students sought to discover how words were used in creating stems and options for test items. A third group of students analyzed the components of a TAAS reading passage. During session III, representatives from each group shared their discoveries with members of the other groups.

Sessions IV, V, and VI were designed to develop the reading passage from which questions and subsequent answers were derived. Students determined that in every TAAS reading test there were three major parts: a reading passage, questions, and answers. During sessions IV - VI students developed a reading passage using a painting as a cue for the content of the passage. The painting, *Mountain Landscape with an Approaching Storm* (Vernet, 1775), depicts a scene at a river front where the fishermen and their families are working feverishly to finish their work before the storm comes in. In the background are a village, castle and forest between the fishing scene and the village. During session IV students reflected on the painting and verbally described the action, characters, and setting. They each took notes on group responses.

In session V and VI students expanded their thoughts about the painting. During session V students considered what could have occured before the characters entered the painting. Through guided discussion, the students generated possibilities for an introduction to the story. Students took written notes on the discussion. In the final session of story development, session VI, students discussed and generated possibilities for what could occur should the storm begin, essentially what would happen after the

painting, should it come to life. Students created an ending to the story, and took notes on the determinations of the group.

From the discussions of the four groups, and notes collected from student scribing, four distinct stories were created. The researcher, with permission from the students, wrote a TAAS passage that included the information garnered from each group of students' notes. Student editors corrected the passage and a final copy was issued to each student in the respective groups. This generated story served as the basis for question and answer development.

Based on their previous perceptions of items, combined with the created passage, students incrementally developed four test items. During session VII through XI the students were presented an example of a particular type of test item. Test item types included those that are tested by the TAAS exam which are: word meaning, supporting ideas, summarization, relationships and outcomes, inferences and generalizations, and point of view, fact and non-fact. The item came directly from the created story and included: a correct answer, three distracters, and two impossible answers. The complete unit of study can be found in the Appendix B.

The philosophy behind the proposed study stems from Aristotelian roots. Aristotle contends that people perceive things through a hierarchical development from sensation through memory into experience. From experience people can create art and develop an understanding of that art through scientific inquiry. The proposed method of artfully creating test items purports to reach a level of scientific understanding of items, thereby providing the artisans (students) with a comprehension of the phenomenon, testing. Consequently, logic would hold that a deeper understanding leads to greater success when confronted with the phenomenon.

Time Schedule

The project began at the onset of the new school year, 1996, which resumed in mid-August. The principal was re-contacted and meetings with teachers arranged to discuss the ramifications of the study regarding random selection of classes, the parameters of the instruction, space needs, and points of coordination to ensure the success of the project. The actual instructional time spanned twelve weeks during the months of September, October, November, and December. During these twelve weeks each experimental group was seen twelve times. Each meeting consisted of twenty minute sessions of instruction and occured at the same time each day for each class, during the thirty minute advisory period. At the end of the twelve week training period, all students were tested immediately, using the TAI and the diagnostic TAAS.

CHAPTER IV

RESULTS

Rationale for Analyses

Two multivariate analyses of covariance (MANCOVA) with eleven and seven dependent variables, respectively, and seven covariates were performed on the data. In the first analysis, the dependent variables were the total scores on the TAAS mid-year diagnostic, the six subtest scores, and the three parts of the test anxiety inventory. In the second analysis, the dependent variables were total mastery and mastery on the six subtests of the TAAS mid-year diagnostic (T197). The covariates were the total mastery scores and mastery on the six subtests of the TAAS fall diagnostic (T996). The independent variable consisted of training in the Test Item Construction Method which had two levels, training and no training. Univariate analyses for each of the separate dependent variables were also performed following the multivariate analyses. The multivariate analyses initially treated the dependent variables as single units, achievement and anxiety. The decision model used for all analyses was $\alpha \leq .05$, meaning that any obtained probabilities greater than .05 were considered non-significant.

The analysis procedure examines the dependent variables using the covariate as a mechanism for adjusting group differences in the dependent variable based on prior testing. Models for an analysis of variance and analysis of covariance demonstrate graphically the adjustment. $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ denotes the ANOVA model and shows that the only sources for error lie in the α(treatment difference) or the ε (within error) terms. $Y_{ij} = \mu + \alpha_j + \beta_{ij} + \varepsilon_{ij}$ represents the model for the ANCOVA and the addition of covariate β_{ij}, which examines the same sources of variability, provides an additional source of

accounting for variance, thereby reducing within error variance, a desirable condition for analysis of variance.

The MANCOVA analyses adjusted the mean scores on the mid-year TAAS diagnostic to account for differences in the groups as determined by the covariate, the fall TAAS diagnostic. These adjusted means were considered in subsequent analyses. The results of these analyses, using adjusted means, provide a more relevant picture of differences in the treatment effect. The contribution made by the covariate to the elimination of within error variance increases the variance in the main effect, thereby rendering the F test for the treatment as representative of true differences in the treatment.

The multivariate analysis of covariance has several advantages over single univariate tests. The MANCOVA permits a test of the potential interactions among the multiple critera. This is not feasible if each criterion variable is tested in isolation (Huck, 1974). The purpose of analysis of variance is to remove any unexplained within-group variability. By adding a covariate to the test, the within-group variability will be "reduced by an amount dependent on the strength of the relationship between the dependent variable and the covariate" (Maxwell, Delaney, 1990, p. 356). The covariate provides additional information about the subjects that exceeds the information garnered from analysis of the dependent variable alone. The covariate balances the groups under examination, which lends to greater clarity about actual findings.

To conduct the analyses, data were processed through the *Advanced Statistics Version of SPSS 6.1* (1994). The program examines the data entered and determines whether all procedures are necessary based on the linear dependency of the covariates. Determined by the initial linearity decision, all multivariate tests may be disregarded. In all cases, univariate tests are performed. Table 1 lists the descriptive statistics for the dependent variables T197 (test scores TAAS mid-year diagnostic), T197ROB (reading objectives mastered TAAS mid-year diagnostic), and ANX (anxiety measures).

Table 1

Descriptive Statistics for mid-year TAAS tests and anxiety inventory

	Experimental Group		Control Group	
Type of Measure	Mean	Standard Deviation	Mean	Standard Deviation
T197rraw = total score	36.85	10.94	35.84	10.95
T197SC1 = word meaning	6.77	2.23	6.85	2.20
T197SC2 = supporting ideas	4.95	2.13	4.87	2.13
T197SC3 = summarization	5.85	2.42	5.60	2.28
T197SC4 = relationships	3.69	1.60	3.60	1.67
T197SC5 = inference	8.92	2.77	8.80	2.53
T197SC6 = points of view	6.67	2.65	6.49	2.81
ANXTOT = anxiety total	45.98	4.87	46.84	13.17
ANXEM = anxiety emotion	17.90	4.87	17.53	5.34

(table continues)

	Experimental Group		Control Group	
Type of Measure	Mean	Standard Deviation	Mean	Standard Deviation
ANXW = anxiety worry	19.60	5.06	19.18	5.79
T197MST = mastery of all objectives	.375	.490	.409	.497
T197ROB1 = word meaning	.400	.496	.409	.497
T197ROB2 = supporting ideas	.475	.506	.386	.493
T197ROB3 = summary	.675	.474	.614	.493
T197ROB4 = relationships	.525	.506	.591	.497
T197ROB5 = inference	.350	.483	.432	.501
T197ROB6 = points of view	.350	.483	.455	.504

Note: These means and standard deviations represent the unadjusted means.

Hypothesis Testing

It was hypothesized that students receiving an alternative test preparation method, Test Item Construction Method (TICM), would score higher on the TAAS mid-year diagnostic test (1a), subtests (1c), and demonstrate a lower level of anxiety about testing (2). The multivariate analysis for these hypotheses was not performed due to linear dependency among the covariates. This indicates that there were no significant differences observed across the dependent variables, even when the covariates were considered. Univariate analyses of the effect of group membership are shown in Table 2. No significant differences between groups were observed.

Table 2

Effect Of Group Membership Univariate F-tests with (1,47 *df*)

Tests conducted with no covariates

Variable	Hypoth.SS	Error SS	Hypoth.MS	Error MS	F	Prob.F
T197RAW	4.24	3323.136	4.24	70.705	.060	.808
ANXEM	23.328	1286.859	23.328	27.379	.852	.361
ANXTOT	23.857	6198.318	23.857	131.879	.180	.673
ANXW	33.477	1409.110	33.477	29.9811	.116	.296
CONF	.488	38.264	.488	.814	.599	.443
T197SC1	.355	187.237	.355	3.98	.089	.767
T197SC2	.244	150.768	.244	3.20	.076	.784
T197SC3	.046	149.500	.046	3.18	.014	.904
T197SC4	.018	89.513	.018	.90	.009	.921
T197SC5	.136	191.768	.136	4.08	.033	.856
T197SC6	1.58	303.318	1.58	6.45	.245	.623

*$p < .05$

Tables 3-12 present the regression analyses for within-residual error term, individual univariate tests for the eleven dependent variables. In each analysis, except for subtest 2, different covariates contributed significantly to error reduction in the dependent

variables: total scores, subtest scores 1,3,4,5,6, and on all test anxiety scores. However, the contributions failed to eliminate within error variance enough to render the differences observed in the treatment significant. These are reported in order to ascertain which covariates contributed to error reduction and to what extent.

Table 3

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable Total scores - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RAW	-.445	-.439	.441-1	.011	.317
T996SC1	1.234	.262	.7681	.609	.114
T996SC2	-.780	-.161	.875	-.891	.377
T996SC3	2.764	.600	.8583	.223	.002*
T996SC4	2.153	.276	.1601	.857	.070
T996SC5	.961	.242	.8521	.129	.265
T996SC6	.919	.184	.8281	.111	.272

*$p < .05$

Covariate explanation: T996 designates the fall TAAS diagnostic for total score (RRAW) and subtests (SC1-SC6) which are defined in Table 1. Table 3 results indicate that T996SC3 (summary) significantly contributed to error reduction observed in the total scores for the T197RRAW (total scores, mid-year diagnostic).

Table 4

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable SC1 - reading subtest - word meaning (TAAS mid-year diagnostic)

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	-.045	-.208	.105	-.434	.666
T996RSC1	.423	.420	.1822	.327	.024*
T996SC2	.078	.076	.208	.379	.706
T996SC3	.310	.314	.2041	.523	.135
T996SC4	.441	.263	.2751	.602	.116
T996SC5	<.001	<.001	.202	.002	.998
T996SC6	-.262	-.245	.197	-1.337	.188

*$p < .05$

Table 4 results indicate that T996SC1 (word meaning) significantly contributed to error reduction observed in score 1 of the T197RRAW (total scores, mid-year diagnostic).

Table 5

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable SC2 - reading subtest - identify supporting ideas (TAAS mid-year diagnostic)

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	.034	.183	.094	.363	.718
T996RSC1	.005	.006	.164	.032	.974
T996SC2	-.267	-.301	.186	-1.43	.159
T996SC3	.291	.345	.183	.595	.117
T996SC4	-.002	-.001	.247	-.010	.992
T996SC5	.184	.253	.181	.016	.315
T996SC6	.118	.129	.176	.670	.506

*p < .05

Table 5 results indicate that no covariates contributed to error reduction in the subtest T996SC2 (summary of various texts). The means for the SC2 subtests were 4.86 for the experimental and 4.88 for the control groups respectively.

Table 6

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable SC3 - reading subtest - summarize a variety of texts (TAAS mid-year diagnostic)

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	-.185	-.918	.094	-1.97	.054
T996SC1	.226	.242	.163	.391	.171
T996SC2	.114	.119	.186	.619	.539
T996SC3	.611	.669	.182	.363	.002*
T996SC4	.275	.177	.246	.120	.268
T996SC5	.398	.506	.181	.208	.032*
T996SC6	.142	.143	.176	.809	.423

*p < .05

Table 6 results indicate that T996SC3 (summary) and T996SC5 (inference) significantly contributed to error reduction observed in the total scores for the SC3 (summarization).

Table 7

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable SC4 - reading subtest - perceive relationships and recognize outcomes (TAAS mid-year diagnostic)

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	-.030	-.217	.072	-.416	.680
T996SC1	.114	.177	.126	.906	.370
T996SC2	-.110	-.167	.144	-.766	.448
T996SC3	.285	.454	.141	.026	.048*
T996SC4	-.008	-.007	.190	-.042	.967
T996SC5	.075	.139	.140	.539	.592
T996SC6	.210	.309	.136	.548	.128

*$p < .05$

Table 7 results indicate that T996SC3 (summary) significantly contributed to error reduction observed in the total scores for the SC4 (relationships/outcomes, mid-year diagnostic).

Table 8

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable SC5 - reading subtest - analyze for inferences (TAAS mid-year diagnostic)

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RAW	.012	.053	.106	.122	.903
T996SC1	.123	.109	.184	.669	.507
T996SC2	-.364	-.315	.210	-1.73	.090
T996SC3	.399	.362	.206	.938	.059
T996SC4	.578	.309	.279	.075	.043*
T996SC5	.064	.068	.205	.317	.753
T996SC6	.323	.271	.199	.628	.110

*$p < .05$

Table 8 results indicate that T996SC4 (relationships/outcomes) significantly contributed to error reduction observed in the total scores for the SC5 (inference, mid-year diagnostic).

Table 9

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable SC6 - reading subtest - recognizing points of view (TAAS mid-year diagnostic)

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	-.232	-.836	.133	-1.744	.088
T996SC1	.341	.265	.2321	.474	.147
T996SC2	-.232	-.175	.264	-.878	.384
T996SC3	.866	.688	.259	.345	.002*
T996SC4	.869	.407	.3502	.480	.017*
T996SC5	.237	.219	.257	.924	.360
T996SC6	.387	.284	.250	.551	.128

*$p < .05$

Table 9 results indicate that T996SC3 (summary) and T996SC4 (relationships/outcomes) significantly contributed to error reduction observed in the total scores for the SC6 (recognition of points of view).

Table 10

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable ANXTOT - test anxiety - total score

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	.969	.901	.602	.609	.114
T996RSC1	-2.698	-.541	.048	-2.57	.013*
T996SC2	-1.392	-.272	.195	-1.165	.250
T996SC3	-.836	-.171	.171	-.714	.479
T996SC4	-.417	-.050	.584	-.263	.793
T996SC5	-.494	-.117	.163	-.425	.672
T996SC6	.050	.009	.131	.045	.964

*$p < .05$

Results reported in table 10 represent the univariate analysis of the total score on the test anxiety inventory. The total score represents the students' perception of their overall anxiety when taking a test. Indicators point to the RSC1 (word meaning) covariate as a contributor to error reduction observed in the experimental and control groups on their total test anxiety score.

Table 11

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable ANXEM - test anxiety - emotionality

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	.427	.862	.274	.558	.126
T996SC1	-1.342	-.584	.478	-2.811	.007*
T996SC2	-.663	-.280	.545	-1.218	.229
T996SC3	-.187	-.083	.534	-.351	.727
T996SC4	-.068	-.018	.722	-.095	.924
T996SC5	-.206	-.106	.530	-.389	.699
T996SC6	.010	.004	.515	.021	.983

*$p < .05$

Results reported in table 11 represent the univariate analysis of the emotionality score on the test anxiety inventory. The emotionality score represents the students' perception of their responses of the autonomic nervous system that are roused while taking tests. Indicators point to the RSC1 (word meaning) covariate as a contributor to error reduction observed in the experimental and control groups on their total test anxiety score.

Table 12

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate Analysis

Dependent variable ANXW - test anxiety - worry

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996RRAW	.34087	.67389	.2871	.187	.241
T996SC1	-1.04269	-.44470	.500	-2.086	.042*
T996SC2	-.76650	-.31828	.570	-1.345	.185
T996SC3	.05257	.02292	.559	.094	.925
T996SC4	-.39967	-.10285	.755	-.529	.599
T996SC5	-.13552	-.06856	.555	-.244	.808
T996SC6	.05673	.02280	.539	.105	.917

*$p < .05$

Results reported in Table 12 represent the univariate analysis of the emotionality score on the test anxiety inventory. The worry score represents the students' "concerns about the consequences of failure." (Speilberger, 1980, p. 5) Indicators point to the RSC1 (word meaning) covariate as a contributor to error reduction observed in the experimental and control groups on their total test anxiety score.

It was speculated in hypothesis 1c that students receiving alternative test preparation practice (TICM) would display a higher mastery level on the TAAS mid-year diagnostic than students who did not receive the TICM treatment. A multivariate analysis of covariance with seven criteria and seven covariates was performed on the data. The dependent variables were reading objective mastery and reading subtests objectives mastery scores for the mid-year TAAS diagnostic. Mastery was indicated by a score of 1 and non-mastery a score of 0. The covariates were the reading objectives mastered and reading subtests' objectives mastered for the TAAS fall diagnostic.

Table 13

Effect of Within+Residual Regression

Multivariate Tests of Significance

T197 ROB - reading objectives mastered - TAAS mid-year diagnostic

Test Name	Value	Approx. FHypoth.	DF	Error DF	Prob.
Pillais	1.430	1.944	49	371.00	<.001*
Hotellings	2.922	.70075	49	317.00	<.001*
Wilks	.1401	2.34393	49	243.03	<.001*

*p < .05

Table 13 presents the multivariate analysis of the effect of the covariates on reducing within-group error variance. Results of all three analyses indicate that there were significant contributions by covariates to within-group error reduction, but these analyses do not indicate which covariates are the significant contributors. Tables 14 - 20 present the individual univariate analyses of the dependent variables and the relationship to the covariates, performed to determine specifically which covariates contributed to error reduction. In all cases, except ROB 3 (summarization), covariates bore some significant contributions to error reduction.

Table 14

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate

Dependent variable: T197RMST - overall mastery of reading objectives - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996MST	.508	.424	.1972	.581	.013*
T996ROB1	-.027	-.019	.183	-.148	.883
T996ROB2	.096	.098	.146	.65	.513
T996ROB3	.016	.016	.137	.121	.904
T996ROB4	.268	.275	.1381	.942	.057
T996ROB5	-.143	-.131	.186	-.770	.444
T996ROB6	.015	.014	.144	.106	.916

*p < .05

Table explanation (same for all tables 14-20): T996 stands for TAAS September 1996. MST = total mastery, ROB 1 = mastery of objectives for word meaning, ROB 2 = mastery of objectives for identify supporting ideas, ROB 3 = mastery of objectives for summary, ROB 4 = mastery of objectives for relationships, ROB 5 = mastery of objectives for inference, ROB 6 = mastery of objectives for points of view. The results in Table 14 indicate that T996MST (overall mastery of objectives) from the fall TAAS diagnostic significantly contributed to error reduction observed in T197MST (overall mastery of objectives, mid-year diagnostic).

Table 15

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate

Dependent variable .. T197ROB1 - objectives mastered for word meaning - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996MST	.556	.478	.208	2.677.	.010*
T996ROB1	.063	.047	.193	.330	.743
T996ROB2	.142	.150	.154	.926	.359
T996ROB3	-.113	-.118	.144	-.786	.436
T996ROB4	.130	.137	.146	.893	.376
T996ROB5	-.431	-.406	.196	-2.202	.032*
T996ROB6	.016	.015	.152	.108	.914

*p < .05

The results in Table 15 indicate that T996MST (overall mastery) and ROB 5 (inference) from the fall TAAS diagnostic significantly contributed to error reduction of T197MST (overall mastery of objectives, mid-year diagnostic).

Table 16

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate

Dependent variable .. T197ROB2 - objectives mastered for supporting ideas - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996MST	-.139	-.115	.194	-.720	.475
T996ROB1	-.162	-.116	.144	-1.687	.097
T996ROB2	-.242	-.246	.144	-1.687	.097
T996ROB3	.153	.153	.135	1.141	.259
T996ROB4	.079	.080	.137	.580	.564
T996ROB5	.519	.469	.183	2.835	.006*
T996ROB6	.331	.309	.142	2.340	.023*

*p < .05

The results in Table 16 indicate that ROB6 (recognizing points of view) and ROB 5 (inference) from the fall TAAS diagnostic significantly contributed to error reduction of T197ROB2 (identify supporting ideas).

Table 17

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate

Dependent variable .. T197ROB3 - objectives mastered for summarization - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996MST	.245	.208	.1971	.242	.220
T996ROB1	.001	.001	.184	.008	.994
T996ROB2	.109	.114	.146	.749	.457
T996ROB3	.268	.277	.1371	.959	.055
T996ROB4	.168	.176	.1391	.215	.230
T996ROB5	-.079	-.073	.186	-.425	.672
T996ROB6	-.005	-.044	.144	-.035	.972

*$p < .05$

Table 17 results indicate that no significant differences were observed between the groups on the subtest ROB 3 (summary of various texts). No covariates contributed to any differences between the groups. The means for the ROB3 subtests were .687 for the experimental and .60 for the control groups respectively.

Table 18

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate

Dependent variable .. T197ROB4 - objectives mastered for percieved relationships and recognizing outcomes - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996MST	.046	.037	.198	.233	.817
T996ROB1	-.131	-.093	.185	-.714	.478
T996ROB2	-.053	-.053	.147	-.364	.717
T996ROB3	.300	.296	.138	2.177	.034*
T996ROB4	-.007	-.007	.139	-.052	.958
T996ROB5	.163	.146	.187	.876	.385
T996ROB6	.342	.316	.145	2.365	.022*

*$p < .05$

The results in Table 18 indicate that ROB6 (recognizing points of view) and ROB 3 (summary) from the fall TAAS diagnostic significantly contributed to error reduction of T197ROB4 (realtionships/outcomes).

Table 19

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate

Dependent variable .. T197ROB5 - objectives mastered for inference - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996MST	.315	.265	.1941	.623	.111
T996ROB1	-.258	-.188	.181	-1.426	.160
T996ROB2	-.144	-.149	.144	-.999	.322
T996ROB3	.212	.217	.135	1.572.	.122
T996ROB4	.445	.460	.137	3.255	.002*
T996ROB5	-.060	-.056	.183	-.331	.742
T996ROB6	-.013	-.012	.142	-.094	.925

*p < .05

The results in Table 19 indicate that ROB5 (inferences) from the fall TAAS diagnostic significantly contributed to error reduction of T197ROB4 (relationships and outcomes).

Table 20

Regression analysis for WITHIN+RESIDUAL error term

Individual Univariate

Dependent variable .. T197ROB6 - objectives mastered for recognizing point of view - TAAS mid-year diagnostic

COVARIATE	B	Beta	Std. Err.	t	Prob. t
T996MST	.260	.217	.214	1.217	.229
T996ROB1	-.253	-.183	.199	-1.275	.208
T996ROB2	-.053	-.054	.159	-.337	.737
T996ROB3	.112	.113	.149	.754	.454
T996ROB4	.370	.379	.1502	.461	.017*
T996ROB5	-.026	-.024	.202	-.133	.894
T996ROB6	-.161	-.152	.156	-1.033	.306

*p < .05

The results in Table 20 indicate that ROB (point of view) from the fall TAAS diagnostic significantly contributed to error reduction of T197ROB4 (relationships and outcomes).

A second MANCOVA, for the treatment effect, was performed on the reading objective data from the TAAS fall diagnostic, for the treatment effect. Table 20 presents the F tests for this analysis.

Table 21

Effect OF Group Participation

Multivariate Tests of Significance

Test Name	Value	Exact F Hypoth.	DF	Error DF	Prob. F
Pillais	.17522	1.42644	7	47	.217
Hotellings	.21245	1.42644	7	47	.217
Wilks	.82478	1.42644	7	47	.217

It is noted in Table 21 that there were no significant differences between the groups observed in the mastery objectives. Table 21 delineates the univariate F tests. There were no significant differences observed in mastery levels between the experimental and control groups.

Table 22

Effect Of Group Participation

Univariate F-tests with (1,53) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Prob.F
T197RMST	.04120	10.05544	.04120	.18973	.217	.643
T197ROB1	.07822	11.17927	.07822	.21093	.370	.545
T197ROB2	.61121	9.76918	.61121	.18432	3.315	.074
T197ROB3	.03016	10.08898	.03016	.19036	.158	.692
T197ROB4	.02426	10.19577	.02426	.19237	.126	.724
T197ROB5	.00278	9.80036	.00278	.18491	.015	.903
T197ROB6	.16441	11.85950	.16441	.22376	.734	.395

The results of the MANCOVA data analyses call for a rejection of the first set of hypotheses as there were no differences observed between the experimental and control groups. The experimental group did not score significantly higher on any of the mid-year TAAS diagnostic measures (total scores, subtest scores, or mastery levels on tests). The second hypothesis is also rejected. The experimental group did not score significantly lower on the test anxiety inventory. Based on these results it is noted that the Test Item Construction Method, as presented in this study, had no bearing on how students performed on state mandated (district interpreted) tests that have the same type of items on which subjects were trained to write during the treatment.

Other Considerations

Students participating in the study as experimental subjects produced four test items across the twelve weeks of training. The items developed aligned with the items tested by the state and were dictated by the Texas Education Agency TAAS Measurements Speficications for English Language Arts and Reading Objectives (1993). Students designed four questions (detail, sequence, comparison and contrast, and inference) that corresponded to the TEA/TAAS tested areas (identify supporting ideas in written texts, preceive relationships and recognize outcomes, and analyze information to make inferences, respectively.)

The student items were scored by an expert in tests and measurement who used the following criteria for scoring: relationship of question to the story, no imbedded tip-offs in the question, concise wording, correct answer accurate to the story, and plausible distracters. The rater awarded a maximum of two points for each of the five criteria. Consequently, the possible score for each item ranged from one to ten, or from four to forty for all four items.

Various statistical procedures were conducted on the student item score data. A reliability analysis of the rater's scoring, applying a Cronbach Alpha internal consistency measure, yielded a standardized item alpha of .8116, indicating a high level of consistency across items. Table 23 shows the means and standard deviations for the different items produced by the students in the experimental group.

Table 23

Student developed test items: Means and Standard Deviations

Test Item Developed	Mean	Standard Deviation
compare and contrast	6.56	1.98
detail	7.15	2.49
inference	6.43	2.62
sequence	7.62	1.99
total items written	69.87	15.89

On the average, students were able to develop the items successfully, thus displaying some measure of test writing skill.

The total score was calculated as a ratio of scored items to possible scores for items developed. Several students developed fewer than four items. Consequently, their total score reflected that ratio of their total score to the total possible for the number of items they completed. Correlations between the total score for items written and scores for individual items indicated a high level of agreement, with correlations of .60, .61, .77, and .77 for comparison and contrast, detail, inference, and sequence items respectively.

A multiple regression analysis was conducted on the data of student scores as related to the overall score on the TAAS mid-year diagnostic. The dependent variable, total score, was regressed against all students' item scores and total scores for item writing. The

correlations ranged from .375 to .458. A backward stepwise regression procedure yielded no significant contributions by the students' test item scores to the variance in the dependent variable. R square values ranged from .001 to .089 indicating no differences in the TAAS mid-year diagnostic accounted for by the independent variables, students' test items.

Further correlations were conducted between the subtests on the TAAS mid-year diagnostic and the students' developed test item scores. The correlations are listed in Table 24. There were no correlations that indicated a strength of relationship between items written and like test items.

Table 24

Correlation between student-developed test items and TAAS test items

TAAS Item Type	Student Item	Correlation
Identify supporting ideas	Detail	.098
Identify supporting ideas	Sequence	.182
Perceive relationships and recognize outcomes	Compare and contrast	-.1516
Analyze information and make inferences	Inference	.2803

Based on the multiple analyses of student test item writing as related to the TAAS mid-year diagnostic, it is noted that there appears to be no statistical relationship between the two. However, students' scores on developed test items indicated that eighth grade students can be taught to write TAAS test items with some level of proficiency.

A final consideration was student attendance at sessions. The mean student attendance was ten of the twelve sessions. The mean attendance for sections one through

four was: 9.8, 11.27, 11.17, and 9, respectively. Correlations between attendance and TAAS mid-year diagnostic scores, and attendance and student test items were calculated. The results are noted in Table 25. There were no significant correlations between attendance and test scores or test item development scores.

Table 25

Correlations between attendance and test scores and test item development scores

Item correlated with attendance	Correlation
T197SC1	.199
T197SC2	-.088
T197SC3	-.010
T197SC4	.107
T197SC5	.049
T197SC6	-.197
T197RRAW	.001
compare and contrast	.2265
detail	.0457
inference	.2456
sequence	.1212

Summary

The results of the data analyses indicate that there were no statistical differences observed between groups on any measures applied. In 86% of the means examined, the experimental group maintained higher adjusted means. However, these higher means did not account for significant differences between the groups. Students in the experimental group may have received some benefit from the treatment, but this was undetected by the

measurement tools used in this study. The study was designed to determine whether training students to write test items would make them more successful at test taking than their non-trained counterparts. The findings in this study did not support any of the hypotheses proposed.

CHAPTER V

DISCUSSION

The present study considered the effect of training in test item writing on eighth grade students' test taking ability. The following discussion examines the results of the study and illuminates the findings. The discussion includes deliberations of the assumptions and limitations of the study that offer germane information. Additional material, garnered from students' responses during training, provide different insights into the effect of the method. Implications for future research and the value of the findings to the enhancement of the body of knowledge in test preparation methodology close the discourse.

The MANCOVA procedure applied to the total test scores, subtest scores, objectives mastered, and anxiety inventories of the students in the study found no significant differences between the groups on any measure. However, among several covariate/dependent relationships, patterns emerged. In one example, the word meaning covariate (T996SC1) significantly contributed to the within error reduction in four of the eleven categories examined in the analysis. The dependent variables influenced by that covariate included the three anxiety components (total, emotional, worry) and the word meaning dependent variable (T97SC1). The relationship between the two word meaning subtests indicated testing alignment between parallel forms of the test. The alignment between the subtest and the anxiety measures, however, merely indicated a relationship between the two. Rendering an explanation regarding their relationship, would border on speculation and extend beyond the realm of this report.

A similar covariate/dependent variable relationship became apparent in another incident. In the first, the T996SC3 (summarize a variety of written texts) significantly contributed to the within error variance of the dependent variables T197SC3 (summarization), T197SC4 (relationships), T197SC5 (inference), T197SC6 (recognize point of view), and T197RRAW (total score). In spite of the limited nature of the relationship, one that did not contribute to the difference in treatment effect, T996SC3 stands out as a contributor to within error reduction.

An observation of the means and adjusted means of the dependent measures reveals interesting information about the groups. In all instances, the experimental group maintained higher means than the control group, although not significantly higher at the .05 level. The experimental group's display of higher means indicates the potential for raising the scores within the experimental group, possibly following an alternative treatment plan. A more extensive presentation of alternatives follows in a subsequent section of this chapter.

Higher means among the experimental subjects, on the emotionality and worry scales of the anxiety inventory, supports earlier research (Tunks, 1996) involving students who successfully tested in the school. They, too, displayed concern over their ability to pass the TAAS test and noted their autonomic reactions in a TAAS testing environment. Likewise, students in the current study, experimental group, when asked during the treatment phase of the study, to consider what they knew about TAAS tests, indicated primarily worry and limited emotionality.

Concerns about testing, expressed on the KWL charts used during the training, provided students with the opportunity to reflect on the multiple aspects of testing. A KWL chart divides perception about a phenomenon into what students know, want to know, and have learned. Students completed different sections of the chart throughout the TICM training. The following statements appeared repeatedly throughout the initial responses in

the K section noted on the KWL charts that experimental students completed during treatment.

They write long passages to get you bored; that is one of their tricks.

They are tricky on purpose so that you fail.

You have to pass this test, or you don't go onto high school.

This test is important, and we're required to take it.

The KWL charts provide a plethora of information attesting to changes in perception regarding the TAAS test. A content analysis of the charts suggests six categories of statements that include statements or words dealing with: content, expectations, fear, personal choice, attitudes, and perception of the test. Content statements related to comments about the component parts of tests such as questions and answers. Expectation statements refer to things student expect to have to do, either while taking the test or writing one. Statements of fear relate to students' concerns over passing the test. An interesting category, personal choice, emerged with statements about the students' choice to become test writers. Students' attitudes toward the test include statements about their affective judgmental reactions to the test and to the process of writing tests. Finally, students' perception of test taking and writing indicated their understanding of testing and the process of writing a TAAS test.

The "K" section of the KWL chart, completed during the first session, asked students to state what they knew about TAAS test taking. Responses fell within each of the six categories, but attitude toward testing dominated the responses. Overwhelmingly, students found the TAAS test boring, hard, long, tiring, deceptive, and time consuming. In general, their attitude toward the test reflected an affective negative proclivity. Despite this, a number of students noted content aspects of the test, where the structure of the test entered discussions. However, the minimal mention of the expectations of testing that came forth included comments about the requirements -- primarily reading, answering

questions, and preparation. Clarity of understanding of the intent of the test eluded most. One student suggested that the TAAS test challenged you and helped determine what you knew. The statements regarding fear appeared previously.

In contrast, statements and words examined in the "L" (what did you learn about writing TAAS tests) indicated a shift in thinking among students. While some statements regarding an affective reaction to TAAS testing such as, "writing tests is boring", "this is really hard", their appearance paled in comparison to the gain in statements regarding content. Seventy-five percent more statements regarding content emerged during reflective discussions in the final session. Student portfolios burgeoned with specifics that related directly to what they studied across the twelve sessions. Samples of common statements included:

We learned how to write different types of questions.

We learned how to find important information in the paragraphs of the story.

We learned how to make a story from a painting.

We learned how to predict what will happen in a story and write a question about it.

We learned how to use context clues to find answers.

We learned that you have to know the level of test you are writing.

We learned how to find possible answers in the story.

We learned how to put questions in categories.

We learned how to brainstorm in order to write tests.

We learned how to formulate questions.

We learned the importance of characters in stories.

We learned that questions, answers, and stories go together.

We learned the you have to go step by step to write a test.

We learned that you have to go back in the story sometime to find a question.

We learned that to make a test, you have to write alot.

We learned that you have to look for details to develop questions and answers.

The other categories provided insights into what students thought and believed about test writing. Personally, several students mentioned that they would not choose test writing as a profession, partly because you might have to work with people you don't like. This handful of students noted that test writers, "must be nerds." Also on the personal side, one student commented that, "Learning to write TAAS tests could help in taking TAAS tests later." Interestingly, mention of fear of testing disappeared in the discussions during the "L" contemplations. A number of students commented on the amount of time required to write a TAAS test.

A common theme within the perception category pertained to thinking. Students noted that making tests required thinking on many levels. The following statements support the positive connotation about thinking observed in the "L" section of the charts.

To write tests means that you have to use higher learning and knowledge.

You need different skills to write tests than to take them.

Thinking is very important when writing TAAS tests.

You have to imagine the action when you write the story.

When you look at the painting, you have to draw inferences about the characters and the action.

You have to think a lot when you write TAAS tests.

You have to use your imagination when you write the story.

You have to be a good thinker when you write TAAS tests.

These reflections on the work in their portfolios imply that students changed in their understanding of the TAAS test across time.

This finding supports Wittrock's (1991) Generative Learning Theory (GLT). He noted that motivated learners, inspired by their own reconstruction of a concept, more clearly comprehend the phenomenon. The exercise of mindful abstraction, regarding the

perception of what students learned about TAAS test writing, prompted students to generate their own thoughts and move from an affective/judgmental perspective to a higher cognitive plane. From an Aristotelian philosophic perspective, the students' understanding of testing and the processes required to create tests, elevated from sensation, (gut reactions to testing) to art, (knowledge of the cause) and finally to a scientific level (understanding).

Wittrock (1992) suggests that student engagement in the process of developing understanding motivates learners to generate their own learning. A further examination of the portfolios supports this supposition. Available learning time includes all the learning time available for an instructional session (Stallings, 1975). The available time divides into three segments: allocated time, engaged time, and academic learning time. The most important to student learning, academic learning time, represents the time when students display on task academic behaviors relevant to the academic outcome (Romberg, 1980). Evidence of academic learning time manifests in student engagement and productivity. An additional benefit of engaged active learning, minimal discipline and high levels of achievement.

A content analysis of students' portfolios intimated that students actively participated in the learning process at a high level. The "think" sheets provided at each session required students to engage in conversations either with the researcher or fellow students in cooperative learning groups and resulted in completed sections of the "think" sheets. An examination of the completed sheets indicates that students actively participated in the process. The overwhelming majority of the sheets flourished with details generated from conversations regarding test item development. Very few of the completed sheets reflected passive participation.

Active participation motivates learners to engage in learning. During the development of the story (a reading passage for the student-developed TAAS item), students participated in an inquiry method of instruction that evoked student contributions

to the story. In their ruminations of what they learned, students commented on numerous occasions how the process of developing a story from a painting prompted their thinking. Their creation of the story from their imagination with a single painting as a prompt gave the students ownership and greater understanding of the story from which they drew question and answer items.

During the development of questions and answers, students showed high levels of motivation. Their discussions of the content of particular paragraphs, used during their deliberations, stemmed partly from their familiarity with the story from which they derived the items. A second motivational factor involved the use of cooperative learning structures (Kagan, 1994). During deliberation sessions, students operated in groups of three or four. The structures placed students in the position of contributing to the process. The use of manipulatives during the structured group time provided students with concrete representations of ideas and members' additions to the thought pool. An assumption put forth in this study suggested that students' own experiences with the process and the materials provide the impetus for test item generation.

The products that resulted from TICM supported another assumption of this study -- that junior high students' ability to create test items exists. Experimental students exhibited fairly high levels of understanding of test item development as attested by the mean mastery level of 70%. This level of mastery indicates that students understood the inner workings of test items to create cogent test items on their own. The use of mastery instruction methodology provided the grounding students required to observe, practice, and independently attempt test item writing.

An examination of students' portfolios indicates that students' experiences and generative learning processes aligned with Wittrock's Generative Learning Model application which Wittrock (1992) and Kourilsky (1992) successfully demonstrated differences in test scores among experimental groups. However, the experimental

students in this study exhibited no significant differences from their control group counterparts on standardized measures. To account for this discrepancy, a closer look at the limitations anticipated bears reconsideration.

As stated previously, time represents a limitation of the study. Though other limitations exist and serve to limit the generalizability of the study, time presents the most poignant explanation for the conundrum. Time, regarding this study, divides into several descriptors. These include the time of day, amount of time, the timing of the test, and timing of the study. Each played a possible role in the lack of differences observed between groups.

Students met with the researcher during the first period (advisory) in the morning. This period provided teachers with thirty minutes to get organized for their day. During this time students generally complete homework assignments, chat with friends, or sleep until first period. This time served as the meeting time for school organizations such as student council. However, once weekly students relinquished this time to meet with the researcher to develop TAAS test items. In general, the overall attitude toward participation, seemingly disdain, generated mostly due to departure from their friends in a non-structured setting to a structured setting to be with others whom they may not choose to affiliate.

In addition, to the time of day, first period in the morning, other factors affected that time and consequently the amount of time allotted for instruction. The advisory period provided the principal and others who needed to impart vital information and an opportunity to speak to the entire school via the public address system. Some days, particularly when a gang related problem needed addressing, the time to work on test item development diminished to as little as 15 minutes. This minimal amount of time left the researcher and students in a position of moving quickly through ideas and materials, and possibly disserviced students.

Essentially, the instructional time averaged 20 minutes. When calculated across 12 sessions, students received a total of four hours of instruction across an entire semester. Twenty minutes a session disallowed time to teach for transfer, another important aspect of the generative learning theory. To establish a relationship between writing TAAS test items and taking the TAAS test, mandates teaching for high-road transfer (Fogarty et al, 1992). Students attended regularly and attentively completed instructional activities planned, but received no guidance in transfer, due to the limited amount of time allotted. Any attempts to recap or recapture thoughts from a previous week became difficult because of the week long (sometimes two) time lapse between sessions.

Another aspect of time, timing of the TAAS diagnostic test, possibly contributed to the lack of difference between groups. Instruction in the TICM terminated in mid-December, 1996. Due to delays from the district testing department, the school waited until the last week in January to administer the test. The seven weeks between the termination of training and the administration of the test suggests that the timing distance between the two possibly incited decay. With no high-road means to relate the two, training and testing, students reverted to their original habits of test taking.

The final "time" consideration, timing of the study, played an interesting role in the process. In a high-stakes testing environment, as the test draws near, the anxiety level elevates and student awareness heightens (Tunks, 1966). The training in this study occurred during a lull period in TAAS test preparation. The school lacked the enthusiasm for testing observed during the background study, which occurred in the spring two weeks prior to the administration of the spring 1996 TAAS test. Experimental students' recollection of TAAS testing in seventh grade focused on affective/feeling aspects of the test (boring, hard, long, tiring). Their motivation for the 1997 TAAS test administered in the late spring of their eighth grade year failed to garner their attention during the fall.

Implications

However, the TICM solicited their attention in so far as students produced viable test items based on a story they generated from an external source, a painting. This finding leads to the implication that junior high students in high-stakes testing environments possess the ability to respond to alternative forms of test preparation. Shepard (1990) called for alternative methods of test preparation that exceeded the limits of drilling testwiseness. The TICM approaches testing knowledge and understanding from a generative theoretical base and places the learner in the position of viewing the test from an insider's perspective, as opposed to a confrontive view, afforded by testwiseness and/or measurement driven instructional methodology. Consequently, TICM answers Shepard's charge and provides what she calls "higher order" thinking in test preparation (1991).

A second implication for this study emanates from higher order thinking that resulted from the process. Students' responses on the KWL charts imply that, across time and through experience, they gained in ability to think on a higher plane about TAAS testing. Their comments changed dramatically from affective musings to resourceful statements that implied a higher level of thinking about the phenomenon, testing. The implication suggests that using TICM fosters higher order thinking about testing. Training in higher order thinking provides learners with extended perceptions and relationship building capabilities not witnessed in lower order drills (Shepard, 1991).

Essentially, TICM provides tools for life-long learning as opposed to learning the tricks to pass the test. Life-long learning encompasses the ability of learners to visualize and extend beyond the quest for the one "right answer" (Hyerle, 1996). Students using the TICM experimented with possibilities for answers and generated plausible and implausible story lines, test items, and suppositions about test writing. The generative processing used in TICM purports to engender skills required in life-long learning that include decision making, self-assessment, realignment, and interconnectedness. Students in TICM training

displayed those skills. Therefore, the implication that participation in TICM instruction promotes greater understanding of the phenomenon testing, became apparent in this study.

Contributions to the Literature

The findings in the study contribute to the field of test preparation inquiry for several reasons. First, the method examined presents a distinct approach to test preparation, differing from current test preparation methods that focus primarily on testwiseness and/or measurement driven instruction. TICM uses a learning model (GLM) that employs constructivist techniques, thereby placing the responsibility for learning in the hands of the student. This learner-centered method encourages life-long learning.

A second contribution to the field answers the charge for alternative methods for test preparation that encourage higher-order thinking. The findings from this study demonstrate that students engaged in TICM increase their level of thinking across time. This valuable piece of information provides the field with a demonstration of student thinking about testing that extends into an entirely new direction.

The study examined the use of TICM in an inner-city, predominately Hispanic, low SES, low achieving junior high school and determined that students of this description met the qualifications as beginning test writers. This third contribution to the field of study provides an example of a challenged learning environment in which students, entrusted with the responsibility of their learning, generated test items that reflected mastery. This finding corroborates Wittrock's tenets of the generative learning theory that purports understanding a phenomenon through generative processing without regard to learning conditions or student status.

Finally, the study demonstrates the use of an experimental design in an intact school setting. Understanding differences in treatment bear serious consideration. The field of test preparation study mandates elevated studies that press the issue of verifiable differences

among tested populations. This study provides a model for design and analysis that fulfills that mandate. Consequently, the method of research contributes to the field as well.

Recommendations

To test the efficacy of TICM as a viable method in effecting change in standardized test scores, several recommendations for further study emerge. First, in a future study of the method, a resolute commitment to providing ample amount and placement of time, warrants consideration. Replication of the study across a semester seems logical, but might better serve students if the preparation time juxtaposed to a realistic testing situation. In addition, longer sessions for training may provide students with the time needed to ponder the weightiness of test item development and provide the opportunity to create more sample items. Replication with different populations might produce different results as well.

A second consideration for future study stems from the option to teach the method and incorporate transfer techniques. Bridging techniques (Fogarty et al, 1992) include anticipating applications, generalizing concepts, using analogies, parallel problem solving, and metacognitive reflection. The current study incorporated metacognitive reflection in the final step of the process. Future studies that embody more bridging techniques, concomitant with the TICM, may bring about significant changes in student test performance.

This study used cooperative, inquiry, and mastery learning models for instruction. Consequently, student work displayed in portfolios reflected group work. A third possible study generates from an examination of instructional model and the related effectiveness in promoting test item development skills among individuals. Cooperative structures and discovery aspects of inquiry learning provided learners with opportunities to contribute to group idea building and consensus on item construction. However, the materials in the portfolios offered a somewhat tainted view of individual student effort. Other methods of

recording and reporting students' understanding and skill level in test item development warrant further deliberation and study.

A fourth possible platform for research lies in the manipulation of TICM across different content areas other than reading. Statewide testing encompasses math, social studies, and science. TICM structure allows for exploration into other content areas possibly requiring minor adjustments to accommodate specifications for testing in other areas. Research of the use of TICM across disciplines opens a vista of possibilities.

Finally, to discern the feasibility of TICM as a viable method, necessitates teachers applying the method in authentic settings. Unfortunately, this form of research generally precludes the option for experimental design. Current experience in this study demonstrates that teachers care not to divide their classes into experimental and control groups, without assurances of equal treatment later. However, a static group design that employs covariates in the analysis, increases the potential for attainment of finding differences, albeit ungeneralizable.

No matter the generalizability, validity, or reliability of qualified research, the reading public currently operates on a crisis foundation regarding test scores. Their sources of information, the daily newspaper, radio, and television, characterize schools, districts, states, and nations as worthy based on reports of students' scores on standardized measures. Schools, accountable for meeting the mark and raising the bar, accept a high-stakes testing status and drive their students to higher scores and greater financial recognition. The reading public accepts this measure of success and status as "exemplary" schools receive their rewards without recognizing the undercurrent of reduction in curriculum and instruction required to achieve these levels of success.

Educational researchers, challenged by the reduction of education to "teaching for the test", must continue to examine and explore the possibilities for working within the current system for the enhancement of student learning. TICM presents a model for

incorporating sound instructional models that align with state mandated curriculum and testing. It challenges students to consider testing through generative processes that require higher order thinking. The findings of this study bear consideration for future study of both the method and the implication for students as not only test-takers but life-long learners.

APPENDIX A

The Background Study

The background study was conducted in the spring of 1996, at a Junior High School in North Texas. The school personnel and students were in the process of preparing for the TAAS test which was to be administered the week following the two weeks of observations. The research consisted of an qualitative research study that employed ethnographic methodology. The techniques used in the study included observation, interview, and documentation. Four teachers, referred to as 1w, 1f, 2w, and 3m in the report, labels used for anonymity purposes, permitted the researcher to observe each classroom for one full day of instruction. Teachers 1w and 1f classrooms were housed in the main building, while teachers 2w and 3m taught in portable buildings on the grounds of the school.

Interviews

Following each day of observation, each teacher provided the researcher with an interview. Each teacher was asked to "walk the interviewer" through their approach to preparing students for the TAAS. Teachers freely described their strategies and application of district recommended approaches to preparation practice. The researcher used a set of questions that served as guidelines in maintaining consistent coverage of information across all interviewees. The questions pertained to the teachers' strategies, reasons for selecting the strategies, their beliefs about the effectiveness of the strategies for student achievement on the tests, and the value of the process of preparation to overall learning for the students.

Students were interviewed at the end of the observations both individually and in groups. Students discussed their attitudes and beliefs about the TAAS tests and preparation techniques. Individual student interviews consisted of an emic style of questioning allowing the student to freely express their beliefs. Students reflected on preparation practice, the value of such practice for preparation, their beliefs about the effectiveness of

preparation in preparing them for the TAAS tests, and the value of the TAAS tests as a measure of their knowledge and skills as eighth graders.

Following individual interviews, students were asked to join in a group discussion about testing. The structured group interview examined the attitudes of eighth grade students toward their efficacy as test evaluators and writers of test items. Students discussed, through a structured interview technique, alternatives to test preparation that they would consider to be more effective than those previously experienced. Students suggested ways a teacher might approach test preparation differently and still be effective in preparing students. Students proffered other forms of assessment that might provide teachers, the district, parents, and the state with needed information.

As a final part of the discussion, the group broached the idea of test construction. Students explored inherent components of test items for the TAAS. Students and researcher explored, in open discourse, how eighth grade students perceive the process of construction of items for such a test. Finally, the students advised the researcher as to the best approach a teacher might to take when teaching students how to write test items.

Data Analysis

Class observations, interviews, and environmental observation yielded interesting findings. Several categories relating to test preparation and attitudes toward testing emerged. These categories included: 1) evidence of curriculum related to TAAS objectives, 2) evidence of teaching practice or instruction related to TAAS, 3) evidence of teacher discourse to students related to TAAS, 4) evidence of teacher discourse with students related to teacher's attitude toward TAAS, 5) evidence of teacher attitude displayed in meetings, informal conversations, or interviews, 6) evidence of student attitude displayed during instruction, 7) evidence of student attitude displayed in group or individual

interviews and informal comments, and finally, 8) official evidence from school documents of attitudes toward TAAS, and official school curricular support and training.

Signs in the School

Walking through the halls of the school an outsider is struck by the plethora of TAAS excellence encouragement. Across the entire main building there were signs every four feet along all walls except in the arts academy. The signs varied in color, size, quality, wording, and art elements, but carry a single message, "Do well on the TAAS test." A clever sign maker had taken the acronym TAAS and configured it to represent the words **T**ake **A**im **A**t **S**uccess. This saying was represented with arrows flying toward targets, with 100% written across the arrow head. Other versions used the take aim motto and represented the outcome as shooting stars. Another set of small posters read "You can pass the TAAS." The verbiage on the signs made clear the intent, the development of positive mental attitudes.

It is unknown whether the sinage had any impact on the students or teachers. At no time was the graphic display of encouragement mentioned by anyone during interviews. Explanations for the lack of discussion of the sinage may stem from the limited amount of time students at the school have to contemplate the messages in the signs. Possibly the students were overwhelmed by the repetitiveness of the signs that the signs are considered part of the natural environment, warranting no particular comment. This finding is interesting considering the abundance of signs.

Notably, these signs urging TAAS success were posted only in the main building. Some students spend all their time in the portables with the exception of science class and lunch. Many students were observed during lunch hours standing on the blacktop outside the portables and never entered the main building for lunch. In essence, the signs of encouragement were for those who took classes in the building. Having observed only

parts of two instructional teams, it was difficult to determine whether this lack of TAAS sign encouragement in the portables had an effect on students' attitudes toward the test.

A Tale of Two Cities

An unexpected observation produced "A Tale of Two Cities". There were some subtle differences between students attending classes in the building and the portables, during between class exchanges. It was noted that students exchanged classes in the building with less excitement than the students in the portables. Teachers in the building stood outside their rooms during class exchange and encouraged students to hurry on to class and kept a watchful eye on the students who entered. Students tended to sit quietly until class began. In contrast, the portables with the multiple entrances between classes allowed the students to wander freely across the quad portable (a portable with four rooms joined by doors at the adjoining points of the four rooms). During the four minute exchange time, there was a fair amount of interaction among the students. Teachers in the portables went about preparing for the next class until the tardy bell rang when they took role.

The nature of the design of the portables appeared to contribute the type of exchanges observed among the students. It was impossible for the teacher to stand at multiple entrances to greet students. Students could enter and depart from multiple doors resulting in increased opportunity for verbal exchange and camaraderie among the students exchanging classes within the portable environment. Students coming from across the main building science, physical education, arts, or computer classes, were faced with a challenge to be on time in the classes in the portables. Consequently, in the portables, tardiness was more evident than in the main building classes.

Analysis of Instructional Approach

The main theme throughout the school, success on the TAAS, was present in all interviews, observations, and documentation. The administration provided teachers tools to produce successful test takers through mandated structures such as the metacognitive days, timelines for TAAS, teaming, diagnostic profiling, and incremental testing throughout the year. Metacognitive days were days that teachers were required to schedule weekly where they took the students through a series of seven steps that encouraged the student to reflect on a specific element of a content area for a forty-five minute period. These "metacognating days", as the teachers called them, were based in Socratic questioning techniques. The seven steps were listed on the walls of three of the rooms where observations were conducted. Through interview and observation, the researcher learned how different teachers and students dealt with the drive for success on the TAAS test.

The two teachers in team A, both young and new to the system, approached the challenge head-on. Their classrooms represented models of measurement-driven instruction (MDI). According to Popham (1991), MDI is an appropriate approach in a high-stakes testing setting where the measure is criterion-referenced, clearly delineated in objective form, and clearly articulated to the teachers. Bracy (1991) agreed that while MDI enhances higher scores, learning and the expanse of the curriculum is jeopardized. These two teachers in team A strove throughout the year to provide their students with examples of test items and item structuring their students needed to succeed. They both spoke confidently about what they had done, and their only concern was for the length of the test as related to student fatigue.

The language arts teacher in team A explored the use of highlighters as a technique in guided reading of test items. Both team A teachers trained their students in the use of highlighters for marking salient points in test item reading. While observing class 2w it

was noted that students tended to highlight everything including the illustrations and page numbers. They even created borders for their work. The implication here is that the students were both unaware of the purpose of the highlighter and possibly sought some creative outlet. Students in 3m, however, had progressed to the point of understanding the purpose of highlighting. Their daily work papers showed marked progress in TAAS practice work across time following the introduction and understanding of the highlighter technique.

In contrast, the teachers from team Bused no testwiseness techniques with their students. Neither were observed in the practice of coaching. In four out of five class sessions observed, teacher 1w lectured on content and encouraged discussion about the topic. Students took notes and interacted with the teacher, engaging in lengthy discourse about the ramifications of the decisions made by the persons in the topic under study. Likewise, in 1f students were asked to engage in a short discussion of a film they had watched the day before and then were assigned the task of writing a reaction. The students in these classes were generally higher functioning than the ones observed in team A. It could be inferred here that higher level students require less drill and practice prior to TAAS testing.

A pattern in preparation seemed apparent in all settings. In classes of lower-ability students, drill and practice constituted the method of choice. This instructional style evidenced even 1w. The overhead the centerpiece, and worksheets were the student focus. In classes of higher ability students, teachers taught more openly and operated on the premise that these students had the testwise skills needed. Even in the case of the drilling team A, the higher ability students exhibited more freedom to work on their own and engaged in more detailed discussions. The lower-ability students seemed dependent on the teacher for guidance in thought. They would wait for prompting before answering, even if they knew the answer. Student papers on the walls in all the rooms were for the "one

hundred club only." In other words, only student papers demonstrating 100% accuracy on a TAAS exercise (worksheet) gained the status of display with other one-hundred club members. The less confident, lower ability students' work remained in the grade book and student folders. The general impression was that students in lower ability groups have accepted their status and deal with the pressure through either total silence, avoidance, or discomfort when answering a question.

Students in both teams made a concerted effort to come up to the expectations of the teachers. Students interviewed admitted that there was pressure to succeed from the school administration, teachers, parents, and themselves. They had their eyes set on particular high schools and did not consider summer school or repetition of eighth grade a reasonable option. They were willing to accept whatever pressure was necessary to get through the tests successfully. The students in team A were not interviewed but seemed to listen carefully when the teachers emphasized the importance of being ready.

Independent of the method selected and no matter the ability level of the student, teachers expressed concern for their students' academic welfare. They made constant reference to the schedule of the test and its impact on student concentration ability across so many days and tests. They questioned whether they had given the students everything they needed to be ready for the test. Their anxiety attested to their commitment to the students and the profession. The students at the school were perhaps on the brink of decision about gangs, pregnancy, high school, jobs at McDonalds or McDonald Douglas, and many other issues. The teachers saw their role as the champions of knowledge and learning. They fervently hoped that students would thrive academically.

Student Thoughts

Students interviewed were from team B and enrolled in the advanced drama section in the arts academy. In interviews, several students embodied the teachers' drive for

academic excellence. Their descriptions of preparation practice aligned with classroom observations and teacher comments. Students confirmed that TAAS preparation had gone on all year. They affirmed that teacher 1w did not drill, but had high expectations. They indicated an appreciation this teacher's concern for their educational welfare and welcomed the challenges of class 1w. Students implied that they understood the need to work hard now so that could succeed later in life. Several gave examples of relatives who were employed at minimum wage because of lack of initiative in school and vowed to take a different path.

The overwhelming consideration for the students centered on the fear of summer school attendance. Summer school implied failure, an experience uncommon to the students interviewed. Most of them had auditioned or applied into magnet high schools and had been accepted. Although attendance at summer school would not preclude their admission to these schools in the fall, the students did not want the stigma of a failure. In their young artists' minds there were set backs in life, but not true failures. Their zeal for excellence pushed them forward to succeed.

When questioned about how they dealt with the anxiety of failure, they noted three approaches. One group denied the pressure existed, but in the same breath they stated their fear of summer school. They followed these statements about the test being easy and something they could do with no work. A second group was grateful to have the test preparation training the teachers provided. One student, very academically capable, stated that she didn't know how she could pass the test without all the work the teachers gave her. The final group just chose to be anxious. These students talked about how they were nervous for all tests, even though they do well. One student reported that she worried all the time and that there wasn't anything she could do about it.

Group Structured Interview (Students)

In the focus group discussion, students offered alternative ways to assess learning and understanding. Among their suggestions was hands on learning. Notably, during class observations, students who were physically involved in the learning process, showed greater on- task behavior. In the math class 2w, students sat with worksheets in front of them but were not allowed to mark on them. Disruptions to the drill, which included talking across the room and laughter outbursts, required the teacher to stop the lesson, call students down, and bring order back to the room several times. The higher functioning math class caused enough of a ruches and resistance that the teacher acquiesced and allowed students to work alone. They were doing something physical. Although teacher 3m drilled, the students had flash cards with choices. The physical aspect of responding with flashcards seemed to be a critical aspect for these students.

Summary

In summary, the teachers and students in the school deal with the need to succeed through various means. All of the means lead to the end of preparing for the TAAS tests. Each person had to cope with the multiple pressures of high-stakes testing. The consequences of failure comprise embarrassment to the school in the media, teacher tiering by the district, summer school attendance for the students, and upset for parents with children returning to the eighth grade.

The school administration has an enormous challenge to guide approximately two thousand students to academic excellence as measured by a single instrument administered annually. Observations, interviews, and documentation of the final two weeks before the administration of the TAAS test indicate that the administration has taken a pro-active role in the forwarding of the academic success of its student population. It was obvious that the teachers observed understood the level of functioning and academic needs of their students and strove to provide them with appropriate instruction in preparation for the TAAS test.

Students, in all observed settings, responded positively to the execution of the recommended curriculum.

This study reflects a limited perception of a complex situation. It is recommended that a longitudinal observational study of numerous teachers over an extended period of time be conducted to more clearly determine the patterns of instruction as related to student ability in a testing situation. In addition, it would be interesting to explore in more detail the "tale of two cities", the differences between life in the portables and the building with regard to student and teacher attitude toward testing, instruction, and learning. Test scores provide one view of a student, teacher, school, and district. Through in depth observation, documentation, and interview, a clearer picture emerges and provides the needed information to validate the instructional and administrative practices applied in a high-stakes testing environment.

APPENDIX B

Instructional Unit

The Test Item Construction Method

The following format was followed when presenting the instruction of constructing TAAS test items.

- **Session I**
  - Outcome - Establish students' knowledge, perceptions, and preconceptions about TAAS test items.
  - Process - Discuss what students know about TAAS tests. Discuss with the students the possibility of students as test item designers. Discuss what information and skills they would have to possess to write TAAS test items. Ask them to complete the column "What I Know About TAAS tests" on the KWL chart. Ask the students to place the chart and the TAAS items in their portfolio.
  - Assessment - students participated in discussions, completed the first portion of the KWL chart. (KWL chart found in Appendix C)

Session II

- Outcome - Identify characteristics of reading passages, stems, distracters, incorrect answers, and correct answers
- Process - 1) Retrieve from the portfolios the items discussed the previous week. Divide students into cooperative learning groups. Have each group examine one part of a TAAS reading test (passage, questions, and answers). In group discussions have students generate specifics about each portion of the test item. Encourage discussion about the parts of the passage such as introduction, body, conclusion,

character development, the rise and fall of action and resolution of issues posed in the passage. In addition, encourage students in the quesiton and answer groups to notice characteristics of the stem including the wording, the types of questions, and the type of thinking required to approach the question. When considering the answers guide students to note the correct answer and its juxtaposition to the distracters. Have the students complete the section of "What Do I Know About Test Items" that pretains to their group work.

- Assessment - students identify components of a TAAS test item and document their shared information. (Chart for WKTI is found in Appendix C)

Session III

- Outcome - Complete the WKTI sheet through shared knowledge about TAAS test item component partswith other teams. Complete what need to know portion of KWL chart.
- Process - Reform groups so that new groups constitute members from each previous cooperative group representing the areas of passage, questions, and answers. Group members verbally share their previous group findings with new group members. All members write down the shared infomation resulting in a completed chart. Referring to that chart, students then consider what they would need to know to write a TAAS test. Each member making a contribution to the "think pool" places a mark on their paper indicating their contribution. All members make contributions to the process of deliberation. Students listen and respond by writing down each others' ideas on the KWL chart.
- Assessment - Students will document shared information on the

WKTI chart. Students will document group discussions of things needed to write TAAS tests on the KWL chart.

Session IV

- Outcome - Students will verbally identify the setting, action, and characters observed in a painting.
- Process - Students will observe a painting with considerable action, interesting setting, and multiple characters. Students will be asked to describe what they see. They will be encouraged to explore the setting, including all aspects of time, place, season, circumstance, and potential for change. In addition, they will be asked to describe the action in the painting including the action of the characters and the environment. Finally, students will be asked to consider the characters in the painting, their relationship to each other, the environment, the circumstance in the painting, and the potential for change. During these discussions, students should take notes on the WKAP form (found in Appendix C). Students should be issued chits or chips to register comments. When the chits or chips run out those persons must wait until all chits are spent before making additional contributions to the discussion.
- Assessment - Students will complete center portion of WKAP form.

Session V

- Outcome - Students will contemplate the painting from session IV and consider what could have happenened prior to the scene in the painting.
- Process - Students, while observing the painting, will review their notes from session IV and consider names for characters and parts of the scene (village names, castle names, forest, rocks, etc.). Students will

be asked to depart from the painting and consider the left side of the painting (a blank sheet of paper). They will be asked to consider the setting, character, and action and conjur what could have occured prior to the scene observable in the painting. Students will be issued chits as in session IV and the same procedure for discussion will ensue with the addition of creating an entry point for the setting, characters, and action. Students will be encouraged to document the discussion on their WKAP form.

- Assessment - Students will contribute to and document discussion using the WKAP form.

Session VI

- Outcome - Students will contemplate the painting from session IV and consider what could have happenened after the scene in the painting.
- Process - Students, while observing the painting, will review their notes from session IV and V and consider what line the story has taken to this point. Students will be asked to depart from the painting and consider the right side of the painting (a blank sheet of paper). They will be asked to consider the setting, character, and action and conjur what could possibly occur after the observable scene in the painting. Students will be issued chits as in session IV and V and the same procedure for discussion will ensue with the addition of creating an ending point for the setting, characters, and action. Students will be encouraged to document the discussion on their WKAP form.
- Assessment - Students will contribute to and document discussion using the WKAP form.

- **Between sessions** - A story will be written that incorporates all ideas from discussions. Students will edit the story and a final story will be issued to each student prior to session VII.

Session VII

- Outcome - Recognize and distinguish appropriate words for creating stems and answers for test items; create an item that addresses supproting ideas (details).
- Process - Provide each student with a sample of TAAS reading test item that is generated from the created story. Ask the students to discuss what the item signifies to them and their perceptions of the item with regard to their component parts (stem, distracters, correct answers) and where they can find the information in the created passage. Have the student highlight the part of the passage and compare the passage with the question and answers. Ask the students to discuss how they could have written the item differently. 1) Ask the students to retrieve, from their portfolio, the items discussed in the previous sessions. While looking at the items, discuss with the students the importance of words when writing items. Point out the words that imply specificity and those that are more global. 2) Ask the students to highlight another passage in the story and lead a discussion about what type of supporting idea question could be generated from the highlighted passage. Guide students through the process of generating a question, using well chosen words in the question. Once a group question has been created, guide the group through the wording of the correct answer. In groups have students consider distracters that could accompany the correct answer. Have them consider the wording so that the distraction is

realistic and plauible. Have groups report distracter to larger group. The larger group documents small group work and all students write the question, answer, and distracters on the Question Guide Form. As a large group have students generate two implausible answers. All answers are labeled: correct *, distracter ~, implausible #.

- Assessment - Students recognize and distinguish words and apply knowledge to creation of supporting ideas item while labeling their answers on the Question Guide Form (found in Appendix C).

Session VIII

- Outcome - Recognize and distinguish appropriate words for creating stems and answers for test items; create an item that addresses supproting ideas (sequence).
- Process - Review the process for creating a detail question in session VII. Discuss words that indicate sequence such as before, after, first, second, third, now and later, first and last. Have the student highlight the part of the passage and compare the passage with the provided question and answers. Ask the students to discuss how they could have written the item differently. 1) Ask the students to retrieve, from their portfolio, the items discussed in the previous sessions. While looking at the items, discuss with the students the importance of words when writing items. Point out the words that imply specificity and those that are more global. 2) Ask the students to highlight another passage in the story and lead a discussion about what type of supporting idea question using sequence could be generated from the highlighted passage. Guide students through the process of generating a question, using well chosen words in the question. Once a group question has been created,

divide the group into smaller groups and have them generate the wording of the correct answer. Check these answers and consider review for wordng if necessary. Proceed by having small groups generate distracters that could accompany the correct answer. Have them consider the wording so that the distraction is realistic and plauible. Have groups document small group work on the Question Guide Form All answers are labeled: correct *, distracter ~, implausible #.

- Assessment - Students recognize and distinguish words and apply knowledge to creation of supporting ideas sequene item while labeling their answers on the Question Guide Form.

Session IX

- Outcome - Recognize and distinguish appropriate words for creating stems and answers for test items; create an item that addresses relationships and outcomes (comparison and contrast).
- Process - Present students with a cartoon about comparison and contrast. Discuss words that indicate distinguishing differences such as change, differences, alike, same, before and after, big and small, etc. Have the student highlight the part of the passage and compare the passage with the provided question and answers. Ask the students to discuss how they could have written the item differently. 1) Ask the students to retrieve, from their portfolio, the items discussed in the previous sessions. While looking at the items, discuss with the students the importance of words when writing items. Point out the words that imply specificity and those that are more global. 2) Ask the students to highlight another passage in the story and lead a discussion about what

type of supporting idea question using sequence could be generated from the highlighted passage. Divide the group into smaller groups and have them generate the wording of the question, correct answer, distracters, and implausible. Check these answers and consider review for wordng if necessary. Have groups document small group work on the Question Guide Form All answers are labeled: correct *, distracter ~, implausible #.

- Assessment - Students recognize and distinguish words and apply knowledge to creation of relationship and outcome item while labeling their answers.

Session X

- Outcome - Recognize and distinguish appropriate words for creating stems and answers for test items; discuss the criteria for creating an item that addresses inferences.
- Process - Engage students in a discussion of the dress code of the school. Have them describe a typical student's attire and the acceptable code in the school. Present students with a scenerio of a student who enters the school dressed completely differently from all of the other students in the building. Engage the students in a discussion of the reasons why that student could be dressed as they are. Have the students draw on their knowledge of the school code to drive the discussion. From that discussion, lead students to know that they were able to infer certain things about the person dressed out of code based on facts and conjecture, but mostly facts. Have the students cite other examples of inference they might have noticed in their lives. Have students review the TAAS passage that uses inference as its base

question. Have them consider the words used and the fact that the answer must be generated other than from a specific word in the passage. Have the students highlight a passage from the story. In small groups have the students generate ideas about the aspects about the highlighted passage that can be used to create an inference item.

- Assessment - Students recognize and distinguish words and discuss the source of inference questioning and the potential for a question from a given passage.

Session XI

- Outcome - Recognize and distinguish appropriate words for creating stems and answers for test items; create an item that addresses inferences.
- Process - Have students review their small group discussions from the previous session. Have students reconsider the passage and the implications for an inference question. Have each group generate a question, answer, distracters, and implausible answers. Check each groups' progress and guide where necessary. Have students record their responses on the Question Guide Form.
- Assessment - Students will write a cogent inference question and answers and put these on the Question Guide Form.

Session XII

- Outcome - Reflect and respond in writing to discussion of what the student learned and what the student knows about TAAS test writing.
- Process - Have students in small groups. Issue each student five chips. Have students open their portfolios and review all of the materials within for evidence of things they have learned or now know about

writing TAAS tests. Encourage students to enter into a discussion of what they have learned or what they know about writing TAAS test items. Each student in turn places a chip on the table and offers an idea. The entire small group writes down their idea. The result is a pattern of placement, representing contribution to discussion from all members. When the small groups have exhausted their discussion, a representative from each group reports five ideas their group generated. No ideas can be repeated in the reporting from groups.

- Assessment - Verbally contribute to the discussion, write group ideas down, and report to other groups.

APPENDIX C

Name _____ Homeroom _____ Student ID _____

What do I **KNOW** about the different parts of a TAAS reading test?

Reading Passage

Questions

Answers

Name _____ Homeroom _____ id # _____

Example: Detail TAAS test item

According the to passage what was Lidia Tabitha's job . [STEM]

A. a fisherwoman at Fishing Falls +
B. a kitchen maid at the castle *
C. a chicken keeper in the village ~ +
D. a basket maker in the castle ~ +
E. a Dallas Cowboys Cheerleader#
F. a cook in the castle +
G. a chamber maid in the castle ~ +
H. the queen of the castle #

[RESPONSES]
* correct
+ possible (distracters)
~ + possible, unlikely
# not possible, ridiculous
****

Practice: Detail

[STEM]

[RESPONSES]

1.

2.

3.

4.

5.

6.

7.

8.

Name _____ Homeroom _____ id # _____

Example: Sequence TAAS test item

When was the first time Lidia Tabitha knew of the abondoned well? [STEM]

A. after she went to the village +
B. at the beginning of the story *
C. before she went to the river ~ +
D. during the rain storm ~ +
E. when she was swimming in the river#
F. when she was in the cave +
G. when she fell into the river ~ +
H. when she was jumping off the waterfall #

[RESPONSES]
* correct
+ possible (distracters)
~ + possible, unlikely
# not possible, ridiculous
****

Practice: Sequence

[STEM]

[RESPONSES]

1.

2.

3.

4.

5.

6.

7.

8.

Name _____ Homeroom _____ id # _____

Example: Compare and Contrast TAAS test item

How did the fishing day change in the story? [STEM]

A. from windy to rainy +
B. from sunny to stormy *
C. from day to night ~ +
D. from morning to afternoon ~ +
E. from rain to sunshine +
F. it didn't change ~ +

[RESPONSES]
* correct
+ possible (distracters)
~ + possible, unlikely
# not possible, ridiculous
****

Practice: Compare and Contrast

[STEM]

[RESPONSES]

1.

2.

3.

4.

5.

6.

Name _____ Homeroom _____id # _____

Inference/Prediction

Asks a question that suggests an idea that is not exactly clear in the story. The test taker has to know the whole story to consider what could happen or could have happened.
[STEM]

[RESPONSES]
* correct
+ possible (distracters)
~ + possible, unlikely
# not possible, ridiculous
****

Practice: Inference/Prediction

[STEM]

Look at paragraphs 12 and 14. Based on the information in the paragraphs, think about what clues lead to the identity of Lidia Tabitha's rescuer. Write a question that addresses that inference.

[RESPONSES]

1.

2.

3.

4.

5.

6.

Name _____ Homeroom _____ Student ID _____		
What do I **KNOW** about TAAS tests?	What do I **WANT** to know about writing TAAS tests?	What have I **LEARNED** about making TAAS tests?
What do I **KNOW** about writing TAAS tests?		

Name _____ Homeroom _____ Id # _____

What happened before the painting?

What is happening in the painting?

What would happen if the painting continued?

People/animals

Setting

Action

Please provide the following information:

Name_____ Date_____

Gender (*Please circle*): **Male Female** Score: T_____ W_____ E_____

Directions

A number of statements which people have used to describe themselves are given on the following page. Read each statement and then circle the appropriate number to the right of the statement to indicate how you *generally* feel:

1 = Almost Never, 2 = Sometimes, 3 = Often, 4 = Almost Always.

There are no wrong or right answers. Do not spend too much time on one statement but give the answer which seems to describe how you generally feel. Please answer every statement.

	ALMOST NEVER	SOMETIMES	OFTEN	ALMOST ALWAYS
1. I feel confident and relaxed while taking tests	1	2	3	4
2. While taking examinations I have an uneasy, upset feeling	1	2	3	4
3. Thinking about my grade in a course interferes with my work on tests	1	2	3	4
4. I freeze up on important exams	1	2	3	4
5. During exams I find myself thinking about whether I'll ever get through school	1	2	3	4
6. The harder I work at taking a test, the more confused I get	1	2	3	4
7. Thoughts of doing poorly interfere with my concentration on tests	1	2	3	4
8. I feel very jittery when taking an important test	1	2	3	4
9. Even when I'm well prepared for a test, I feel very nervous about it	1	2	3	4
10. I start feeling very uneasy just before getting a test paper back	1	2	3	4
11. During tests I feel very tense	1	2	3	4
12. I wish examinations did not bother me so much	1	2	3	4
13. During important tests I am so tense that my stomach gets upset	1	2	3	4
14. I seem to defeat myself while working on important tests	1	2	3	4
15. I feel very panicky when I take an important test	1	2	3	4
16. I worry a great deal before taking an important examination	1	2	3	4
17. During tests I find myself thinking about the consequences of failing	1	2	3	4
18. I feel my heart beating very fast during important tests	1	2	3	4
19. After an exam is over I try to stop worrying about it, but I can't	1	2	3	4
20. During examinations I get so nervous that I forget facts I really know	1	2	3	4

BIBLIOGRAPHY

Airasian, P.W. and Madaus, G.F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20 (2), 103 -118.

Anders, P. and Richardson, V. (1992). Teacher as game-show host, bookkeeper, or judge? challenges, contradictions, and consequences of accountability. Teachers College Record, 94 (2), 382-96.

Anderson, S.B. (1981). Testing and coaching. Paper presented at the annual meeting of the American Association of School Administrators. (ERIC Document Reproduction Service No. ED 206 726)

Archbald, D.A. and Porter, A. C. (1990). A retrospective and an analysis of roles of mandated testing in educational reform. Washington, D.C.: Office of Technical Assistance. (ERIC Document Reproduction Service No. ED 340 782)

Bangert-Drowns, R.L. Kilik, K.A., and Kulik, C.L.C. (1983). Coaching and its effects on test preparation. Review of Educational Research, 53 (4), 571-85.

Bracey, G. (1994) The fourth Bracey report on the condition of American Education. Phi Delta Kappan, October, 114-127.

Bracey, G. (1995) The fifth Bracey report on the condition of American Education. Phi Delta Kappan, October, 149-160.

Bracy, G.W. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. Phi Delta Kappa, 68 (9), 683-86.

Brandt, R. (1989). On misuse of testing: A conversation with George Madaus. Educational Leadership, 46 (7), 26-30.

Bridge, C. (1996). The implementation of Kentucky's primary program 1995: a progress report. Paper presented at the annual conference of the American Educational Research Association. New York City, New York, April 11, 1996.

Brooks, J. and Brooks, M. (1993). In search of understanding: The case for the constructivist clasroom. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Callenbach, C. (1973). The effects of instruction and practice in content-independent test-taking techniques upon standardized reading test scores of selected second grade students. Journal of Educational Measurement, 10 (1), 25-30.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. Educational Measurement: Issues and Practice, 7, (2), 5-9.

Canner, J. (1992). Regaining the public trust: a review of school testing programs, practices. Educational Measurement: Issues and Practices, 9(3), 15-22.

Cohen, S. A. and Hyman, J. S. (1991). Can fantasies become facts? Educational Measurement: Issues and Practice, 10 (1), 20-23.

Cushing, K. (1989). Testwise or test foolish: effects of riverside materials on test taking skill instruction. Paper presented at the annual meeting of the national council of measurement in education. (ERIC Document Reproduction Service No. ED 317 589)

Dangler, S. A. (1994). Intervention and remediation strategies for the Ohio ninth grade proficiency tests: implementation and percieved success in northwest Ohio secondary schools. (ERIC Document Reproduction Service No. ED 382 673)

Daniel, P. (1996). The impact of five state mandated reform initiatives on the instructional programs in Kentucky schools; school-based professional development supporting education reform. Paper presented at the annual conference of the American Educator's Research Association. New York City, New York, April 11, 1996.

Diamond, J.J. and Evans, W.J. (1972). An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 9 (2), 145-50.

Dolly, J. P. and Williams K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. Educational and Psychological Measurement, 46 (3), 619 - 25.

Driver, R., Asoko, H., Leach, J., Mortimer, E., and Scott, P. (1994). Constructing scientific knowledge in the classroom. Educational researcher, 23, 7, 7-12.

Ducote, K. (1982). Motivation and testwiseness interactions. Paper presented at the annual meeting of the Southwest Educational Research Association. (ERIC Document Reproduction Service No. ED 225 211)

Edleman, J. (1981). The impact of the mandated testing program on classroom practices: Teacher perspectives. Education, 102 (1), 56-59.

Flippo, R.F. and Borthwick, P. (1981). Should testwiseness curriculum be a part of undergraduate teacher education? Paper presented at the Annual Meeting of the American Reading Forum. (ERIC Document Reproduction Service No. ED 218 591)

Fogarty, R. and Bellanca, J. (1991) Patterns for thinking patterns for transfer. Palatine, Illinois: Skylight Publications.

Fogarty, R., Perkins, D. and Barell, J. (1992) How to teach for transfer. Palatine, Illinois: Skylight Publications.

Frankel, S. (1983). Study of test burden at the elementary and intermediate schools. Rockville, Maryland: Montgomery Count Public Schools. (ERIC Document Reproduction Service No. ED 251 486)

George, P. (1985). Coaching for tests: A critical look at the issues. Curriculum Review , 25 (1), 23-26.

Guifford, C.S., and Fluitt, J. L. (1980). How to make your students testwise. The American School Board Journal, 167, (10), 29-30.

Hakstain, A.R. (1971). The effects of type if examination anticipated on test preparation and performance. Journal of Educational Research, 64 (7), 319-24.

Haladyne, T. M., Nolan, S., and Haas, N.S. (1991). Raising standardized achievement test scores and the origins of test score pollution. Educational Researcher, 20 (5), 2-7.

Hall, J. and Kleine, P. (1990). Preparing students to take standardized tests: have we gone too far? (ERIC Document Reproduction Service No. ED 334 249)

Harlen, W. and Osborne, R. (1983). Toward a teaching model of primary science. (ERIC Document Reproduction Service No. ED 252 397)

Herman, J.L. (1989). Research and development priorities for educational testing and evaluation: Testimony of the cresst national faculty. Los Angeles, CA.: National Center for Research on Evaluation Standards and Student Testing, Office of Research and Improvement. (ERIC Document Reproduction Service No. ED 352 381)

Herman, J.L. and Golan, S. (1993). The effects of standardized testing on teaching and schools. Educational Measurement: Isues and Practice, 12 (4), 20-25.

Herman, K. L. and Golan, S. (1990). Effects of standardized testing on teachers and learning: another look. Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing. (ERIC Document Reproduction Service No. ED 341 739)

Houston, P. and Schneider, J. (1994). Drive-by critics and silver bullets. Phi Delta Kappan, May, K1-K12.

Huck, S, Cormier, W, and Bounds, W. (1974). Reading statistics and research. New York: Harper Collins Publishers.

Hunkle, D., Wiersma, W. and Jurs, S. (1994). Applied statistics for the behavioral sciences. Boston: Houghton Mifflin Company.

Hyerle, D. (1996). Visual tools for constructing knowledge. Alexandria, VA: Association for Supervision and Curriculum Development.

Jones, P. and Ligon, G. (1981). Preparing students for standardized testing: a literature review. (ERIC Document Reproduction Service No. ED 218 519)

Joyce, B., Weil, M., and Showers, B. (1992). Models of teaching. Boston: Allyn and Bacon.

Kagan, S. (1994). Cooperative learning. San Juan Capistrano, CA: Kagan Cooperative Learning.

Kannapel, P., Coe, P., Aagaard, L., and Moore, B. (1996). "I don't give a hoot if somebody is going to pay me $3600:" local school district reactions to Kentucky's high stakes accountability program. Paper presented at the annual conference of the American Educational Research Association. New York City, New York, April 11, 1996.

Kher - Durlaghji, N. and Lacina - Gifford, L.J. (1992). Quest for test success: preservice teachers' views of high stakes tests. (ERIC Document Reproduction Service No. ED 353 338)

Kilian, L. (1992). A school district perspective on appropriate test - preparation practices: A reaction to Popham's proposals. Educational Measurement: Issues and Practice 11, (4), 13-15,26.

Kourilsky, M. and Wittrock, M.C. (1992). Generative teaching : An enhancement strategy for the learning of economics in cooperative groups. American Educational Research Journal, 29 (4), 861-76.

Laney, J. (1990). Generative teaching and learning of cost-benefit analysis: an empirical investigation. Journal of Research and Development in Education, 23 (3), 136-44.

Lenze, J. (1993). Learner generated versus instructor induced visual imagery. (ERIC Document Reproduction Service No. ED 363 327)

Ligon, G. (1981). Preparing students for standardized testing: one district's perspective. Austin, TX. (ERIC Document Reproduction Service No. ED 218 319)

Madaus, G. (1985). Test scores as administrative mechanisms in educational policy. Phi Delta Kappan, 66 (9), 611-17.

Madaus, G. (1989). Standardized testing: Harmful to educational health. Phi Delta Kappan, 70 (9), 688-697.

Martin, S. (1996). Conversation regarding real-estate practices in Dallas, Texas.

Matthews, B. (1996). The implementation of performance assessment in Kentucky classrooms. Paper presented at the annual conference of the American Educational Research Association. New York City, New York, April 11, 1996.

Maxwell, S. and Delaney, H. (1990). Designing experiments and analyzing data: a model comparison perspective. Pacific Grove, CA: Brooks/Cole Publishing Company.

McAuliffe, S. (1993). A study of the differences between instructional practice and test preparation. Journal of Reading, 36 (7), 524-31.

McKeon, R, Ed. (1947). Introduction to Aristotle. New York, New York: McGraw-Hill, Inc.

McNeely, P. (1996). Conversations about TAAS testing in the Dallas Public Schools.

Mehrens, W.A. (1989). Preparing student to take standardized achievement tests. Washington, D.C.: American Institutes for Research (ERIC Document Reproduction Service No. ED 314 427)

Mehrens, W.A. and Kaminski, J. (1989). Methods for improving standardized test scores: fruitful, fruitless, or fraudulent? Educational Measurement: Issues and Practice, 8 (1), 14-22.

Mehrens, W.A. and Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. Journal of Educational Measurement, 23 (3), 185-96.

Millman, J. (1981). Protesting the detesting of PRO testing. National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 210 302)

Neill, M. and Medina, N. J. (1989). Standardized testing: Harmful to educational health. Phi Delta Kappan, 70 (9), 688-97.

Nolen, A. B., Haladyna, T.M., and Haas, N.S. (1992). Abuses of achievement test scores. Educational Measurement: Issues and Practice, 11 (2), 9-15.

Norman, C.A. (1980). Measurement and testing: An nea perspective. National Research Memo. (ERIC Document Reproduction Service No. ED 213 762)

Norusis, M. (1994). SPSS advanced statistics 6.1. Chicago: SPSS, Inc.

Norusis, M. (1994). SPSS professional statistics 6.1. Chicago: SPSS, Inc.

Palmer, I. (1984). The ethics of test preparation at intensive english language programs. (ERIC Document Reproduction Service No. ED 248 727)

Paris, S. G., Lawton, T.A., Turner, J. and Roth, J. (1991). A developmental perspective on standardized achievement testing. Educational Researcher, 20 (5), 12-19.

Peckman, I. (1987). Statewide direct writing assessment in California. Paper presented at the annual meeting of the conference on College Composition and Communication. (ERIC Document Reproduction Service No. ED 281 230)

Phillips, A. (1983). Test taking skills: Incorporating them into the curriculum. Medford, OR: Jackson County Education Service District. (ERIC Document Reproduction Service No. ED 235 200)

Phillips, D. C. (1995). The good, the bad and the ugly: the many faces of constructivism. Educational researcher, 24 (7), 5-12.

Popham, W.J. (1987). The merits of measurement driven instruction. Phi Delta Kappan, 68, 679-82.

Popham, W.J. (1991). Appropriateness of teacher preparation practices. Educational Measurement: Issues and Practice, 10 (4), 12-15.

Prell, J. (1986). Improving test scores: teaching testwiseness. Review of the literature research bulletin. Bloomington, IN: Phi Delta Kappa Center on Evaluation, Development and Research. (ERIC Document Reproduction Service No. ED 280 900)

Robinson, J. and Wronkovich, M. (1993). Attaining better proficiency test scores using a multi-grade approach. American Secondary Education, 22 (2), 30-31.

Romberg, T. A. (1980). Salient geatures of the BTES framwork od teacher behaviors. In C. Denham and A. Lieberman (Eds.), Time to learn (pp. 73-93). Washington, D.C.: Department of Health, Education, and Welfare, National Institute of Education.

Samson, G.E. (1985). Effects of training in test-taking skills on achievement test performance: a meta-analysis. Journal of educational research, 78 (5), 261-66.

Savage, D. (1984). Test scores: Are they less rosy under scrutiny?. The American School Journal, 171 (8), 21-24.

Scruggs, T.E., White, K. and Bennion, K. (1986). Teaching test-taking skills to elementary grade student: a meta-analysis. The Elementary School Journal, 87 (1), 69-82.

Seaton, T. (1992). The effectiveness of test preparation seminars on performance on standardized achievement tests. (ERIC Document Reproduction Service No. ED 356 233)

Selden, R. (1985). Measuring excellence: The dual role of testing in reforming education. Curriculum Review, 25 (1), 14-18.

Shaycott, M. (1979). Handbook of criterion referenced testing: Development, evaluation, and use. Palo Alto, CA.: American Institutes for Research in the Behavioral Sciences. (ERIC Document Reproduction Service No. ED 217 048)

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test?. Educational Measurement: Issues and Practice, 9 (3), 15-22.

Shepard, L.A. (1991). Effects of high stakes testing on instruction. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 337 468)

Shepard, L.A. (1991). Will national tests improve student learning? Paper presented at the American Educational Research Association Public Interest Invitational Conference. (ERIC Document Reproduction Service No. ED _____)

Skinner, R.E. (1994). LEP readers: Standardized testing versus informal testing. (ERIC Document Reproduction Service No. ED 368 210)

Smith, J.K. (1982). Converging on correct answers: A peculiarity of multiple choice items. Journal of Educational Measurement, 19 (3), 211-220.

Smith, M.L. and Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. Educational Measurement: Issues and Practice, 10(4), 7-11.

Spielberger, C. (1980). Test anxiety inventory. Palo Alto, CA: Mind Garden.

Stallings, J.A., (1975). Relationships between classroom instructional practices on child development. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. (ERIC Document Repreduction Service No. ED 110 200)

Stedman, L. (1993). The condition of education: Why reformers are on the right track. Phi Delta Kappan, October, 215-225.

Stevens, J. (1992). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, publishers.

Summers, J. (1983). Improving test taking skills. Terre Haute, IN: Indiana State University, Curriculum Research and Development Center. (ERIC Document Reproduction Service No. ED 230 573)

Tunks, J. (1996). [What price success: a study of taas test preparation practices in a junior high school]. Unpulished raw data.

Webster, W.J. (1995). The connection between personnel evaluations and school evaluation. Studies in Educational Evaluation, (21), 227-54.

Webster, W.J. (1995). The connection between personnel evaluations and school evaluation. Studies in Educational Evaluation, (21), 227-54.

Webster, W.J., Mendro, R.L., and Almaguer, T.O. (1994). Effectiveness indices: A "value added" approach to measuring school effect. Studies in Educational Evaluation (20), 113-145.

Webster, W.J., Mendro, R.L., Bembry, K., and Orsak, T.H. (1995). Alternative Methodologies for identifying effective schools. Paper presented in a Distinguished Paper Session at the American Educational Research Association Meeting, San Francisco, CA.

Wittrock, M. C. and Carter, J. k. (1975). Generative processing of hierarchically organized words. American Journal of Psychology, 88 (3), 489-501.

Wittrock, M.C. (1974). Learning as a generative process. Educational Psychologist, 174 (11), 87-95.

Wittrock, M.C. (1991). Generative teaching of comprehension. The Elementary School Journal, 92 (2), 169-84.

Wittrock, M.C. (1992). Generative learning processes of the brain. Educational Psychologist,27 (4), 531-541.

Wittrock, M.C. and Alesandrini, K. (1990). Generation of summaries and analogies and analytical and holistic abilities. American Educational Research Journal, 27 (3), 489-502.

_____ (1993). Texas assessment of academic skills: Student performance results. Austin, Texas: Texas Education Agency.

_____ The effects of testing project: The effects of testing on teaching and learning. Los Angeles, CA.: Center for Research on Evaluation, Standards, and Student testing. (ERIC Document Reproduction Service No. ED 327 572)

_____ (1982). Helping students to do their vest on standardized achievement tests. Los Angeles, CA.: L.A. Unified school district; California Research and Evaluation Branch. (ERIC Document Reproduction Service No. ED 230 611)

_____(1982). Teaching students to be testwise: A handbook for teachers who administer or construct tests grades K-12. Montgomery County Board of Education. (ERIC Document Reproduction Service No. ED 220 542)

_____ (1987). A comparison of the effectiveness of four test preparation programs. Chicago, Ill: Final report - Chicago Public Schools, Department of Evaluation Bureau of ECIA Program Evaluation. (ERIC Document Reproduction Service No. ED 318 739)

_____ (1993). Texas assessment of academic skills: English language arts reading objectives and specifications. Austin, TX: Texas Eduction Agency.