National Dissemination of the CAT Instrument: Lessons Learned and Implications

Barry Stein, Tennessee Technological University
Ada Haynes, Tennessee Technological University
Michael Redding, Tennessee Technological University

NSF recognizes a need to develop better tools to help evaluate higher order thinking skills such as critical thinking that are essential for STEM disciplines and for a competitive national workforce. This project addresses that need by providing an innovative tool to assess critical thinking skills and engage a broad spectrum of educators and researchers with its use. The project has benefited a broad range of institutions (community colleges, public and private four- year colleges and universities), and educational researchers that are engaged in efforts to improve students' critical thinking skills.

The overarching goal of the project is to nationally disseminate the CAT instrument to improve the assessment of students' critical thinking skills and to help identify effective practices for improving those skills. A key activity of the project involves the two-day regional training workshops that prepare representatives from collaborating institutions to administer and score the CAT instrument on their own campuses with their own faculty. The project also provides extensive consulting to institutions and educational researchers to develop efficient assessment plans, to provide statistical analyses of the CAT results, and to provide feedback about the accuracy of institutional scoring sessions.

Many educators believe that faculty must be involved in the assessment of student work in order for assessment to drive changes in teaching methods. The CAT instrument is scored by an institution's own faculty using short answer essay questions that reveal strengths and weaknesses in student's thinking. The regional training workshops prepare representatives from each institution to use the detailed scoring guide and carefully designed procedures for scoring to provide a consistent and reliable method of evaluating student responses that helps gain faculty buy-in and encourages faculty to explore methods improving student learning.

Dissemination has exceeded the original target by over 700% and has involved over 250 institutions. The CAT instrument is viewed as both a valid assessment tool and a faculty development tool that can improve instruction. Faculty participation in CAT scoring sessions increases the use of active learning strategies and reduces the emphasis on the rote retention of factual information. Educational researchers have found the CAT is sensitive to changes that occur in formal and informal learning situations over a semester or less. A wide variety of NSF projects have found significant gains in CAT scores demonstrating the transfer of skills developed within a discipline's content to the interdisciplinary content of the CAT instrument.

The project has enhanced the capacity to assess critical thinking skills and evaluate educational practices designed to improve those skills across a broad range of higher educational institutions. Over 250 institutions have collaborated including HBCU's, other minority serving institutions, community colleges, as well as a broad range of public and private universities. Involvement has been supported through electronic dissemination (e.g., www.CriticalThinkingTest.org, YouTube, social media, email), conference presentations, professional publications, and regional training workshops. Consequently, a growing body of educational research is using the CAT instrument to evaluate innovative educational practices to improve critical thinking skills.

Background

Various constituent groups in our society are in widespread agreement about the importance of critical thinking. For instance, the Higher Education Research Institute (HERI) found that over 99% of faculty across the country felt that teaching critical thinking is "essential" or "very important."[1] Employers also recognize the importance of critical thinking and problem solving skills. A recent survey by the American Association of Colleges and Universities (AACU) found that 75% of employers want colleges to place more emphasis on critical thinking, real world problem solving, communication, and creativity. Furthermore, 93% of employers felt that these skills were more important than a specific college major.[2] Several reports from the National Research Council also identify critical thinking, non-routine problem solving, and communication skills as essential for success in 21st century careers.[3,4]

Despite the clear agreement on the importance of critical thinking and problem solving skills, these skills are not frequently assessed in most college courses.[5,6] Higher education courses have a pervasive tendency to emphasize the rote retention of factual information.[7,8]

One explanation for the widespread emphasis on the rote retention of factual information is that constructing a factual knowledge test is much easier than designing an assessment that evaluates critical thinking skills. Most faculty have little or no training in developing classroom assessments that promote the development of critical thinking skills.[9,10,11] Faculty are often unaware of the impact of assessment on student learning.[10] Unfortunately, when faculty use rote retention tests to assess student performance, they are inadvertently encouraging students to devote most of their time and energy to memorizing information. Excessive reliance on factual knowledge assessments can sabotage the impact of using active learning pedagogies and may lead to dissatisfaction with programs of study and lack of persistence in that field of study. How faculty assess student learning can have a greater impact on learning than the particular teaching pedagogy that is used.[12]

The Approach

The CAT instrument was designed to assess a broad range of skills that faculty across a wide range of disciplines and institutions associate with critical thinking and that are considered essential for success in life and work in the 21st century.[2,3,4] These skills are identified in Table 1 and include both critical and creative thinking skills as well as non-routine problem solving and effective communication.

Table 1: Skill Areas Assessed by the CAT Instrument

Evaluating Information
Separate factual information from inferences
Interpret numerical relationships in graphs
Understand the limitations of correlational data
Evaluate evidence and identify inappropriate conclusions
Creative Thinking
Identify alternative interpretations for data or observations
Identify new information that might support or contradict a hypothesis
Explain how new information can change a problem
Learning and Problem Solving
Separate relevant information from irrelevant information
Integrate information to solve problems
Learn and apply new information
Use mathematical skills to solve real-world problems
Communication
Communicate ideas effectively

The CAT instrument was designed to give faculty a clear understanding of their student's strengths and weaknesses in the skill areas above using mostly short answer essay questions that are scored by an institution's own faculty using a very detailed scoring guide. The questions used in the CAT involve general content but can be easily adapted to each discipline's content. These features enable the CAT instrument to be used as a model to help faculty design better course assessments in their own disciplines.

The CAT instrument is separated into two parts. In part I of the CAT instrument, a variety of real world scenarios are used to present information (or data) and possible interpretations of that information. The principles of dynamic assessment are used to provide increasingly deeper prompts to elicit as much critical and creative thinking as possible.[13] Thus, students might be asked how strongly information supports an idea or hypothesis, followed by a prompt to identify other reasonable alternative interpretations of the data or information, followed by a prompt to identify new information or data needed to evaluate the alternative explanations.

In part II of the CAT, a non-routine real-world problem scenario is presented. This problem is accompanied by a simulated Google or database search with resulting article titles that must be evaluated for relevance to the particular problem scenario. Students are then given access to the short articles and must learn and apply the new information to the problem situation to identify and explain the best solution for the problem. Deeper learning is examined with an additional prompt that asks how the recommended solution would change if there were significant changes to the problem constraints.

A variety of studies were conducted in previously funded work to help establish the validity and reliability of the CAT instrument.[14] The high face validity of the instrument has contributed to strong interest in using the CAT among faculty who might otherwise be skeptical of the value of assessment tools. The cultural fairness of the test has been evaluated in three ways. A multiple regression analysis of CAT performance revealed that once the effects of entering SAT score, GPA, and whether English was the primary language were taken into account, neither gender, race, nor ethnic background were significant predictors. A cultural differential item functioning (DIF) analysis was performed to examine possible question bias; a review of the DIF results did not reveal any items with prevalent cultural bias. In addition, qualitative data from faculty were collected regarding cultural and gender fairness. In the development stage, questions where faculty perceived bias were examined and modified as necessary.

New methods had to be developed to ensure that faculty at each collaborating institution could reliably and accurately score the test. The scoring guide underwent considerable revision in the early stages to ensure that faculty could accurately score ambiguous student responses. Each question has a unique scoring rubric that is customized to the question content and skills being evaluated. We learned early on that to achieve reliable scoring the scoring process should limit the number of factors being examined in each question. We also developed a training/scoring process in which faculty would be trained to score a question and then immediately score a set of student responses for that question. Combining the training and scoring helped avoid much of the forgetting that would occur if the two steps were separated by a longer period of time. Within this training/scoring process faculty read selected student responses aloud to the group and practice applying the scoring rubric to make sure that everyone is calibrated in using the scoring rubric for each question. We have found that this calibration activity is essential to maintain accurate and reliable scoring.

Each student response is scored by a minimum of two scores and a third scorer is used if the first two scorers disagree. If two scorers agree, then that score is assigned to the response. If all three scorers disagree, then the mathematical average of the three scores is assigned to the response. During the scoring workshops, faculty enter their assigned scores in a predesigned scannable score sheet. The actual determination of a score for each response is done later when automated scanning and analysis is performed. The reliability of first and second scores has been quite good with our latest analyses indicating R = .92 (n = 14,600 tests). One of the lessons we learned is that high reliability in scoring across first and second scorers does not always mean that responses have been accurately scored. We have rescored and continue to rescore randomly selected subsets of tests from each institutional scoring session using our expert scorers. These accuracy checks are then used to provide feedback to each institution with instructions on what may have contributed to inaccuracies if applicable. We thought that this type of feedback would only be needed in the early stages of institutional use, but we learned that institutions that have scored accurately for several years may suddenly have problems because of changes in personnel or due to deviations from our prescribed scoring procedures. We now realize that accurate scoring requires the constant monitoring of accuracy with feedback. Overall, we have found that the average error in institutional scoring sessions is about 5.4% (n = 280 institutional scoring sessions).

We developed a two-day intensive regional training workshop to prepare representatives from each institution to lead faculty through the scoring process using the scoring guide and calibration process described above. These workshops turned out to be essential to ensure accurate scoring. We originally thought that each new training workshop would only include representatives from new institutions, however, that was not the case. People frequently wanted to come back to the training workshops to refresh their understanding after a few years and many of the people we trained moved on to other institutions and needed to be replaced by new trainees.

Since participants in our regional training workshops would not always be able to immediately follow-up those sessions with their own scoring workshops, we developed a narrated multimedia training module that reviews how to score each question with some sample responses. This training module turned out to be more important than we expected, and we later found that some institutions used the training module to help conduct scoring sessions at their institution with considerable success.

Findings

We have collaborated with over 250 institutions across the United States, including a wide range of community colleges, 4-year institutions, and R-1 universities (of these 22 are Hispanic Serving, 17 are Historically Black Colleges and Universities, 1 is a Tribal College, and 1 is a Women’s College). This is over 700% of the original target of the project. Thus far over 700 faculty members, administrators, and staff have been trained to lead scoring workshops on their own campuses. It is estimated that over 4,000 faculty have participated in CAT scoring sessions at institutions around the country and over 135,000 student tests have been administrated. Strong interest in using the CAT instrument is allowing the core functions of the project to become self-sustaining.

A wide variety of institutions and other NSF projects are using the CAT instrument to assess the effects of various types of high-impact and active learning strategies on improving students' critical thinking skills and to validate other instruments. The sensitivity of the CAT instrument to these effects has been particularly useful in identifying effective practices. Additional evidence for the **criterion validity** of the CAT instrument is emerging from collaborations with other NSF projects that have found significant gains on specific CAT skills, which were targeted by NSF projects.[15,16,17]

The CAT has been found to be sensitive enough to assess changes in critical thinking skills in formal and informal learning environments in as few as three weeks.[15] The CAT can also be used to evaluate changes over a sequence of courses or a program of study.[14] No evidence of a floor effect or a ceiling effect (lowest possible score = 0, highest possible score = 38) exists at any of the institutions tested, including community colleges and Ivy League universities.

Thus far, the CAT team has collaborated with about 40 NSF projects. About half of these projects are showing significant gains on the CAT instrument. These results are quite encouraging and demonstrate the potential usefulness of the CAT instrument as a tool for measuring treatment effects in NSF projects that have targeted critical thinking/real-world problem solving. The use of the CAT instrument in these projects is advancing the knowledge base of research on effective STEM education. A growing body of published research by these projects reveals how the CAT is being used to assess project outcomes and how STEM education can improve students' critical thinking. In some cases, NSF projects are showing gains in students' critical thinking in one course that are equal in magnitude to gains across an entire 4-year college education. These projects' use of the CAT is helping build the knowledge base of effective STEM practices and is showing that considerably more can be done to improve students' critical thinking skills.

The CAT instrument is also being used at many institutions for faculty development purposes. Recent research indicates that participation in CAT scoring sessions has a positive effect on the willingness of faculty to use more active learning strategies and to place less emphasis on the rote retention of factual information in course assessments.[18] Some faculty are actively involved in using the CAT instrument as a model for developing better course assessments in their disciplines. We are currently exploring ways to help faculty develop better assessments for their courses that emphasize critical and creative thinking.

Bibliography

1. Higher Education Research Institute, (2009). *The American College Teacher: National norms for 2007-2008 HERI Faculty Survey.* Los Angeles: Higher Education Research Institute.
2. Association of American Colleges and Universities. (2013). *It Takes More than a Major: Employer priorities for college learning and student success*. Retrieved from AACU website: http://www.aacu.org/leap/documents/2013_EmployerSurvey.pdf
3. Pellegrino, J.W. & Hilton, M.L. (Eds.) (2012) Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. National Research Council Report.
4. Koenig, J.A., Rapporteur (2011). Assessing 21st Century Skills: Summary of a Workshop. National Research Council Report.
5. Anderson, L., & Sosniak, L. (1994). *Bloom's taxonomy: A forty-year retrospective.* Chicago, IL: University of Illinois Press.
6. Stein, B., Haynes, A., & Harris K. (2009). Assessing and improving critical thinking using the CAT instrument, *ABET Best Assessment Processes Symposium XI.* Symposium conducted at the meeting of Accreditation Board of Engineering and Technology, Indianapolis, IN. Retrieved from http://www.tntech.edu/cat
7. Stein, B., Haynes, A., & Harris K. (2009). Assessing and improving critical thinking using the CAT instrument, *ABET Best Assessment Processes Symposium XI.* Symposium conducted at the meeting of Accreditation Board of Engineering and Technology, Indianapolis, IN. Retrieved from http://www.tntech.edu/cat
8. Kvale, S. (2007). Contradictions of assessment for learning in institutions of higher learning. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term,* (pp. 3–13). New York, NY: Routledge.
9. Fox, M. A., & Hackerman, N. (2003) Evaluating and Improving Undergraduate Teaching in Science, Technology, Engineering, and Mathematics,
10. Hutchings, P. (2010). *Opening doors to faculty involvement in assessment.* National Institute for Learning Outcomes Assessment. Champaign, IL: University of Illinois.
11. Petress, K. (2007). How to make college tests more relevant, valid, and useful for instructors and students. *College Student Journal,* 42, 1098-1011.
12. Boud, D., & Falchikov, N. (2007.) Introduction: Assessment for the longer term. In D. Boud & N. Falchikov (Eds.), *Rethinking Assessment in Higher Education: Learning for the Longer Term* (pp. 3-13). New York, NY: Routledge.
13. Lidz, C. S. (1987). *Dynamic assessment: An interactional approach to valuating learning potential*. New York: Guilford Press.
14. Stein, B. Haynes, A. & Redding, M.(2007). Project CAT: Assessing critical thinking skills. In D. Deeds and B. Callen *Proceedings of the National STEM Assessment Conference.* NSF and Drury University
15. Brittany J. Gasper, Stephanie M. Gardner (2013). Engaging Students in Authentic Microbiology Research in an Introductory Biology Laboratory Course is Correlated with Gains in Student Understanding of the Nature of Authentic Research and Critical Thinking. Journal of Microbiology & Biology Education, vol. 14, no.1.
16. Frisch, J. K., Jackson, P. C., & Murray, M. C. (2013). WikiED: Using web 2.0 tools to teach content and critical thinking. Journal of College Science Teaching, 43 (1), 70-80.K. Schneider, A. Bickel, and A. Morrison-Shetlar (2014). Planning and Implementing a Comprehensive Student-Centered Research Program for First-Year STEM Undergraduates. Journal of College Science Teaching, 44, (3), 37-43.
17. Gottesman, A.J. & Hoskins, S.G. (2013). CREATE Cornerstone: Introduction to Scientific Thinking, a new course for STEM-interested freshmen, demystifies scientific thinking through analysis of scientific literature. CBE-Life Sciences Education, vol. 12 no. 1, p 59-72.
18. Lisic, E. (2015). Creating change: Implementing the critical thinking assessment test (CAT) as faculty development to improve instruction. Ph.D. Dissertation.

Biographical Information

Barry Stein is chairperson and professor of psychology and Co-Director for the Center of Assessment and Improvement of Learning. He is a cognitive psychologist who has been active in the area of learning and cognition for over 35 years. He is co-author with John Bransford of *The Ideal Problem Solver*. He is the PI for three NSF grants related to the CAT instrument. Ada Haynes is a professor of sociology and Co-Director of the Center for Assessment and Improvement of Learning. She was formerly the Director of TTU's Quality Enhancement Plan for improving students' critical thinking/real-world problem solving through active learning. Her use of critical thinking techniques in the classrooms has been responsible for her winning the TTU Outstanding Faculty Teaching Award, the Honors Faculty Award, and the Service-Learning Teaching Award. She is the Co-PI for three NSF grants related to the CAT instrument and is an evaluator for other NSF grants. Mike Redding is a professor of Biology and a core STEM faculty member at TTU. In addition to his teaching and research, he has worked extensively with local, state, and international science fairs to increase student interest in science careers and improve science education. He is the Co-PI for three NSF grants related to the CAT instrument.