The Effect of Training in Test Item Writing on Test Performance of Junior High Students

JEANNE TUNKS
Teacher Education and Administration, College of Education, University of North Texas, Denton, TX 76203, USA

SUMMARY  High stakes testing, a phenomena born out of intense accountability across the United States, produces instructional settings that marginalize both curriculum and instruction. Teachers and other school personnel have minimized instruction to drill and practice in an effort to raise standardized and criterion referenced test scores. This study presents an alternative to current practice that engages students in learning and increases their awareness of the internal aspects of standardized tests. The Test Item Construction Model (TICM) guides students through the process of studying test item stems and subsequently creating items using a 12 week process of incrementing from understanding to creating test items. Students grew in their understanding of the test item stems and the generation of these. An ANOVA did not yield significant differences between random groups of trained and untrained test writers. However, students in the experimental group demonstrated gains in understanding of test items.

Mandated testing, a fairly new phenomenon in US 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 US over the control of world power. This matter became the impetus behind the US 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.

Prior to the drive for accountability, testing was done on a local level. In 1939, Flannagan 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, 44 of the 50 states in the US 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 have awarded merit pay to teachers who have raised standardized test scores in their districts. This most recent occurrence in testing has fostered the use of the term ‘high-stakes testing’.

High-stakes testing reached a national concern by 1999, prompting the American Psychological Association and the National and the National Council on Measurement in Education to write a position statement on high-stakes testing. The 1999 Standards for Educational and Psychological testing ‘guides policy makers, testing professionals, and test users in high-stakes testing programs’ (www.aera.net). In support of the statement, the American Educational Research Association (AERA), the US’s largest professional research organization, took a position on high-stakes testing in July 2000 (aera.net). The policy takes issue with decisions based on a single test, provision for adequate resources and opportunities to learn, validation for intended use of scores, disclosure of negative effects of testing, alignment between the curriculum and the test, validity of passing scores, opportunities for meaningful remediation, student differences considerations, sufficient reliability, ongoing evaluation of intended and unintended effects of high-stakes testing. In every category, the AERA makes recommendations for policy changes among high-stakes testing states and districts. In spite of these statements and policy recommendations, many states, led by Texas, continue to pressure teachers and students to raise scores higher each year. The current means of preparing students, primarily drill and practice continues to dominate the approach to rising to the level of expectation of high-stakes mandates.

Purpose

The purpose of the study was to examine the effect of training in test item writing on test performance of junior high students. The alternative method used teacher students to write test items, as opposed to confronting them through testwiseness. A comparison was made between the scores of students who prepared for tests with the alternative method and of those who prepared using the traditional method of testing preparation, i.e. traditional drill and practice, testwiseness preparation, and test-taking trials. The alternative practice employed in this study was grounded in active constructivism, as represented in the theory of generative learning (Wittrock, 1974).

Research Hypotheses

Several hypotheses were considered in this study. It was hypothesized that: (1) students receiving alternative test preparation practice (Test Item Contraction Method—TICM) would score higher on mandated state tests (Texas Assess-
ment of Academic Skills—TAAS) in the following areas: (a) TAAS total score means; (b) TAAS reading subtest score means; (c) TAAS reading mastery means; and (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 with students who received current practice in test preparation.

**Concerns about High-stakes Testing**

Mandated testing practices manifest several concerns, among these, 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 & 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 & Alesandrini, 1990; Brooks & Brooks, 1993).

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

The use of tests for purposes other than diagnosing learner needs, prescribing instructional activities, and measuring student progress in curriculum content (Norman, 1980) constitutes a third concern. Test scores were published in newspapers throughout the US (Shepard, 1991), establish a prowess of academic excellence in communities touting higher scores. 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). External pressures extend beyond the community into the classroom and directly to the student.

The learner who expected to comply with the mandates of testing in a high-stakes setting comprises the final most important concern. 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 & Medina, 1989; Paris et al., 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.

**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 exist. These practices constitute circumvention schemes to which school personnel resort to survive the pressure. Another type of unacceptable activity involves reporting scores. Cannell (1988), in a survey of all 50 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, 50% 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 & Kleine, 1990; Palmer, 1984; Ducote, 1982; Ligon, 1981; Seaton, 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; Frankel, 1983; Phillips, 1983; Ligon, 1981; Anderson, 1981;
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 eighth-grade students for state mandated tests. Specifically, the 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, TICM, based in constructivist principles and generative learning techniques, fosters greater understanding of testing expectations. The strategy focuses on comprehending test items through the construction of such items. Current methodology of confronting items and developing testwiseness strategies that promote short-term thinking for short-term gain contrasts with the TICM. 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 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.

Wittrock’s (1991) generative teaching model, 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 TICM guides students in the practice of constructing test items and eventually provides the basis of success in test-taking on like items and establishes 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.

**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 2158 students, of whom 6.6% were Anglo, 85.9% Hispanic, 6.6% African American, 0.10% Asian, and 0.80% American Indian. The percentage of students on free lunch was 73%. Thirty-two per cent 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 per cent of the students were ranked below the 50th percentile on grade equivalency measures. Fifty-three per cent were ranked above the 40th percentile and 36 above the 30th percentile on standardized test scores.

Students, particularly eighth graders, and their teachers were under considerable pressure to produce higher scores on the state mandated tests, TAAS. The pressure manifests in retention in eighth grade among students who fail to meet state mastery levels in all areas tested. In addition, students failing to meet standards face the requirement of attending summer school to drill further for excellence in TAAS testing. Finally, teachers who fail to produce increases in scores, particularly among minority students, face the possibility of remedial attention and eventual dismissal.

**Design**

The study used an experimental post-test 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 Thursday was written on slips of paper and one was left blank. Each of the five team members (teachers) secretly selected one piece of paper, thereby eliminating one intact class from the selection pool. The four remaining 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 of 25, 12 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 among control group students, although they did not consider the non-selected group as control, the researcher agreed to return in the spring of 1997 to give the remaining students the treatment, TICM training.

During the experimental portion of the study, the researcher met with each group of 12 students during the same time of the day on each of four week days across a 12-week period. Students in the experimental group were extracted from class and received 20 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 mid-year 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.

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 20-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 compared favorably with the Sarason’s Test Anxiety Scale. Test–retest reliability stability coefficient across two weeks and one month were 0.80 and higher for both high
school and college students. The coefficient dropped for the high school students after six months to 0.62.

**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 consisting 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. The instrument was tested for internal consistency using the Kuder–Richardson Formula which yielded coefficients ranging from 0.77 to 0.93 across all content areas tested and all grades levels tested.

**Method**

The unit of study (TICM), designed to accomplish the outcome of creating test items, was divided into 12 instructional sessions. Each session accommodated the various aspects of the generative learning model. Within the TICM structure, students examined actual TAAS items provided by the Texas Education Agency, considered their relationship and understanding of these, and consequently developed cognate items. 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, test item development. Each student maintained a portfolio throughout the project, which held assessments and to which they referred for guidance as they progressed toward the outcome, developing test items.

Test items constructed consisted of multiple choice items that related to TAAS instructional objectives in reading and language arts. State mandated items tested students’ competence and understanding of a passage read. Consequently, students generated a passage using a painting and subsequently created multiple choice items similar to those presented on TAAS tests and correlated the items to the passage generated. The types of items presented on state tests include: comparison and contrast, detail recognition, inference, sequence, follow written directions, predict future actions and outcomes, describe plot, setting, character, and mood. During the training sessions, students developed items in the first four competencies tested by the state.
Results

Two multivariate analyses of covariance (MANCOVA) with 11 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 TICM 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 $\alpha \leq 0.05$, meaning that any probabilities obtained greater than 0.05 were considered non-significant.

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), T197ROB (reading objectives mastered TAAS), and ANX (anxiety measures).

Hypothesis Testing

It was hypothesized that students receiving an alternative test preparation 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 owing 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.

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 TICM, as presented in this study, had no bearing on how students performed on state mandated tests that have the same type of items on which subjects were trained to write during the treatment.
Table I. Descriptive statistics for mid-year TAAS tests (component parts) and anxiety inventory

<table>
<thead>
<tr>
<th>Type of measure</th>
<th>Experimental group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>T197raw = total score</td>
<td>36.85</td>
<td>10.94</td>
</tr>
<tr>
<td>T197SC1 = word meaning</td>
<td>6.77</td>
<td>2.23</td>
</tr>
<tr>
<td>T197SC2 = supporting ideas</td>
<td>4.95</td>
<td>2.13</td>
</tr>
<tr>
<td>T197SC3 = summarization</td>
<td>5.85</td>
<td>2.42</td>
</tr>
<tr>
<td>T197SC4 = relationships</td>
<td>3.69</td>
<td>1.60</td>
</tr>
<tr>
<td>T197SC5 = inference</td>
<td>8.92</td>
<td>2.77</td>
</tr>
<tr>
<td>T197SC6 = points of view</td>
<td>6.67</td>
<td>2.65</td>
</tr>
<tr>
<td>ANXTOT = anxiety total</td>
<td>45.98</td>
<td>4.87</td>
</tr>
<tr>
<td>ANXEM = anxiety emotion</td>
<td>17.90</td>
<td>4.87</td>
</tr>
<tr>
<td>ANXW = anxiety worry</td>
<td>19.60</td>
<td>5.06</td>
</tr>
<tr>
<td>T197MST = mastery of all objectives</td>
<td>0.375</td>
<td>0.490</td>
</tr>
<tr>
<td>T197ROB1 = word meaning</td>
<td>0.400</td>
<td>0.496</td>
</tr>
<tr>
<td>T197ROB2 = supporting ideas</td>
<td>0.475</td>
<td>0.506</td>
</tr>
<tr>
<td>T197ROB3 = summary</td>
<td>0.675</td>
<td>0.474</td>
</tr>
<tr>
<td>T197ROB4 = relationships</td>
<td>0.525</td>
<td>0.506</td>
</tr>
<tr>
<td>T197ROB5 = inference</td>
<td>0.350</td>
<td>0.483</td>
</tr>
<tr>
<td>T197ROB6 = points of view</td>
<td>0.350</td>
<td>0.483</td>
</tr>
</tbody>
</table>

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

The student-developed 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

Table II. Effect of group membership univariate F-tests with (1.47 df). Tests conducted with no covariates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Hypoth.SS</th>
<th>Error SS</th>
<th>Hypoth.MS</th>
<th>Error MS</th>
<th>$F$</th>
<th>Prob.$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T197RAW</td>
<td>4.24</td>
<td>3323.136</td>
<td>4.24</td>
<td>70.705</td>
<td>0.060</td>
<td>0.808</td>
</tr>
<tr>
<td>ANXEM</td>
<td>23.328</td>
<td>1286.859</td>
<td>23.328</td>
<td>27.379</td>
<td>0.852</td>
<td>0.361</td>
</tr>
<tr>
<td>ANXTOT</td>
<td>23.857</td>
<td>6198.318</td>
<td>23.857</td>
<td>131.879</td>
<td>0.180</td>
<td>0.673</td>
</tr>
<tr>
<td>ANXW</td>
<td>33.477</td>
<td>1409.110</td>
<td>33.477</td>
<td>29.9811</td>
<td>0.116</td>
<td>0.296</td>
</tr>
<tr>
<td>CONF</td>
<td>0.488</td>
<td>38.264</td>
<td>0.488</td>
<td>0.814</td>
<td>0.599</td>
<td>0.443</td>
</tr>
<tr>
<td>T197SC1</td>
<td>0.355</td>
<td>187.237</td>
<td>0.355</td>
<td>3.98</td>
<td>0.089</td>
<td>0.767</td>
</tr>
<tr>
<td>T197SC2</td>
<td>0.244</td>
<td>150.768</td>
<td>0.244</td>
<td>3.20</td>
<td>0.076</td>
<td>0.784</td>
</tr>
<tr>
<td>T197SC3</td>
<td>0.046</td>
<td>149.500</td>
<td>0.046</td>
<td>3.18</td>
<td>0.014</td>
<td>0.904</td>
</tr>
<tr>
<td>T197SC4</td>
<td>0.018</td>
<td>89.513</td>
<td>0.018</td>
<td>0.90</td>
<td>0.009</td>
<td>0.921</td>
</tr>
<tr>
<td>T197SC5</td>
<td>0.136</td>
<td>191.768</td>
<td>0.136</td>
<td>4.08</td>
<td>0.033</td>
<td>0.856</td>
</tr>
<tr>
<td>T197SC6</td>
<td>1.58</td>
<td>303.318</td>
<td>1.58</td>
<td>6.45</td>
<td>0.245</td>
<td>0.623</td>
</tr>
</tbody>
</table>

*P < 0.05.
TABLE III. Student developed test items: means and standard deviations

<table>
<thead>
<tr>
<th>Test item developed</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compare and contrast</td>
<td>6.56</td>
<td>1.98</td>
</tr>
<tr>
<td>Detail</td>
<td>7.15</td>
<td>2.49</td>
</tr>
<tr>
<td>Inference</td>
<td>6.43</td>
<td>2.62</td>
</tr>
<tr>
<td>Sequence</td>
<td>7.62</td>
<td>1.99</td>
</tr>
<tr>
<td>Total items written</td>
<td>69.87</td>
<td>15.89</td>
</tr>
</tbody>
</table>

answer accurate to the story, and plausible distracters. 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 0.812, indicating a high level of consistency across items. Table 3 shows the means and standard deviations for the different items produced by the students in the experimental group. Students exhibited 70% mastery on test item writing.

A content analysis of students’ portfolios intimated that students actively participated in the learning process at a high level. The ‘think’ sheets and KWL charts 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 and charts. The KWL chart, completed at three critical times across the sessions reflected student perception about what they ‘knew’, ‘wanted to know’, and ‘learned’ about TAAS testing. An examination of the completed sheets indicated that students actively participated in the process of test item development. The overwhelming majority of the sheets flourished with details generated from conversations regarding test item development.

The 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 TAAS 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 while 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.

Gains in understanding and perception about TAAS testing emanate in a comparison between the ‘K’ and ‘L’ sections of the charts. Responses to the ‘K’
section of the KWL chart, completed during the first session 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 contrast, statements and words examined in the ‘L’ indicated a shift in thinking among students. Seventy-five per cent 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.

**Implications**

The TICM approached testing knowledge and understanding from a generative theoretical base and placed the learner in the position of viewing the test from an insider’s perspective, as opposed to a confrontive view, afforded by testwise-ness and/or measurement driven instructional methodology. Consequently, TICM answered Shepard’s charge and provided what she calls ‘higher order’ thinking in test preparation (1991). Essentially, TICM provided tools for life-long learning (Hyerle, 1996) as opposed to learning the tricks to pass the test.

Students using the TICM experimented with possibilities for answers and generated plausible and implausible story lines, tests items, and suppositions about test writing. The generative processing used in TICM purported to engender skills required in life-long learning that included decision making, self-assessment, realignment, and interconnectedness. Students in TICM training displayed those skills. Therefore, the implication that participation in TICM instruction promoted greater understanding of the phenomenon testing, became apparent in this study.

The study demonstrated the use of an experimental design in an intact school setting. Understanding differences in treatment bears serious consideration. The field of test preparation mandates elevated studies that press the issue of verifiable differences among tested populations. This study provided a model for design and analysis that fulfilled that mandate.

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.

Educational researchers, challenged by the reduction of education to ‘teach 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.

REFERENCES


