

# Assessing Critical Thinking Skills

Barry S. Stein, Ada F. Haynes, and Jenny Unterstein  
Tennessee Technological University

## Abstract

Tennessee Technological University has been exploring methods of assessing critical thinking skills as part of a performance funding initiative since 2000. Our experiences over the last three and half years provide useful information about both a process for developing an assessment tool and a product for assessing critical thinking skills. Our approach has been to empower our faculty to both identify and evaluate a core set of skills they believe to be an important part of critical thinking in our graduates. Our initial test has demonstrated good face validity and high criterion validity when correlated with scores on the American College Test (ACT) and the California Critical Thinking Skills Test (CCTST).

## Introduction

Many colleges and universities are looking for better ways to evaluate student learning either in response to state mandates for accountability or to satisfy quality enhancement objectives. While most institutions will use existing assessment instruments, the approach described here sought to create an assessment that was more directly related to faculty objectives and hence could engage the faculty in quality enhancement. Our experiences and evolving assessment product may provide useful information for other institutions that are searching for a similar assessment methodology.

When our university decided to participate in a performance funding initiative to evaluate critical thinking skills we explored a variety of tests that were being marketed for that purpose. Recognizing that any initiative to evaluate critical thinking might ultimately require efforts to improve critical thinking performance, we searched for a test that would have high face validity in the eyes of our faculty. Many attempts to assess educational outcomes rely on tests that assess knowledge or skills that are not deemed important by the instructional personnel. Adopting an assessment benchmark that is not consistent with teaching goals of a universities' faculty may be counter productive – especially if evidence of continuous quality improvement is desired.

We discovered the options were fairly limited and we chose a test that involved short answer essay questions that could be graded by our own faculty. Although selecting an essay test that must be scored by one's own faculty is more costly than objective tests, it has the advantage of allowing faculty to directly experience the weaknesses of their students. The latter outcome is particularly important if an institution is interested in continuous improvement or quality enhancement.

---

Paper presented at SACS/COC Annual Meeting / Nashville, Tennessee / December 6 - 9, 2003. Dr. Barry S. Stein is a Professor of Psychology at Tennessee Technological University and Director of Planning for the university (Bstein@tntech.edu). Dr. Stein also serves as the coordinator for the university's critical thinking initiative and the university's teaching evaluation process. He is coauthor of *The Ideal Problem Solver: A guide for improving thinking, learning and creativity*. Dr. Ada F. Haynes is Professor of Sociology and Political Science at Tennessee Technological University. Dr. Haynes is the author of *Poverty in Central Appalachia*. Jenny Unterstein is a doctoral student at Tennessee Technological University.

Our experiences with that test (Tasks in Critical Thinking) were generally positive although our subsequent analyses of results showed poor criterion validity when related to students ACT scores. Our faculty also had some concerns about the face validity of the instrument. These concerns coupled with the subsequent removal of the test from the market necessitated an alternative approach. While we considered using an objective test such as the *California Test of Critical Thinking Skills*, we found that most of our faculty did not consider the thinking skills assessed in that test to be an important part of their own courses except in the case of certain philosophy, problem solving, or mathematical logic classes (they did think that such tests measured some important thinking skills).

We decided to pursue a somewhat radical and ambitious alternative - that alternative being the development of our own test to assess critical thinking skills. This test was envisioned to assess what our own faculty considered to be important and relevant critical thinking skills. It was acknowledged from the outset that this test would probably not evaluate all aspects of critical thinking, but instead would focus on a core set of critical thinking skills that would be interdisciplinary and reflective of skills needed to be successful in many fields.

### Early Development Efforts

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. It was agreed that the tests would involve mostly essay answers to help assess communication skills and leave opportunities for creative answers to questions that don't always have a single correct response. The essay format would also involve faculty in the scoring of exams and hence promote more interest in improving critical thinking skills. In addition, it was decided that the test would be 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 seriously participate in the previously administered ETS exam because they found the topics irrelevant to their interests and academic focus. It was agreed that the tests should also involve 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. Finally, it was decided that each test should follow a similar format even though they focused on different content areas. This would allow comparisons across different versions of the test – something that was not possible with the previously administered ETS test.

The skills identified by the interdisciplinary committee working on this project included a broad range of skills that underlie effective problem solving, life-long learning, and critically evaluating information.

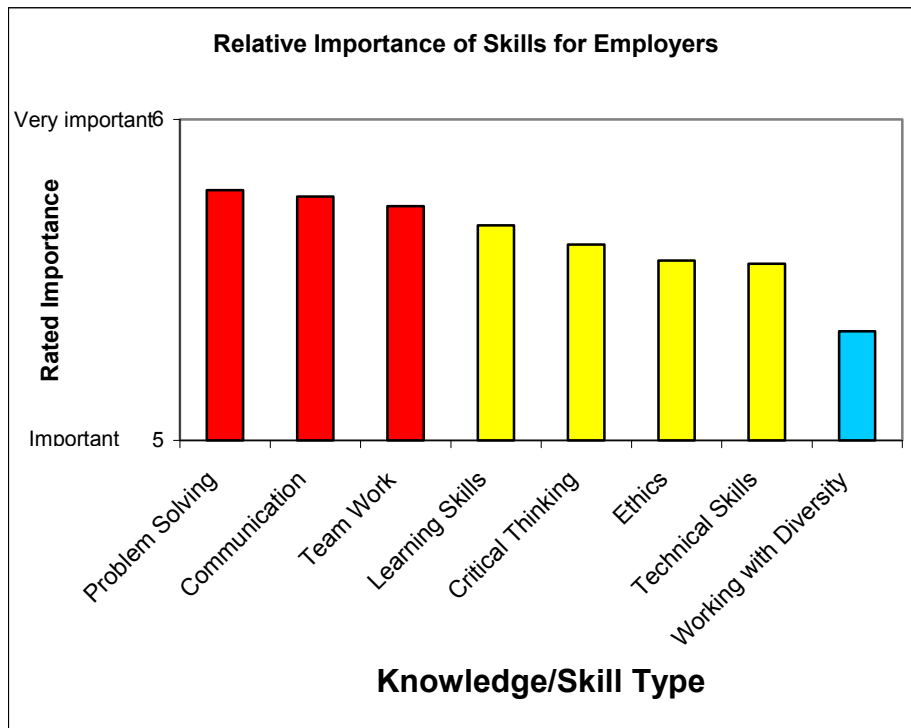
#### Important skills identified by our faculty

1. Interpret numerical relationships in graphs.
2. Identify inappropriate conclusions and understand the limitations of correlational data.
3. Identify evidence that might support or contradict a hypothesis.
4. Identify new information that is needed to draw conclusions.
5. Separate relevant from irrelevant information when solving a problem.
6. Learn and understand information in an unfamiliar domain.
7. Use mathematics skills in the context of solving a larger real world problem.
8. Analyze and integrate information from separate sources to solve a complex problem.

9. Recognize how new information might change the solution to a problem.
10. Communicate critical analyses and problem solutions effectively.

It is interesting that these skills clearly relate to 4 out of 5 of the most important skills identified by employers of our graduates in 2003 (see figure 1).

Figure 1 (2003 TTU Employer Survey)



Faculty explored a broad range of content areas that they thought would be intrinsically interesting to students and settled on three initial content domains. These areas included; water pollution, crime, and Alzheimer’s disease. A general format for questions that could be applied across a broad range of content areas was developed and refined. Each version of the test was pilot tested by faculty in their classes and the group as a whole looked at the range of student responses with the aim of clarifying confusing questions and making modifications that would improve the validity of the test. Scoring criteria were also developed to help graders assign point values to each answer.

To help evaluate the newly developed test a stratified random sample of graduating seniors in four colleges (Arts & Sciences, Education, Engineering, Business) were selected for testing. One hundred and nineteen students took the California Critical Thinking Skills Test (CCTST) and 104 students took the CAT Critical Thinking Test (developed by TTU). Sixty-four students in the sample took both tests. Additionally, approximately 30 students in 4 different classes also participated in preliminary pilot testing to assist in the development of the tests.

Ten faculty members from ten different disciplines (Biology, English, Economics, Engineering, Chemistry, History, Mathematics, Political Science, Psychology/Education, and Sociology) participated in a day-long workshop to score the exams and evaluate the test. The faculty members were given an outline for scoring each answer, but were asked to make modifications if necessary to refine the scoring system to improve its reliability and ease of

application. Faculty discussed the range of answers given by students and found the test to be sensitive to differences in the quality of student answers. The faculty generally thought that the test did a good job of measuring critical thinking and suggested some modifications in certain questions to improve the test and increase the reliability of scoring. This group of faculty also reviewed the preliminary versions of two other tests and made recommendations for improving those tests that were still in development. In general, the faculty were pleased with the test and those not already participating in the test development process expressed an interest in participating in the development of new versions of a critical thinking test at some point in time.

Although the test development process was designed to ensure good content validity from the perspective of our faculty, we sought to evaluate criterion validity by comparing performance on this test to other measures of academic achievement and critical thinking. The process of establishing criterion validity by comparison to other related measures of performance is common practice in test development. The goal would be to show that the new test is reasonably correlated with other related measures of performance, but is not so highly correlated that the new test would appear to be measuring the same skills. Students' scores on this test were compared with their scores on both the ACT (at time of admission) and the California Critical Thinking Skills Test (CCTST). Correlations among the tests were good but not excessively high (TTU Test and ACT,  $r = .659$  and TTU Test and CCTST,  $r = .645$ ).

### Test Scoring Process

Test scoring is a critical aspect of this instrument. Having faculty score the tests provides them with a first hand understanding of students' strengths and weaknesses. This understanding can provide sufficient motivation to explore changes in pedagogy that would help support institutional efforts to improve students' performance. Each scoring session typically involves about ten to twelve faculty for a whole day. Faculty are paid a reasonable stipend for their participation in the scoring workshop and seem to enjoy the process. Faculty who volunteer seem to be genuinely interested in critical thinking and like the financial reward.

It is important to ensure that the scoring is reliable and not arbitrary for the results to be meaningful and valid. We have developed a very specific scoring criteria and continue to refine the guidelines. The criteria provide specific guidelines for awarding points on each question. We try to enlarge the pool of participating faculty by encouraging the participation of new faculty in each scoring workshop, but we also try involve a core of experienced faculty to maintain reliability across scoring sessions.

A minimum of two faculty grade each question. If the scores given by the first two graders differ, the answer is read by a third faculty member. The final score is either based on the two raters that agree with each other, or on the average of all three if there is no agreement. During the scoring workshops, individual tests are rotated across faculty graders after every two or three questions to increase the generalizability of the scoring. We continually analyze inter-rater agreement patterns to identify questions that might be problematic to score, and explore ways of rewording the question or scoring criteria to improve inter-rater agreement.

### More Recent Development Efforts

Although our early development efforts sought to develop three different test versions, we have recently sought to concentrate on the improvement and refinement of just one test. Our efforts to improve this test have been focused in three areas.

- Adding questions that improve the face validity of the test for our faculty
- Improving the clarity of questions to reduce ambiguity
- Improving the scoring criteria to increase agreement between raters

We evaluate new items by adding them to the test and surveying our faculty graders after they have had the opportunity to view the question and the range of responses to the new question. While many questions may appear to increase the validity of the test before they are scored, we have found that faculty judgments change after they have had the opportunity to consider the range of student responses and how clearly they can be differentiated by the graders. We also continue to modify the scoring criteria to improve the agreement between raters.

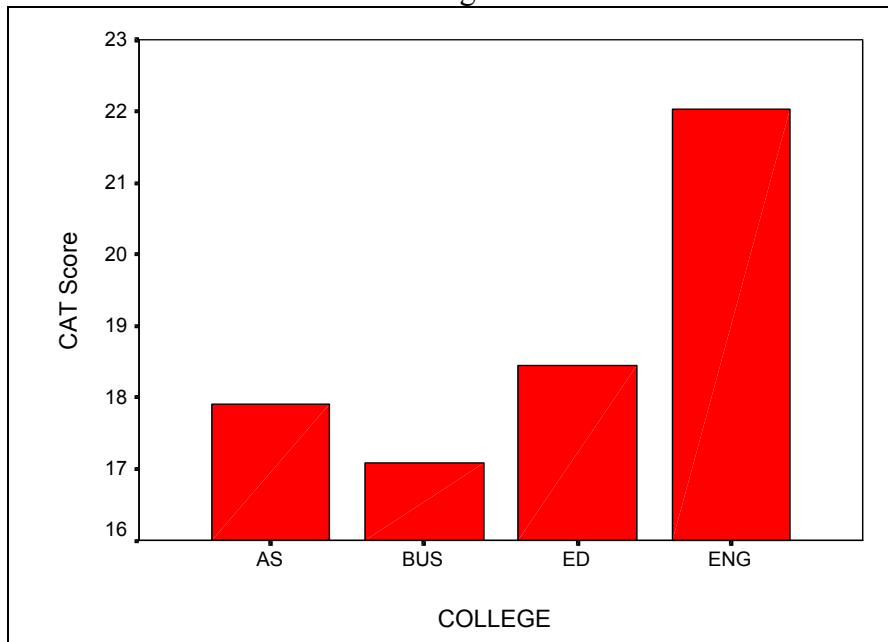
### Test Use

We have administered the test to a variety of groups in our efforts to develop and refine it. The various ways in which we have sampled students from the university population and the results are described below.

#### *Seniors by College*

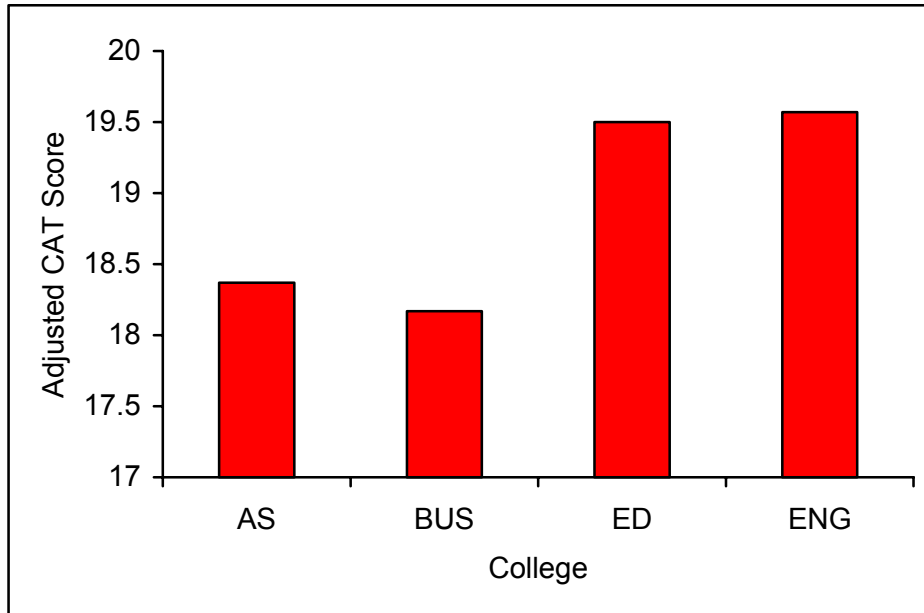
In our initial testing we examined stratified random samples of seniors at TTU from four colleges. Figure 2 shows average performance on the CAT Test by college. As expected, engineering students performed the best on this test. It is important to note that entering engineering students have ACT scores that are the highest of the four colleges tested and ACT scores explain about 40% of the variability on the CAT Test.

Figure 2



An analysis of covariance was performed to remove the effects of entering ACT scores and the adjusted means for the colleges are presented in figure 3. Although the differences between colleges were not significant after adjusting for ACT scores ( $p > .05$ ), the ordering remained similar to graph above.

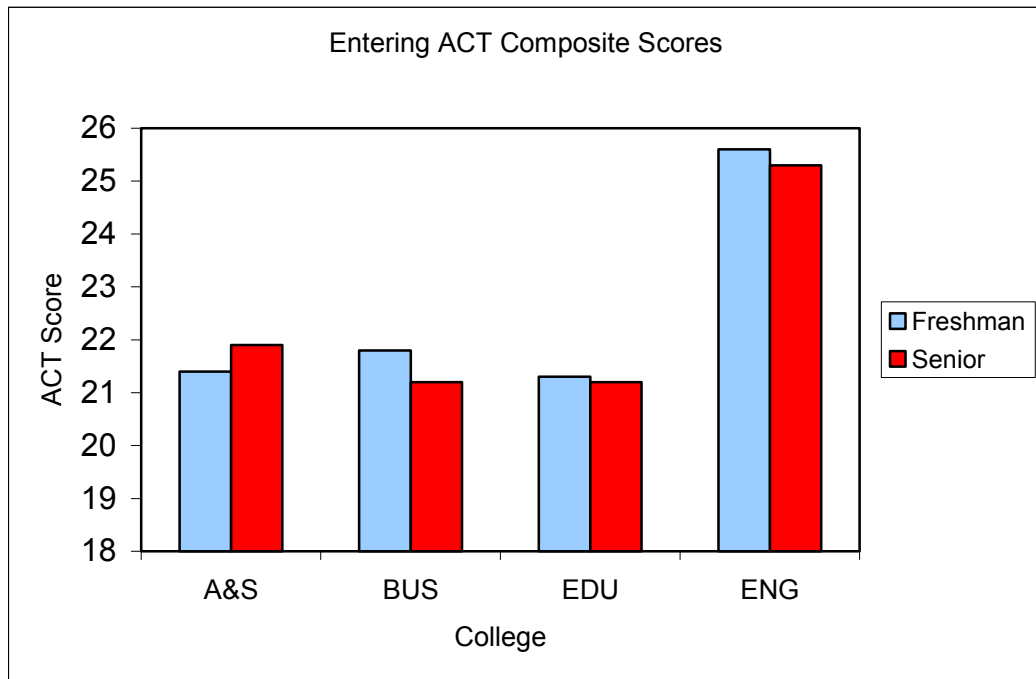
Figure 3



*Freshman vs Senior Testing*

To evaluate the sensitivity of the test and examine the potential improvement in critical thinking skills that might accrue from an undergraduate education at TTU, we compared the performance of entering freshman with seniors. The freshman and senior TTU students were both selected using a stratified random sample from the Colleges of Arts & Sciences, Business, Education, and Engineering. Figure 4 shows the entering composite ACT scores of students in each subgroup within the freshman and senior class and corroborates the equivalence of the freshman and senior samples on ACT scores.

Figure 4

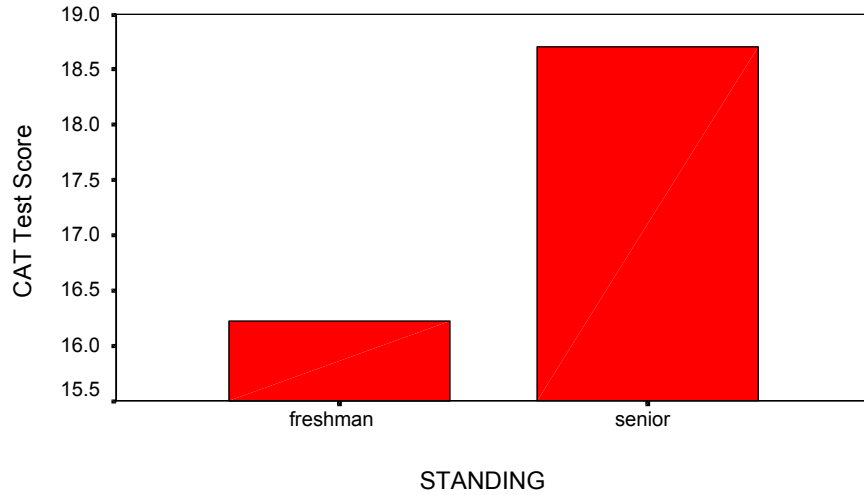


An analysis of covariance was performed on the freshman and senior test scores across the four colleges. Composite ACT score was used as a covariate to adjust for any potential differences between freshmen and seniors' entering ACT score. The results revealed a significant increase in critical thinking test scores from the freshman to the senior class ( $p < .001$ ). This effect is illustrated in figure 5.

Figure 5

### CAT Critical Thinking Test Scores

#### Freshman vs. Seniors



Although no significant interaction between college and class standing was found, figures 6 and 7 illustrate the potential gains in critical thinking broken down by college. Figure 6 shows the actual test scores while figure 7 shows scores that have been adjusted to remove the effects of differences in ACT scores across the samples.

Figure 6

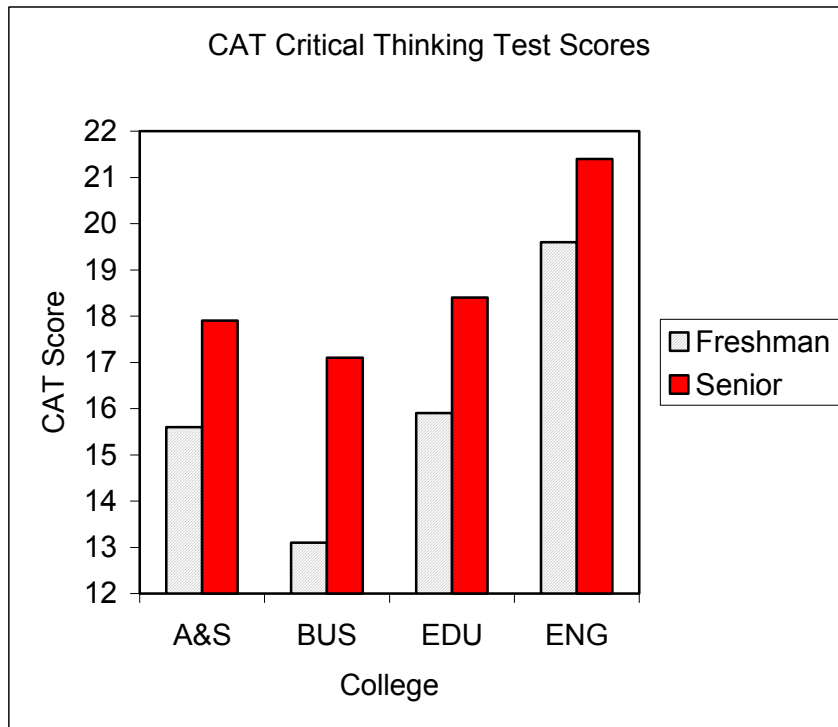
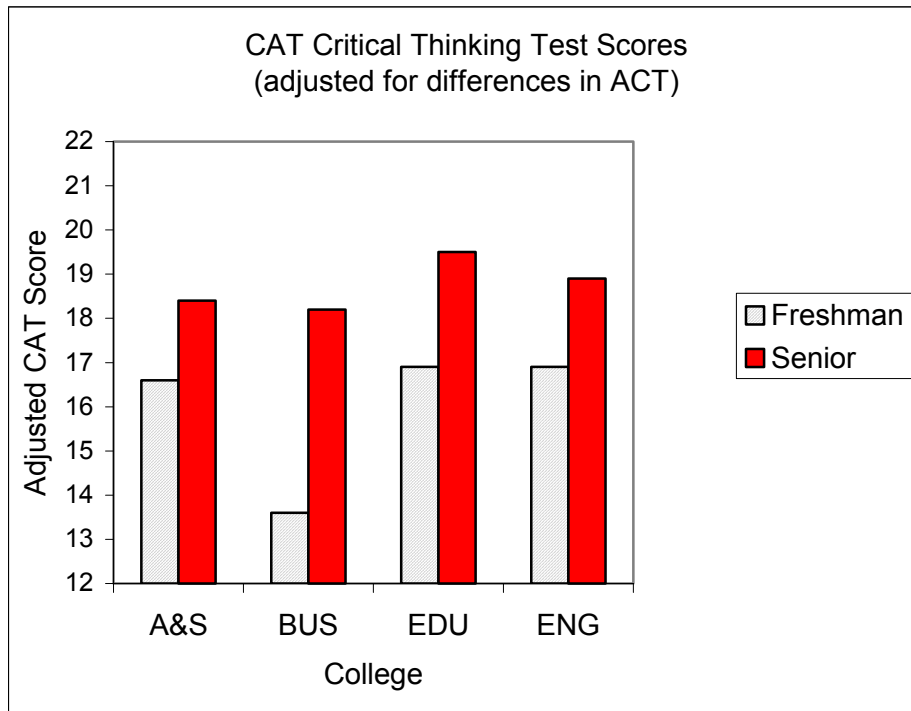


Figure 7



*Pre-testing and Post-Testing of Specific Courses*

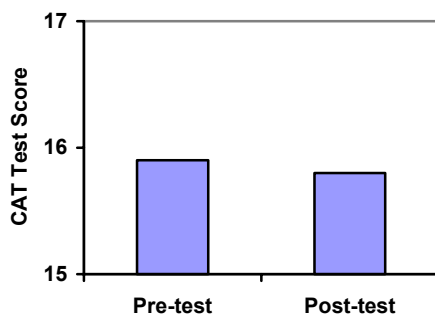
The administration of our CAT critical thinking test as a pre-test/post-test in specific courses was undertaken to investigate three issues. First, we wanted to know whether the test was sensitive enough to measure gains that could be achieved by students in a single course. If this is possible we could use the test results to identify courses or methods of instruction that would be representative of best practices in encouraging students to develop critical thinking skills. Second, assuming a course does not improve critical thinking skills, we wanted to know how reliable the test is when taken twice at two different times. Third, we wanted to evaluate the effects of a course designed to improve critical thinking and problem solving on test performance. If students who took a course designed to improve critical thinking and problem solving showed improvements on the test then this would help support the validity of the test.

To evaluate these issues, we administered the test in two different courses in the social sciences with the consent of the instructors. Both courses were junior level classes. One course was specifically designed to improve students critical thinking and problem solving skills (course #2) while the other course served as a control (course #1). Students in both courses took the pre-test during the first two weeks of the course and then took the post-test during the last week of classes. There was approximately 13 weeks between pre-test and post-test in each course. The pre-tests and the post-tests were scored by the same group of faculty from a broad spectrum of disciplines.

Sixteen students in the control course ( course #1) took both the pre-test and post-test. No significant change was observed in the performance of students between the pre-test and the post-test. The test-retest reliability coefficient = 0.6,  $p < .01$ . Overall test performance was remarkably similar on the pre-test and the post-test in the control course (see figure 8). Nineteen students in the critical thinking/problem solving class took both the pre-test and post-test. A significant improvement ( $p < .05$ ) was observed between scores on the post-test and scores on the pre-test for students in this course (see figure 9).

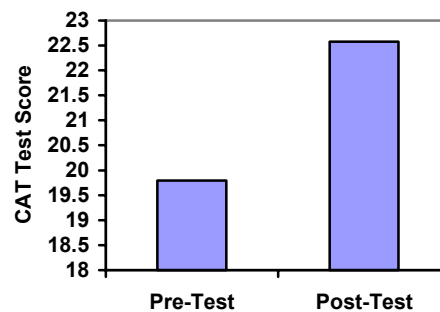
**Figure 8**

**Course #1**  
(Control)



**Figure 9**

**Course #2**  
(Problem Solving/Critical Thinking)



The results of the pre-test/post-test evaluations of these two courses indicate that the CAT Critical Thinking Test may be sensitive enough to detect the positive effects of an individual course on a student’s critical thinking and problem solving skills. These findings suggest that the test could be used to identify courses and pedagogies that promote critical thinking and problem solving skills. The findings don’t indicate how prevalent such courses may be across a university curriculum. The findings also indicate that test performance can be reasonably stable over time particularly when looking at group means. In addition, the findings provide another indication of test validity by showing that it is sensitive to training in critical thinking and problem solving.

### Current and Future Efforts

We continue to explore modifications in the test and scoring criteria that can improve the validity and reliability of the test. These efforts include the involvement of other interested institutions. This year, the University of Memphis will be pilot testing our critical thinking test. We are also exploring potential sources of funding that can be used to help involve other institutions in further testing and refinement of this instrument. We are currently examining whether there is any gender bias in our questions and we plan on extending this analysis to examine potential racial and ethnic bias although there is no indication such problems currently exist in the test.

## Appendix

### Current Critical Thinking Tests, Types, and Weaknesses (Based on Information Obtained from U.S. Department of Education, 2000)

Test	Type	Weaknesses
Academic Profile	Objective (different measures for humanities, social science, and natural science)	-Lacks sensitivity from freshmen to seniors -Proficiency levels change across various skills -Doesn't measure improvement in critical thinking from coursework
California Critical Thinking Dispositions Inventory (CCTDI)	75 Likert scale items	-Not a measure of critical thinking ability or skills
California Critical Thinking Skills Test (CCTST)	34 multiple choice items	-Low reliability -Low Item-Total Correlations -Principle component analysis did not support item classification -Some indications of cultural bias
Collegiate Assessment of Academic Proficiency (CAAP)	32 multiple choice items	-Limited to ability to analyze, clarify, evaluate, and extend arguments
College Outcome Measures Program (COMP)	60 multiple choice items	-DIF favors whites more often -Single factor according to factor analysis
Cornell Critical Thinking Test (CCTT)	50 item multiple choice	-Gender DIF analysis found 3 items favored males, while one favored females -Issues of validity
Critical Thinking Assessment Battery (CTAB)	Combination essay and objective	-No validity studies
Measure of Intellectual Development (MID)	Single essay	-Low reliability
ETS Tasks in Critical Thinking	Nine Essay/Short answer	-Low reliability -No longer available -Bias in scoring guide
Problem Solving Inventory (PSI)	35 Likert statements	-Not sensitive across academic levels -This is a test of confidence and attitude toward problem solving (not skills based)
Reflective Judgment Interview (RJI)	Standardized probe questions	-Gender biased -Limited range of critical thinking covered
Watson-Glaser Critical Thinking Appraisal (WGCTA)	80 multiple choice	-Possible test bias -Lack of cross-validation studies -6 low item correlations with total

## References

- Bloom, B. (1956). *Taxonomy of educational objectives*. New York: David McKay Co. Inc.
- Bransford, J. D., Brown, A. L. & Cocking, R. R. (Eds.) 2000. *How people learn: Brain, mind, experience, and school*. Washington, D.C., National Academy Press.
- Brenna, Mary A. (1991). Libraries for the National Education Goals *Eric Digest*, ERIC Identifier: ED345753. [http://www.ericfacility.net/databases/ERIC\\_Digests/ed345753](http://www.ericfacility.net/databases/ERIC_Digests/ed345753)
- Campione, J.C. & Brown, A.L. (1990). "Guided learning and transfer: Implications for approaches to assessment". In N. Frederiksen and R. Glaser et al. (Eds.). *Diagnostic monitoring of skill and knowledge acquisition* (pp. 141-172). Hillsdale, NJ: Erlbaum.
- Duchesne, R. E. (1996). "Critical thinking, developmental learning, and adaptive flexibility in organizational leaders". Paper Presented at the Annual Meeting of the American Association for Adult and Continuing Education in Charlotte, NC.
- Ennis, R. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership*, 44-48.
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction, Executive summary* Millbrae, CA: California Academic Press.
- Facione, P. A. (1998). *Critical thinking: What it is and why it counts*. Millbrae, CA: California Academic Press.
- Facione, P.A. Facione, N.C., Sanchez, C., & Gainen, J. (1995). The disposition toward critical thinking. *Journal of General Education*, 44(1) ERIC Identifier: ED 337 498
- Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. Baltimore: University Park Press.
- Forawi, S. A. (2001). Critical thinking and the national science standards. *Transactions of the Missouri Academy of Science, Annual*, 62.
- Halpern, D. E. (1993). Assessing the effectiveness of critical-thinking instruction. *Journal of General Education* 42 (4).
- Haas, P. F. and Keeley, S. M. Coping with faculty resistance to teaching critical thinking. *College Teaching* 46 (2), 63-69.
- International Society for Technology Education (2003) *National educational technology standards for teachers: Preparing teachers to use technology*. International Society for Technology Education: Eugene, Oregon. [http://cnets.iste.org/teachers/t\\_book.html](http://cnets.iste.org/teachers/t_book.html)
- Lidz, C. S. (1987). *Dynamic assessment: An interactional approach to valuating learning potential*. New York: Guilford Press.
- National Center for Educational Statistics (2002). *Statistical standards*. Washington, D.C. National Center for Educational Statistics. [http://nces.ed.gov/statprog/stat\\_standards.asp](http://nces.ed.gov/statprog/stat_standards.asp)
- National Survey of Student Engagement (2003) Perspectives from USM First-Year and Senior Students. [www.usm.maine.edu/testing/report1.doc](http://www.usm.maine.edu/testing/report1.doc)
- Paul, R. and Nosich, G. (1992). *A model for the national assessment of higher order thinking*. Washington, D.C.: National Center for Educational Statistics.
- Pawlowski, D. R. and Danielson, M. A. (1998). "Critical Thinking in the Basic Course: Are We Meeting the Needs of the Core, the Mission, and the Students?" Paper Presented at the Annual Meeting of the National Communication Association.

Quellmatz, E. (1985). Needed: Better methods for testing higher-order thinking skills. *Educational Leadership*. 29- 35.

Resnick, L.B. (1987). *Education and Learning to Think*. Committee on Mathematics, Science, and Technology Education, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.  
<http://www.nap.edu>

Samuels, M.T. (2000). Assessment of post-secondary students with learning difficulties: Using dynamic assessment in a problem-solving process. In C.S. Lidz & J.G. Elliott (Eds.). *Dynamic assessment: Prevailing models and applications* (pp. 521-542). Amsterdam: JAI/Elsevier Science

Siegel, H. (1988). *Education reason: Rationality, Critical thinking, and education*. New York: Routledge

Stein, B., Haynes, A., and Understein, J. (2003). "Assessing Critical Thinking." Paper Accepted for Presentation at SACS Annual Meeting in Nashville, Tennessee in December 2003.  
<http://iweb.tntech.edu/cti/SACS%20presentation%20paper.pdf>

Sternberg, R.J. & Grigorenko, E.L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge (UK): University of Cambridge.

U.S. Congress, (1994). Goals 2000: Educate America Act, H.R. 1804, Washington, D.C. U.S. Congress. <http://www.ed.gov/legislation/GOALS2000/TheAct/index.html>

White, E.M. (1993). Assessing higher-order thinking and communication skills in college graduates through writing. *The Journal of General Education* 42(2).

U.S. Department of Education, National Center for Education Statistics. *The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing*, NCEES 2000—172, prepared by T. Dary Erwin for the Council of the National Postsecondary Education Cooperative Student Outcomes Pilot Working Group: Cognitive and Intellectual Development. Washington, DC: U.S. Government Printing Office, 2000.

U.S. News and World Reports (August, 2003). America's Best Colleges: 2004, [http://www.usnews.com/usnews/edu/college/rankings/brief/libartco/libartco\\_campdiv\\_brief.php](http://www.usnews.com/usnews/edu/college/rankings/brief/libartco/libartco_campdiv_brief.php)

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education* 123 (1), 39-56.

Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.