

TTU CAT Instrument 2000 - 2005 Summary Report

Background

Tennessee Tech University began a pilot program during the 2000-2001 academic year to evaluate critical thinking skills of graduating seniors. During the 2000-2001 academic year approximately 200 seniors were given the *Tasks in Critical Thinking* Test developed by the Educational Testing Service (ETS). The students given the test were selected using a stratified random sample of seniors from four colleges at the University (education, arts & sciences, business, and engineering).

Tennessee Tech University selected the ETS test because it was an essay test and could involve faculty in the scoring and discussion of student responses. Such faculty involvement was seen as an essential ingredient in any subsequent efforts to encourage faculty to modify their teaching to improve critical thinking. Many faculty involved in the first scoring workshop gained insight into student deficiencies in critical thinking and discussed the need to modify their teaching approaches to provide students with more opportunities to develop critical thinking skills.

Three factors played an important role in our decision to stop using the ETS test and explore other means of evaluating critical thinking skills. Our statistical analysis of the test results and feedback from faculty involved in the scoring of the test raised serious questions about the validity of the test. Specifically, a variety of ambiguous and perhaps faulty guidelines for scoring responses reflected a failure to adequately refine the test. Secondly, while the test measured some aspects of critical thinking, it was neither comprehensive nor thorough. That is, many important areas of critical thinking were not addressed by the test, and those that were may not have been thoroughly and accurately assessed. Specifically, we found many questions simply asked students to restate ideas that were provided in the reading material without requiring any significant evaluation or critical analysis. We also found little evidence to corroborate the validity of the test when we examined the correlation between the ETS test scores and other measures of student achievement such as the ACT Test or cumulative grade point average. Finally, ETS informed us that they were removing it from the testing market so it would not be available for further use later that year.

We examined several alternative objective tests that had been developed to evaluate critical thinking. None of these tests involved faculty in the scoring of exams, and most of these exams operationally defined critical thinking in a very narrow way. Specifically, the objective tests focus almost exclusively on verbal, categorical, analogical, and hypothetical-deductive reasoning. While many faculty members think these skills are important, they also associate the teaching of those formal reasoning skills with courses in logic, mathematics, or formal problem solving. Consequently, the use of such tests as an assessment tool does not encourage broad faculty involvement in the development of critical thinking skills.

In order to encourage faculty involvement in not only the assessment of critical thinking, but also in the improvement of critical thinking skills, TTU embarked on an ambitious plan of having small groups of faculty work together to identify and develop an assessment tool for measuring critical thinking. The underlying idea was to increase faculty involvement and interest in developing critical thinking by identifying critical thinking skills that they themselves thought were important for their own students. Developing their own tests would give them a vested interest in the outcomes.

This effort began with an attempt to analyze what faculty liked about the previously used ETS exam and what they did not. Although the ETS test had numerous problems, the faculty involved in the first workshop generally thought that this type of test measured something important about students' abilities to evaluate and analyze new information. The fact that the test involved information that the students had never seen before was considered important. The fact that the

test required students to analyze and evaluate information and form conclusions was also regarded as important. An additional feature that was deemed important by some faculty members is that some of the tests asked students to determine what additional information they might need to further evaluate the issue under consideration. These observations became the starting point for developing a new test of critical thinking that would have high face validity and would, we hoped, correlate with other measures of student achievement.

During the 2001-2002 year TTU developed and pilot tested its first critical thinking test. Three groups of faculty worked in teams and as members of a larger group to identify important critical thinking skills and develop questions/materials that would measure those skills. The test relied heavily on essay answers to help assess communication skills (as well as critical thinking skills) and leave opportunities for creative answers to questions that don't always have a single correct response. The essay format also involved faculty in the scoring of exams and helped promote more interest in improving critical thinking skills. In addition, the test was based on topics that the faculty thought students would find intrinsically interesting. The latter decision derived, in part, from observations of some students' unwillingness to participate seriously in the previously administered ETS exam because they found the topics irrelevant to their interests and academic focus. The tests also involved some elements of "dynamic assessment," a procedure whereby students are given opportunities to learn and then use that newly acquired knowledge in new situations. Tests which do not use dynamic assessment measure what a student has already learned and not their potential to master new ideas and content.

Key Areas/Skills Targeted for Assessment

1. Ability to interpret numerical relationships in graphs.
2. Ability to identify inappropriate conclusions and understand the limitations of correlational data.
3. Ability to identify evidence that might support or contradict a hypothesis.
4. Ability to identify new information that is needed to draw conclusions.
5. Ability to separate relevant from irrelevant information when solving a problem.
6. Ability to learn and understand information in an unfamiliar domain.
7. Ability to use elementary mathematics skills in the context of solving a larger real-world problem.
8. Ability to draw inferences between separate pieces of information and formulate conclusions.
9. Ability to recognize how new information might change the solution to a problem.
10. Ability to communicate effectively.

The locally developed test (CAT) was administered to a stratified random sample of seniors at TTU. A subset of that sample also took the California Critical Thinking Skills Test (CCTST) to help evaluate criterion validity. The results of that first pilot test were very encouraging. The TTU test had high criterion validity when compared to CCTST scores ($r = .645$) and ACT scores ($r = .659$) scores. In addition, the test appeared to have high face validity and provided a good range of test scores with no ceiling or floor effects and a distribution that was reasonably close to a "normal" distribution.

During the 2002 – 2003 academic year, TTU continued the refinement and testing of the CAT critical thinking test. During the fall semester of 2002, approximately 200 TTU freshman and senior level students were evaluated with the CAT Critical Thinking Test. The freshman and senior TTU students were both selected using a stratified random sample from the Colleges of Arts & Sciences, Business, Education, and Engineering. Composite ACT score was used as a covariate to adjust for any potential differences between freshman and senior's entering ACT score. The results revealed a significant increase in critical thinking test scores from the freshman to the senior class ($p < .001$). The CAT test was also administered within several classes using a pretest/posttest design. The test results revealed significant gains in one course that focused on critical thinking/problem solving but not another comparable course that was offered at the same time in the social sciences (both courses were junior level social science

courses). The pattern of results discussed above provides evidence that the CAT test is sensitive to gains in critical thinking skills that may accrue from four years of college education and to gains in critical thinking skills that are associated with a single course in critical thinking/problem solving.

During the 2003 – 2004 academic year, TTU continued the refinement and testing of the CAT critical thinking test. Specifically, we examined how performance on the CAT instrument would compare to performance on the Academic Profile Test (ETS) using the short form. A stratified random sample of seniors took both the CAT instrument and the Academic Profile Test. We examined the correlation between scores on the Academic Profile Test, CAT instrument, and entering ACT score. As can be seen in Table 1, the CAT scores are significantly correlated with both the Academic Profile Test scores and the entering ACT scores at approximately the same magnitude. The Academic Profile Test has a slightly higher correlation with the students' entering ACT score. The latter difference probably reflects the fact that the ACT and the Academic Profile Test have considerable overlap in the skills being evaluated. The magnitude of the correlation between the CAT Score and the Academic Profile Test Score provides additional support for the criterion validity of the CAT instrument while also demonstrating that the CAT instrument measures something different from either the Academic Profile Test or the ACT.

Table 1
Correlation Matrix

	TTU CAT Instrument	Entering ACT Score
Academic Profile Test	.558	.693
TTU CAT Instrument		.599

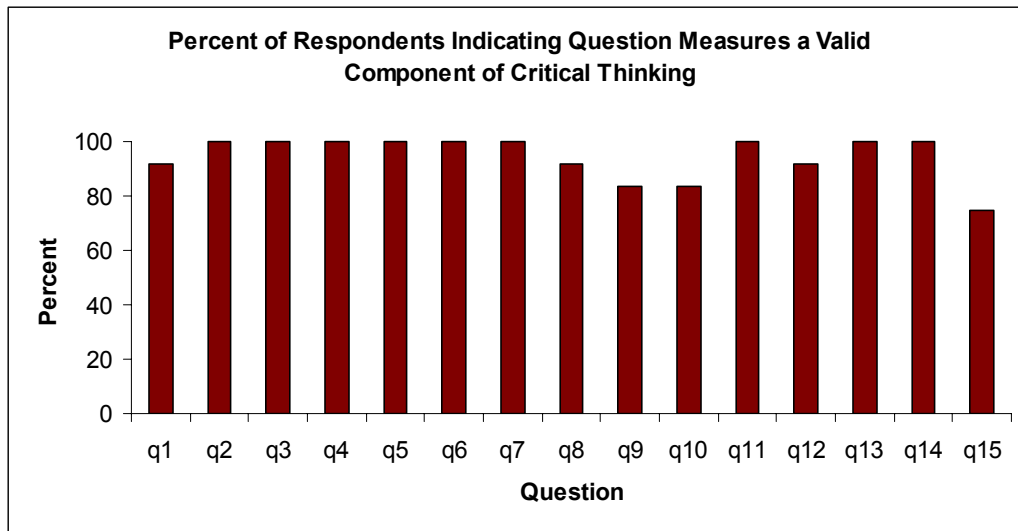
(all correlations significant, $p < .01$)

In our continuing efforts to improve the CAT instrument, we also examine scoring reliability since this has been problematic for many subjectively scored essay tests. Each question is scored by a minimum of two independent evaluators. If these two evaluators do not agree the question is scored by a third evaluator. To evaluate the reliability of scoring, the correlation between first and second evaluator scores for each question is calculated. In the most recent analysis, the average correlation for all questions was .87 which reflects positively on our continuing efforts to improve the test and the scoring criteria.

TTU also submitted a proposal to present a concurrent session at the 2003 SACS/COC annual convention in Nashville that would review TTU's efforts to develop a critical thinking test. This proposal was accepted, and the presentation in December of 2003 was both well attended (standing room only) and enthusiastically received. There appears to be considerable interest in finding better ways to assess critical thinking and in increasing faculty interest and involvement in the process. We received numerous requests for additional information as a result of the SACS presentation.

In the spring of 2004, TTU collaborated with the University of Memphis to administer and score the CAT instrument on their campus. The University of Memphis administered the CAT instrument to a random sample of approximately 130 seniors. Dr. Barry Stein from TTU provided assistance to an interdisciplinary team of faculty at the University of Memphis who scored the test. At the conclusion of the test scoring, faculty were encouraged to discuss their observations and to complete a survey to determine the extent to which each question measured a valid component of critical thinking. The results of the survey are summarized in Figure 1. These ratings reveal that the University of Memphis faculty who participated in the workshop generally considered the questions to measure valid components of critical thinking. These ratings provide additional support for the face validity of the CAT instrument.

Figure 1
Percent of Respondents Judging Questions as Valid
University of Memphis



The scoring workshop at the University of Memphis also provided a good opportunity to evaluate the reliability of scoring for the CAT instrument by people who had no prior experience with the test. Each question is scored by a minimum of two independent evaluators. If these two evaluators do not agree, the question is scored by a third evaluator. To evaluate the reliability of scoring, the correlation between first and second evaluator scores for each question is calculated. The average correlation for all questions was .85 at Memphis and compares favorably with correlations ranging from .83 to .87 observed at TTU.

2004-2005 Year

Overview

During the current academic year, TTU has continued to refine and test the CAT critical thinking instrument. The University received a three-year grant from the National Science Foundation to further refine the CAT instrument with input from six other universities across the country. In addition, TTU began to explore relationships between the widely used National Survey of Student Engagement (NSSE) and performance on the CAT instrument. The University's efforts to develop an effective tool for assessing critical thinking have also set the stage for the University's Quality Enhancement Plan that will involve the campus in efforts to improve critical thinking and real-world problem solving through the use of active learning strategies.

NSF Grant Activities & Findings Related to the CAT Instrument

TTU received a three-year \$499,994 NSF grant to work with six other institutions across the country to refine the CAT instrument this year (www.tntech.edu/cat).

- The University of Texas
- The University of Washington
- The University of Colorado
- The University of Hawaii
- Howard University

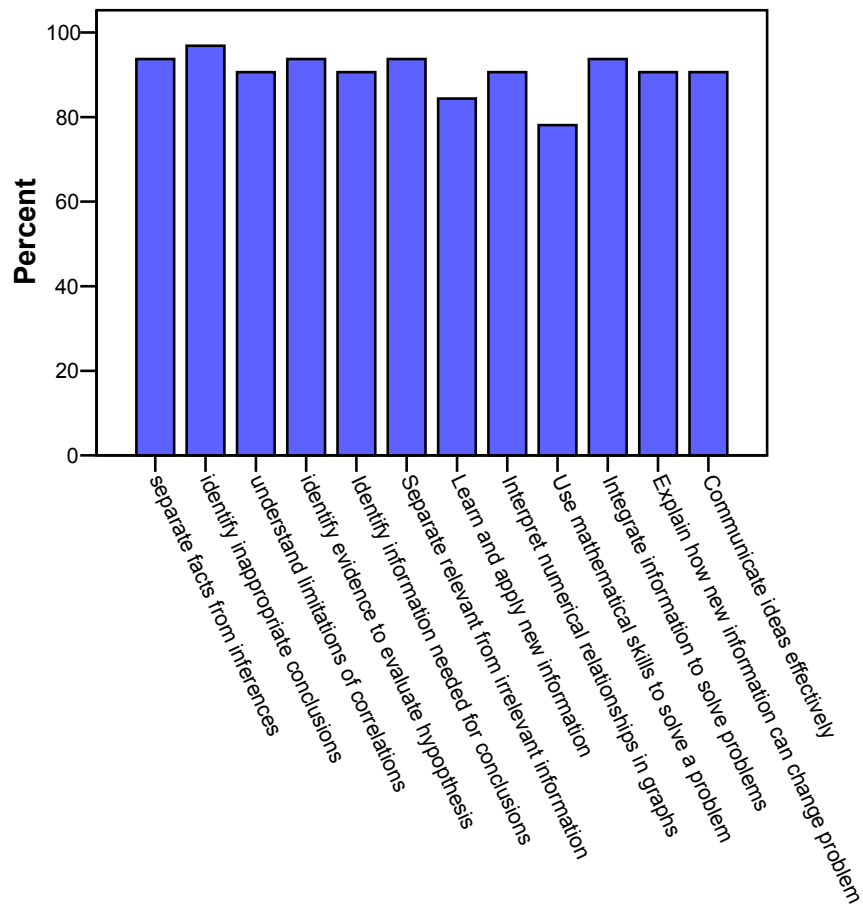
- The University of Southern Maine

During the first year of the grant, TTU worked with the University of Hawaii, the University of Southern Maine, and The University of Texas to administer and score the tests using local faculty graders. These faculty members provided detailed feedback about the test and the scoring process. This information is currently being used to further refine the test and scoring guide. To date, the feedback received from other institutions has been very positive and helpful. The data collected from these institutions that is available for this report are summarized below.

Evaluation of Skill Areas Targeted by the CAT Instrument

Faculty participants in the scoring workshops were asked to indicate which of the skill areas targeted by the CAT instrument they considered to be important components of critical thinking. Figure 2 illustrates the findings of this survey. The findings indicate that the areas of skill targeted by the CAT instrument were generally perceived as important components of critical thinking by most faculty who participated in the three scoring workshops this year. The only area where less than 80% of the faculty felt the area was an important component of critical thinking involved using mathematical skills to solve a complex real-world problem.

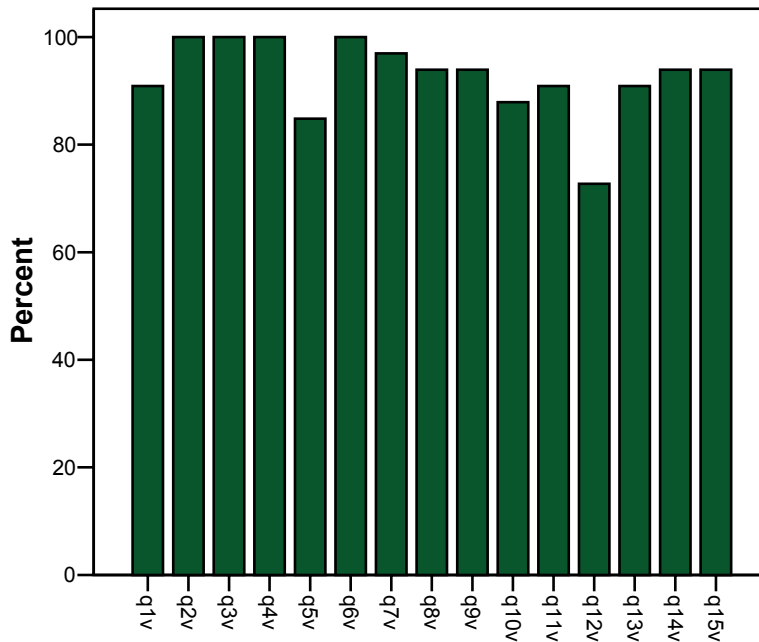
Figure 2
Percent of Faculty that Identify Areas Targeted by CAT as Important Components of Critical Thinking



Evaluation of Question Face Validity

The faculty who participated in the scoring workshops were also asked to evaluate the face validity of each question contained in the CAT instrument. Most faculty felt that the questions included on the CAT instrument were valid measures of critical thinking (see figure 3). The question with the lowest overall support (question 12) involved using a mathematical calculation that was needed on subsequent questions to help solve a complex real-world problem. We received some suggestions for improving question #5 that we will explore to improve its perceived validity.

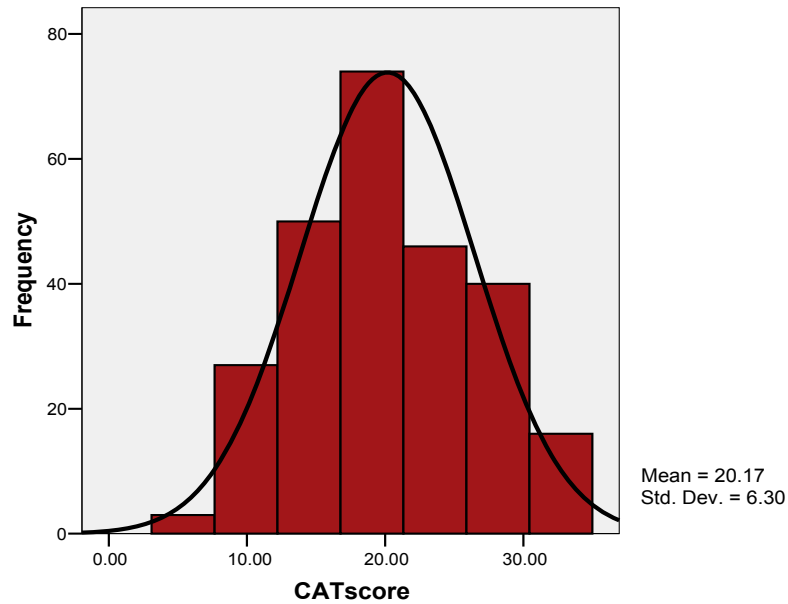
Figure 3
Percent of Faculty Indicating Question Measures a Valid Component of Critical Thinking



Distribution of Scores

Figure 4 shows the distribution of student scores (raw) on the CAT instrument against the normal curve. These scores are similar to those obtained in prior testing at TTU and the University of Memphis. Scores ranged from a low of 7 to a high of 33. There was no evidence of a floor effect or a ceiling effect (lowest possible score = 0, highest possible score = 40). We expect to adjust the weights assigned to each question based on input from the faculty scorers and our external consultant. Once we have finalized question weights, we will explore procedures to standardize the test scores.

Figure 4
Distribution of Student Scores



Correlation with other Measures of Student Performance

Performance on the CAT instrument was correlated with other measures available for the students tested at the participating institutions including entering SAT scores and cumulative grade-point averages. These correlations appear in table 2. The correlations provide support for the criterion validity of the CAT instrument. Entering SAT scores explained 25% of the variability in the CAT instrument. The magnitude of the correlation with the entering SAT score is similar to findings that have been previously observed with the entering ACT score, concurrent performance on the ETS Academic Profile Test, and the California Critical Thinking Skills Test (CCTST).

Table 2
Correlations

	SAT (verbal & math)	Cumulative Grade-point Average
CAT Score	.50 *	.34 *
SAT	-	.52 *

* correlations significant, p < .01

Scoring Reliability

Scoring reliability was evaluated by examining scores assigned by faculty grader one and faculty grader two on each question. The average reliability of scoring across questions is presented in table 3.

Table 3
Scoring Reliability

Location	Scoring Reliability
University of Hawaii	.80
University of Southern Maine	.78
University of Texas	.85
Overall	.81

Preliminary Analysis of Cultural Fairness

Although more extensive analyses of any possible ethnic/racial/gender bias in the CAT instrument are planned, a preliminary analysis of available data provided encouraging results. A multiple regression analysis revealed that once the effects of the entering SAT score were taken into account, none of the predictors related to gender, race, or ethnic background were significant predictors of overall CAT performance.

CAT Performance and NSSE Scores at TTU

A stratified random sample of 120 seniors at TTU received both the CAT instrument and the NSSE survey to evaluate the potential relationship between different types of student engagement activities and performance on the CAT instrument. A scoring workshop was also conducted at TTU to evaluate student performance on the test. Although the data are still being analyzed, preliminary findings indicate that various components of the NSSE are significantly correlated with student performance on the CAT instrument. The table below illustrates some of the correlations between specific NSSE questions and CAT scores. In a regression analysis, the combination of NSSE questions listed below yielded a regression coefficient = .426, $p < .01$. These results provide additional support for the validity of the CAT instrument and indicate some potential areas where strategic initiatives might be focused to improve critical thinking performance.

Table 4
NSSE Correlations

NSSE Question	Correlation with CAT Score
(1i) Put together ideas or concepts from different courses when completing assignments or during class discussions.	.165*
(2a) Memorizing facts, ideas, or methods from your courses and readings so you can repeat them in pretty much the same form.	-.245**
(3b) Number of books read on your own (not assigned) for personal enjoyment or academic enrichment.	.209*
(7h) Plan to participate or already participated in culminating senior experience (thesis, capstone course, project, comprehensive exam, etc.)	.224**
(11e) Institution contributed to thinking critically and analytically.	.157*

* Significant at .05 level (one tailed)

** Significant at .01 level (one tailed)

2000-2005 Summary and Conclusions

Five years ago, TTU set out to evaluate Tasks in Critical Thinking (ETS) as an instrument to assess students' critical thinking skills. The ETS test was selected because it would involve faculty in the scoring of open-ended responses to better help our faculty understand our students' weaknesses. In our first year report, we pointed out some of the weaknesses of this test and were subsequently encouraged to explore other testing alternatives. Given our interest in a faculty scored test, we were left with few options. We decided to embark on a rather ambitious project to try to develop our own test of critical thinking, a test that would capture some of the positive aspects of the ETS test and yet avoid the numerous problems we observed. In less than five years, TTU has made remarkable progress in developing a short-answer essay test to evaluate students' critical thinking skills. We have now administered this instrument to over 1000 students at five universities across the country. The instrument has demonstrated excellent face validity, criterion validity, scoring reliability, and it even appeals to students taking the exam. Our work on this instrument has been recognized by SACS and the National Science Foundation. With funding from the National Science Foundation we are now in the process of refining the test using a distinguished pool of universities across the country. This initiative has exceeded our highest expectations for success and is an excellent example of the positive results that can occur from performance funding.

We will continue to work on this project even though there is no longer a performance funding incentive to do so. In fact, we have found in our search for a QEP topic, that the skills we were attempting to measure with the CAT instrument are the very same skills our faculty, students, and employers think are most important. Consequently, our QEP topic for SACS will focus on improving students' critical thinking/real world problem solving skills through active learning strategies. The CAT instrument will provide one useful assessment of our progress on this QEP.

Adaptability and Feasibility for Statewide Testing

We believe that the CAT instrument could be useful to other institutions in Tennessee. The usefulness of the test relates to two important characteristics.

- It assesses a collection of critical thinking skills that diverse groups of faculty consider important components of critical thinking (and that no other test assesses as completely)
- It serves as a faculty development tool to encourage improvements in pedagogy by involving faculty in the scoring of student responses and making them aware of their students' shortcomings in areas they consider essential for student success.

The recent NSF grant is allowing TTU to further refine the test using a national audience. The enhancements to the instrument that occur as a result of this funding will make it even more useful to other institutions.

Because the test is faculty scored, it will not be financially feasible for most institutions to administer the instrument to all graduating seniors. Although the cost of the test itself is relatively low, the cost associated with paying faculty to score the test makes it prohibitive for testing very large groups of students. A representative sample of 100 to 200 students can be scored by 10 – 14 faculty in a one-day scoring workshop. This size sample can provide an adequate cross section of the institution to break down performance by college and assess efforts to improve critical thinking. The test is best suited for situations in which the institution has a specific goal to improve critical thinking because the involvement of faculty in the scoring will support quality improvement efforts.