In M. Iskander (ed.), *Innovations in E-learning, Instruction Technology, Assessment, and Engineering Education*, 79-82.

Assessing Critical Thinking in STEM and Beyond

Barry Stein, Ada Haynes, Michael Redding, Theresa Ennis, and Misty Cecil
Tennessee Technological University

***Abstract*- This paper reports the latest findings of a project to refine the CAT© instrument, a new interdisciplinary assessment tool for evaluating students' critical thinking skills. Tennessee Technological University partnered with six other institutions across the U.S. (University of Texas, University of Colorado, University of Washington, University of Hawaii, University of Southern Maine, and Howard University) to evaluate and refine the CAT instrument. The results of this project reveal the CAT instrument has high face validity when evaluated by a broad spectrum of faculty across the U.S. in STEM and non-STEM disciplines, has good criterion validity when compared to other instruments that measure critical thinking and intellectual performance, has good reliability, and good construct validity using expert evaluation in the area of learning sciences.** [1]

INTRODUCTION

There is little question that as a result of an increasingly technological and information driven society the ability to think critically has become a cornerstone to both workplace development and effective educational programs. Critical thinking is central to both the National Science Standards [1] and the National Educational Technology Standards [2]. Derek Bok [3] notes that over ninety percent of faculty in the U.S. feel that critical thinking is the most important goal of an undergraduate education. Increasingly, the importance of critical thinking/problem solving skills in the workplace is also being recognized. For example Halpern [4] argues, "virtually every business or industry position that involves responsibility and action in the face of uncertainty would benefit if the people filling that position obtained a high level of the ability to think critically". A recent Tennessee Technological University (TTU) survey of employers revealed that skills typically associated with critical thinking represented four out of the top five skills considered most important [5]. Similarly, a recent CNN poll of employers found that critical thinking is one of the top five essential skills in business that employers look for in potential job candidates [6].

Despite the central importance of critical thinking in the workplace and education, existing assessment tools are plagued by problems related to validity, reliability, and cultural fairness [7]. According to Bransford, Brown, and Cocking [8] "a challenge for the learning sciences is to provide a theoretical framework that links assessment practices to learning theory" (142).

Tennessee Technological University (TTU) has explored and tested a variety of instruments for assessing critical thinking and has spent six years developing the Critical thinking Assessment Test (CAT) that overcomes many of the weaknesses of existing instruments. TTU's approach has been to involve faculty in the identification of a core set of skills they believe to be an important part of critical thinking across all disciplines. An interdisciplinary team that included both STEM (Science, Technology, Engineering, and Math) and non-STEM faculty worked together with a member of the faculty who had extensive research and teaching experience in the area of critical thinking and problem solving to develop the test. The initial CAT test demonstrated high face validity and high criterion validity when correlated with other widely used measures of critical thinking and achievement.

The current paper examines work that TTU conducted in cooperation a diverse set of six schools nationwide to further test and refine this instrument. Through these efforts a test with high face validity, high construct validity, high reliability, and that is culturally fair has been developed. Pellegrino, Chudowsky and Glaser [9] note that in addition to face validity, assessment instruments need to be developed based upon principles of learning and cognition. To meet this standard, preeminent theoreticians and educators in the area of learning sciences participated in the current project and helped refine the instrument.

The CAT instrument is relatively unique in that it is one of just a few critical thinking assessment tools available that uses short answer essay responses to assess critical thinking. While subjective tests can be plagued by problems of scoring reliability, this has not been the case with the CAT instrument. Indeed, there appear to be several important advantages of using essay tests to assess critical thinking. Specifically, many authentic real-world situations that require critical thinking and problem solving do not have a simple answer or a simple set of alternatives from which to choose the best answer. Providing such alternatives on a test can frequently dramatically alter the original problem and greatly simplify the complexity of the task. Many would also argue that the reasons given to support a specific answer are often as important as the answer itself. The short answer essay format allows the CAT instrument to examine such reasoning and thought patterns. Some may consider the fact that the CAT instrument uses an institution's own faculty to score the short answer essay questions a shortcoming. TTU has found that using faculty graders is the best way to make faculty aware of student deficiencies in the area of critical thinking and to motivate faculty to consider changes in pedagogy that might improve students' critical thinking skills.

METHOD

Six institutions from diverse geographical regions across the U.S. and with different missions and student bodies participated in the project (the University

[1] Partial support for this work was provided by the National Science Foundation's CCLI Program under grant 0404911.

of Colorado, the University of Washington, the University of Hawaii, the University of Southern Maine, and Howard University). Each institution administered the CAT instrument to a sample of approximately 100 upper division students across STEM and non-STEM disciplines. Students were paid to participate in the testing.

Materials

The CAT instrument is a 15 question test with mostly short answer essay questions. The questions were specifically designed to evaluate the skill areas identified in Table 1. The questions explore various real-world situations that students find intrinsically interesting. A significant part of the test is devoted to solving a complex real-world problem that involves learning and applying new information – a testing practice sometimes referred to as dynamic assessment. The complex problem is designed to mimic many aspects of real-world problem solving. For example, in addition to solving a complex problem, decisions must be made about the about the types of articles that would be relevant to the solution of the problem as well as deciding how the solution to the problem might change if certain aspects of the problem situation were altered.

Testing Procedure

Students were tested in groups and the testing instrument was provided in booklet form with a package of additional readings. Each student was asked to complete all questions in the instrument and was given up to 60 minutes to complete the task. There was seldom a situation where any student did not have sufficient time to complete the test.

Scoring Procedure

A full day scoring workshop was held at each institution to assess student performance using a detailed scoring guide developed for the CAT instrument. Approximately 10-12 faculty from both STEM and non-STEM disciplines participated in the scoring workshop at each institution. Faculty were paid to participate in the scoring process.

During the scoring sessions, faculty received training on using the scoring guide immediately before each question on the test was scored. Each student's response to that question was then was then scored independently by two faculty members. If there was not agreement between the first two scorers, the question was scored by a third faculty member. The final score for each question was either the common score assigned by two graders or, in the case of three different scores, the final score was computed by averaging the three different scores. This process was repeated for each question on the test. Tests were frequently redistributed to insure that each faculty would see as many different tests as possible.

Faculty participants in the scoring workshop also completed several surveys after they finished scoring the tests. These surveys examined the face validity of the questions on the CAT instrument and whether the skills assessed by the CAT instrument were valid components of critical thinking (see Table 1).

TABLE 1
SKILL AREAS ASSESSED BY THE CAT INSTRUMENT

Separate factual information from inferences that might be used to interpret those facts.
Identify inappropriate conclusions.
Understand the limitations of correlational data.
Identify evidence that might support or contradict a hypothesis.
Identify new information that is needed to draw conclusions.
Separate relevant from irrelevant information when solving a problem.
Learn and understand complex relationships in an unfamiliar domain.
Interpret numerical relationships in graphs and separate those relationships from inferences.
Use mathematical skills in the context of solving a larger real world problem.
Analyze and integrate information from separate sources to solve a complex problem.
Recognize how new information might change the solution to a problem.
Communicate critical analyses and problem solutions effectively.

RESULTS

Evaluation of Skill Areas Targeted by the CAT Instrument

All of the skill areas targeted by the CAT instrument were perceived as valid components of critical thinking by most faculty who participated in the scoring workshops. The area with least agreement (79.4%) concerned using mathematical skills to solve a complex real-world problem (see fig. 1). The latter finding is mostly due to lower ratings by non-STEM faculty in some disciplines.

Evaluation of Question Face Validity

Faculty who participated in the scoring workshops were asked to evaluate the face validity of each question in the CAT instrument as a measure of critical thinking. Most faculty felt that the questions included on the CAT instrument were valid

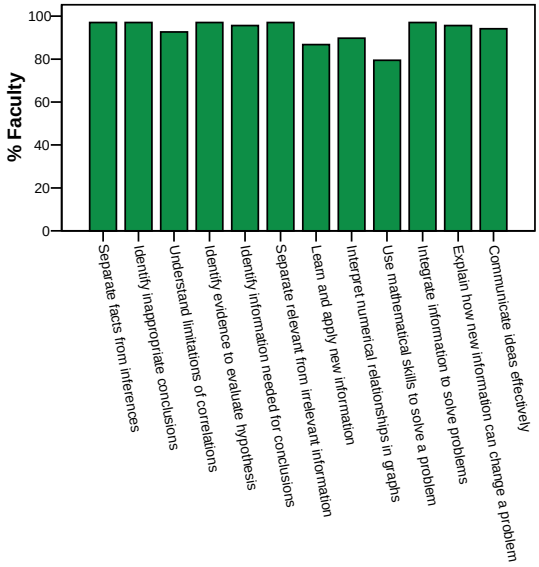

Fig 1. Percent of Faculty that Identify Areas Targeted by CAT as Valid Components of Critical Thinking

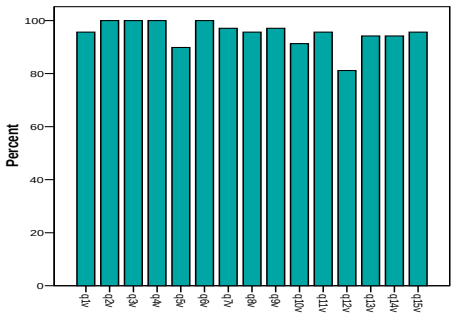

Fig. 2. Percent of Faculty Indicating Question Measures a Valid Component of Critical Thinking

measures of critical thinking (see fig. 2). The question with the lowest overall support (81.2%) involved using a mathematical calculation that was needed to help solve a complex real-world problem.

In addition to the quantitative survey data discussed above, qualitative data were collected from the local testing coordinators and the faculty scorers as well. The comments received from both faculty scorers and the local testing coordinators were overwhelmingly positive. Many faculty felt the test was very useful for identifying student weaknesses and others were interested in using the test within their disciplines to help explore ways of making program improvements.

Distribution of Student Scores and Internal Consistency

Fig. 3 shows the distribution of student scores (raw) on the CAT instrument against the normal curve. Scores ranged from a low of 6 to a high of 36.3. There was no evidence of a floor effect or a ceiling effect (lowest possible score = 0, highest possible score = 40).

The internal consistency of the questions was also examined and found to be reasonably high for an instrument of this type, $\alpha = 0.695$.

Correlation with other Measures of Student Performance

Performance on the CAT instrument has been correlated at various times with other measures of student performance including cumulative grade-point average (GPA), entering

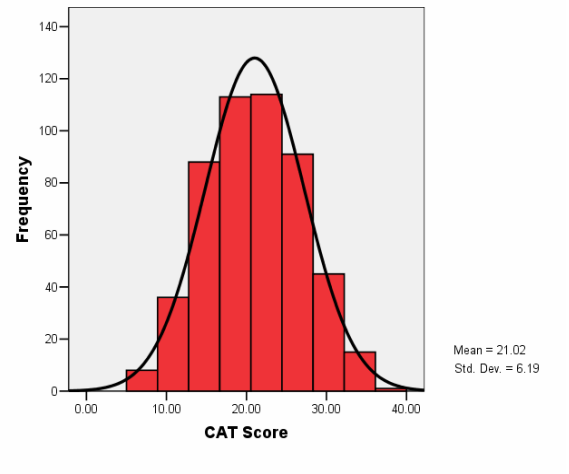

Fig. 3. Distribution of Student Scores

TABLE 2
CORRELATIONS FROM PREVIOUS WORK

	ACT	*AP*	*CCTST*
CAT	0.599*	0.558*	0.645*

* correlations significant, $p < .01$

TABLE 3
CORRELATIONS IN CURRENT STUDY

	SAT	*GPA*
CAT	0.527 *	0.345*

* correlations significant, $p < .01$

SAT or ACT scores, and performance on several other tests that might measure skills associated with critical thinking including the Academic Profile (AP) and the California Critical Thinking Skills Test (CCTST). The correlations from previous work appear in Table 2, while the correlations from the current project appear in Table 3. The correlations provide support for the criterion validity of the CAT instrument. At the same time, the magnitude of the correlations also indicates that the CAT instrument measures something different than the other assessment tools.

Scoring Reliability

Scoring Reliability is often a problem for essay type tests. However, extensive work has been done to refine the scoring rubric for this test to improve scoring reliability. The reliability of scoring was evaluated by examining scores assigned by faculty grader one and faculty grader two on each question. The average reliability of scoring across questions was 0.82.

Preliminary Analysis of Cultural Fairness

The CAT instrument in its current form is designed to assess critical thinking skills of students in U.S. colleges and universities. Nonetheless, it is still important to evaluate whether the instrument has bias against any of the subgroups within this population. Although more extensive analyses of any possible ethnic/racial/gender bias in the CAT instrument are planned, preliminary analyses of available data provide encouraging results. Multiple regression analyses reveal that once the effects of entering SAT score and GPA are taken into account, none of the predictors related to gender, ethnic background, or racial group are significant predictors of overall CAT performance.

CONCLUSION

There have been several significant outcomes of this project. For example, this project has been able to find a relatively high level of agreement across a diverse group of faculty from diverse disciplines about what skills might be considered valid components of critical thinking. The project has also found high levels of support for the face validity of questions in the instrument. There have been few attempts to establish such common ground among the disciplines in higher education about what constitutes “critical thinking.”

The project has also been successful in demonstrating the criterion validity of the CAT instrument when compared to

other measures of academic performance. Finally, the project has managed to circumvent a major problem for essay type tests by providing relatively high rates of scoring reliability with faculty who have had no prior experience grading this type of exam.

The refinement of the CAT instrument is also significant for another reason. The CAT instrument is one of the few interdisciplinary assessment instruments available that also provides an opportunity for faculty development. By participating in the scoring process, faculty become aware of their students' deficiencies and can begin to explore modifications in teaching methods that might address these weaknesses. This becomes increasingly important as accrediting agencies such as the Accreditation Board of Engineering and Technology (ABET) increase their focus on aspects of critical thinking such as life-long learning skills [10].

REFERENCES

[1] S.A. Forawi, "Critical Thinking and the National Science Standards." *Transactions of the Missouri Academy of Science, Annual*, 62.

[2] International Society for Technology Education. National Educational Technology Standards for Teachers: Preparing Teachers to Use Technology. Eugene, Oregon: International Society for Technology Education, 2003.

[3] D. Bok, Our Underachieving Colleges: A candid look at how much students learn and why they should be learning more. Princeton: Princeton University Press, 2006.

[4] E. Halpern, Assessing the Effectiveness of Critical-Thinking Instruction. Journal of General Education vol. 42., pp 251-255, 1993.

[5] B. Stein, A. Haynes, and T.Ennis, (2003). "Assessing Critical Thinking". Paper Presented at Commission on Colleges of the Southern Association of Colleges and Schools Annual Meeting in Atlanta, Georgia in December 2005.

[6] R. Castellini, "Survey: More to See Pay Increase in 2006. http://www.cnn.com/2006/US/Careers/01/04/cb.aol.survey/index.html.

[7] 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, NCES 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.

[8] J. Bransford, A. Brown, and R. Cocking, (Eds.) How People Learn: Brain, Mind, Experience, and School. Washington, D.C.: National Academy Press. 2000.

[9] J. Pellegrino, N. Chudowski, and R. Glaser (Eds.) Knowing What Students Know. Washington, D.C., 2001.

[10] Accreditation Board of Engineering and Technology, Criteria for Accrediting Engineering Programs: Effective for Evaluations During the 2006-2007 Accreditation Cycle. Baltimore, MD: Accreditation Board of Engineering and Technology, 2005