

Quantifying Assessment Of Undergraduate Critical Thinking

Michael Grant, University of Colorado Boulder, USA
Marshall Smith, University of Colorado Boulder, USA

ABSTRACT

Enhancing students' critical thinking capabilities stands as the top goal of undergraduate education, according to faculty from many universities. We assessed the change in critical thinking skills with a sample of 176 students enrolled at either the University of Colorado Boulder (UCB) or Colorado College (CC) by employing the Critical-thinking Assessment Test (CAT) developed with collaboration and support from the National Science Foundation. Students' critical thinking progress was compared by assaying skills during the first and last weeks of the term in classes that expressly emphasized: (1) critical thinking, or (2) civic engagement, or (3) where, according to the class instructors, neither was a point of major emphasis. CAT scores improved significantly for students at both institutions, in different categories of class types, and over the dramatically different lengths of terms (3.5 weeks at CC vs 15 weeks at UCB). Our research contributes to an understanding of changes in critical thinking as part of the undergraduate experience. We demonstrate that the CAT instrument can be an effective tool for assessing critical thinking skills across very different institutions of higher education.

Keywords: Critical Thinking; Learning; Assessment; CAT

INTRODUCTION

The nature and scope of critical thinking in higher education plays a prominent role in scholarly work (e.g. Moon, 2008; Franco, 2016) and practical application. Development and enhancement of undergraduate students' critical thinking skills has been prominent on the list of major objectives for higher education for many years (e.g., Bailin & Siegel, 2003; Mayweg-Paus, Thiebach, & Jucks, 2016; Kivunja, 2015). According to the Higher Education Research Institute's (HERI) faculty survey, "Full-time faculty with undergraduate teaching responsibilities overwhelmingly agree (99.1%) that developing students' ability to think critically is a "very important" or "essential" goal". In fact, since the HERI Faculty Survey first introduced this question in 2004, nearly all faculty have consistently rated this goal as "essential" or "very important" (Eagan, et al., 2014). Further, "nationwide polls have found that more than 90 percent of faculty members in the United States consider [critical thinking] *the most* important purpose of undergraduate education" (emphasis in original, Bok, 2006). In their nationally representative survey of faculty, Lindholm, Szelenyi, Hurtado, and Korn (2005) similarly found that developing critical thinking ability was one of "the most universally emphasized goals for undergraduate education..." Overall, according to higher education faculty, improving critical thinking stands as the most significant function of the undergraduate experience.

Partly in recognition of multiple views of what constitutes critical thinking and how to measure it (e.g., Hatcher, 2011; Mulnix, 2012; Abrami, et al. 2008), the Critical-thinking Assessment Test (CAT) was developed over several years with NSF support and oversight to measure an explicitly broad, inclusive view of the concept that also distinguished among specific aspects of critical thinking (Stein, Haynes, Redding, Ennis, & Cecil, 2007). Here we utilize the inclusive, operational definition of critical thinking that CAT was developed for: "The CAT instrument is designed to assess a broad range of skills that faculty across the country feel are important components of critical thinking and real world problem solving. The test was designed to be interesting and engaging for students. All of the questions are derived from real world situations." We here paraphrase four critical thinking skills CAT assesses according to the CAT Overview website (<https://www.tntech.edu/cat/about/skills>): **(1) Evaluating Information** to include separating factual information from inferences, interpreting simple numerical relationships in graphs, understanding the limits of correlational data and evaluating evidence to identify inappropriate conclusions; **(2) Creative Thinking** to identify

alternative interpretations for data or observations, recognizing new information that might support or contradict an hypothesis, explaining how new information can change a problem; (3) **Learning and Problem Solving** skills to separate relevant and irrelevant information, to integrate multiple sources of information to solve problems and to learn and apply new information and use simple mathematical skills to solve real-world problems; (4) **Communication** to focus on communicating ideas effectively and precisely.

The CAT was developed to assess the wide range of skills consistent with the various predominant components of critical thinking used by critical thinking scholars while emphasizing those skills identified as most important by faculty. For example, Bailin and Siegel (2003) problematized existing definitions of critical thinking including the view that such thinking "is a normative concept that requires mastery of context-specific knowledge to evaluate specific beliefs, claims, and actions" (Abrami, et al., 2008). By providing students with specific knowledge related to a real world scenario, the CAT protocol requires carefully designed steps which faculty collectively use to assess this component of critical thinking. A more detailed discussion of the CAT instrument is available at <https://www.tntech.edu/cat/about>.

The impact of college on undergraduate students generally, and critical thinking skills especially, has been the basis for considerable research for several decades (Pascarella & Terenzini, 1991, 2005; Moon, 2008; Franco, 2016; Ennis, 2008). Only recently has the ability of colleges and universities to improve the critical thinking skills of students been questioned and challenged (Arum & Roksa, 2011; Bok, 2006).

Bok (2006) emphasizes the importance of critical thinking while also claiming that colleges and universities are largely ineffective in developing those skills. Similarly, Arum and Roksa (2011), assert that some higher education institutions are not successfully promoting critical thinking and other complex reasoning skills among undergraduates. They based their conclusion primarily using results of the Collegiate Learning Assessment (CLA) instrument, a widely used assessment test in higher education. In contrast, Shavelson (2010), also using CLA data, found notable improvement in critical thinking and other skills for seniors in comparison to freshmen at a number of institutions. The CLA has been criticized for its overly general approach to measuring critical thinking (Possin, 2008) as well as the applicability of the results (Shavelson, 2010). Further, just prior to publication of Arum and Roksa's book, Abrami et al. (2008) completed a meta-analysis of research exploring the "impact of instruction on the development of critical thinking skills and dispositions". They found a generally positive "effect of instruction on students' CT skills... [but] ... also uncovered some evidence of negative effects". More granular analysis of their results showed that course content, curriculum, instructor training in teaching CT skills, were significant in the size of the positive effects. "To maximize impact requires both the willingness to incorporate CT instruction and explicit strategies and skills to do it effectively" (Abrami, et al., 2008). They concluded that the "[t]ype of critical thinking intervention and pedagogical grounding were substantially related to fluctuations in critical thinking effects sizes" and that "... improvement in students' critical thinking skills and dispositions cannot be a matter of implicit expectation. ... [e]ducators must take steps to make CT objectives explicit in courses ...". The CAT instrument was created to address and mitigate the deficiencies of previously available instruments.

Evaluation of changes in critical thinking skills over the course of a single term remains an under-studied time frame. The CAT instrument has been used to assess student improvement in critical thinking skills over the course of one term for inquiry-based approaches to science education (Gasper, Minchella, Weaver, Csonka, & Gardner, 2012; Gasper & Gardner, 2013); in a science class concentrated on using "a narrowly focused set of readings to promote development of analytical skills" (Gottesman & Hoskins, 2013); and in a class structured around inquiry-driven Web 2.0 instruction activities (Frisch, Jackson, & Murray, 2013). Students from these studies showed improvement in aspects of critical-thinking as measured by the CAT. These results were used to improve future student learning by provoking changes to courses and/or the curriculum. While useful, these few studies leave considerable unexplored perspectives of subjects and contexts for measuring the actual changes in critical thinking skills, occurring over a single term. In this study, our goal was to empirically quantify the extent to which improvement in critical thinking skills could be measured, parsed by type of institution and whether a class explicitly focused on critical thinking development or not.

METHODS

Our research explored the effects of different pedagogical strategies for classes on critical thinking skills at these two very different institutions: Colorado College (CC), a highly selective, liberal arts college which employed a strategy of one subject per 3.5-week blocks (=terms herein) and the University of Colorado Boulder, an R1, comprehensive public research university employing a traditional 15-week term (semester) with multiple courses taken simultaneously. Our study aimed to document any change, positive or negative, in critical thinking skills for students at both institutions over the duration of those single terms. Further, we investigated three categories of classes, as identified by stated pedagogical strategies by the class instructors themselves, to assess the degree of effectiveness in student improvement in critical thinking skills partly as a result of those class experiences.

Students at both institutions in classes completed CAT at the beginning and end of their respective terms. In each case faculty were contacted and asked if we could invite their students to participate in the research. Students who volunteered to participate were given a \$50 honorarium and completed the CAT during personal time, typically in about 40 minutes.

Critical thinking classes were chosen based on faculty explicitly identifying critical thinking as a major focus of that class either by virtue of the course content and/or the teaching strategy. For the Civic Engagement classes, the Institute of Ethical and Civic Engagement at UCB supplied a list of classes to be considered for inclusion where the faculty, in their syllabi, described a primary focus on civic engagement or service learning. While service learning and civic engagement classes can be different, we grouped them together because they strongly overlapped in pedagogical strategy and focus. The control category consisted of classes where faculty agreed that neither of the other two dimensions were particularly emphasized. These included low level music appreciation classes and first term language courses. While these 'control' classes clearly may lead to improvements in critical thinking, the faculty in charge did not explicitly emphasize improvement in critical thinking. A similar approach was developed and applied for selection of classes at CC.

The main goal of our research project: to quantify students' critical thinking skills and improvement thereof as measured by the CAT at the beginning of a term contrasted to their performance level at the end of that term. We employed comparative statistical analyses of students' performance on CAT by class type and institution as well as conducting more fine grained analysis of some specific components of critical thinking.

The extensive development, testing and validation processes that ultimately led to the CAT instrument we used provide compelling evidence for the effectiveness of the tool. CAT appears to be essentially free of well-known drawbacks of many general use instruments such as ethnic biases, ceiling effects, localized special knowledge, etc. Full explanations and discussion of these issues can be found at (<https://www.ntech.edu/cat/about>).

We chose to use the CAT instrument for our project because of the extensive design, validation, and dissemination efforts, as supported by the National Science Foundation for the past several years (Stein, Haynes, & Redding, 2006; Stein, et al. 2010; Stein & Haynes, 2011). Development of the CAT focused on improving on existing instruments that have "very narrow definitions of critical thinking ... and/or [that] did not sufficiently involve faculty in the evaluation of student performance" (Stein, et al., 2006). We judged the faculty assessment of students' work on the CAT to be an extremely important dimension of this study as the process directly connected faculty with a broad range of actual student responses on the instrument.

Focused research demonstrates that the CAT instrument "has high face validity when evaluated by a broad spectrum of faculty across the U.S. ..., has good criterion validity when compared to other instruments that measure critical thinking and intellectual performance, has good reliability, and good construct validity..." (Stein, et al., 2007). In particular, in comparison to the CLA, the CAT provides a more detailed analysis of multiple aspects of critical thinking. We also highlight here the strong negative correlation of CAT scores with those from the National Survey of Student Engagement (NSSE) measuring emphasis on memorization (Stein, et al., 2006).

The CAT was developed specifically to "support the implementation of pedagogical improvements within and across disciplines" (Stein, et al., 2006). The CAT aims to facilitate improvement in developing critical thinking skills through

changes in teaching methods as well as changes by university administrations. Despite the considerable resources devoted to its development, peer-reviewed published results showing how the CAT instrument has been utilized to evaluate critical thinking remain limited. The CAT instrument may also be used as an impetus to encourage faculty to focus on more explicit development of critical thinking skills in their curricula and pedagogical approaches (Gasper & Gardner, 2013; Rowe, et al. 2015).

We focused our analysis on two particular aspects of the students' performance on the CAT. Reflecting our factorial, experimental design, we analyzed CAT results for all students as a single sample with the first factor divided into beginning and end of term scores, and the second factor divided into scores on the three types of classes to generate a two factor, unbalanced, pure Model I design with appropriate follow-ups where needed.

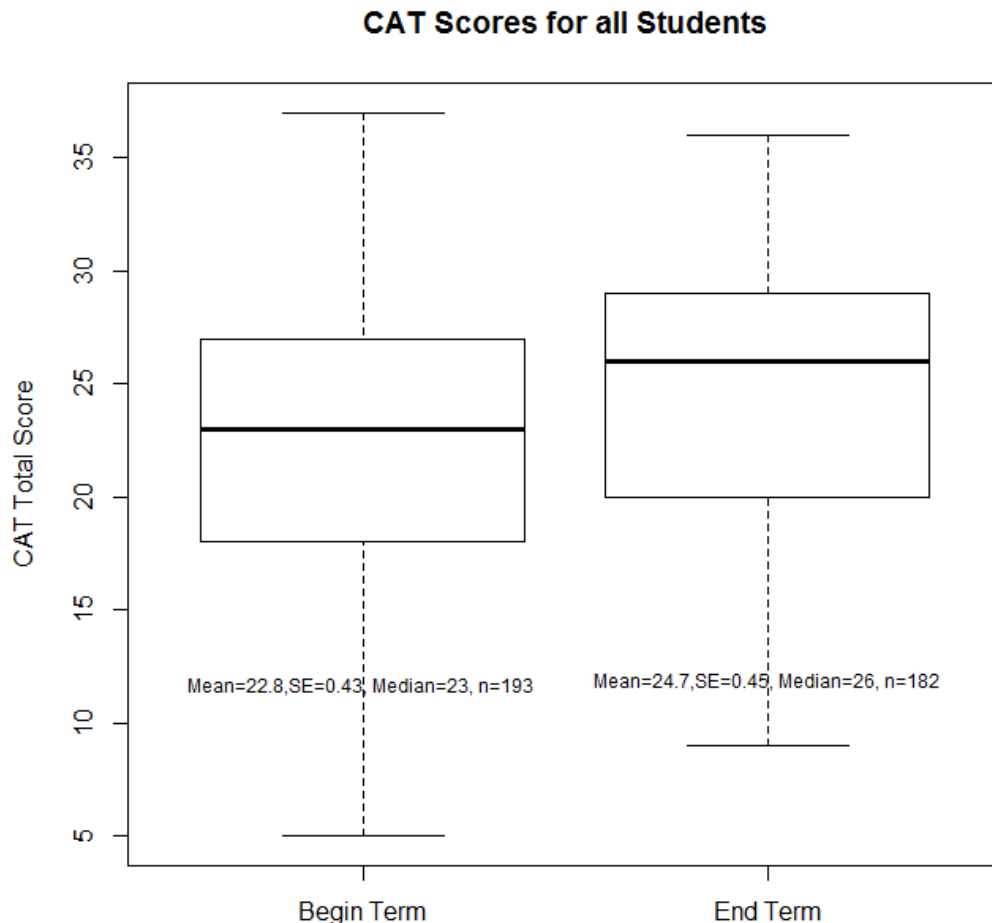
We also analyzed performance on both the overall score on the CAT, as well as performance on individual CAT questions because the various questions invoked different types of critical thinking skills. The maximum possible score for individual questions ranges from one to five points. The maximum possible score overall is thirty-eight points.

RESULTS

CAT Performance for All Students

Figure 1 shows the overall improvement in CAT scores over the duration of the term for all students in all classes at both institutions.

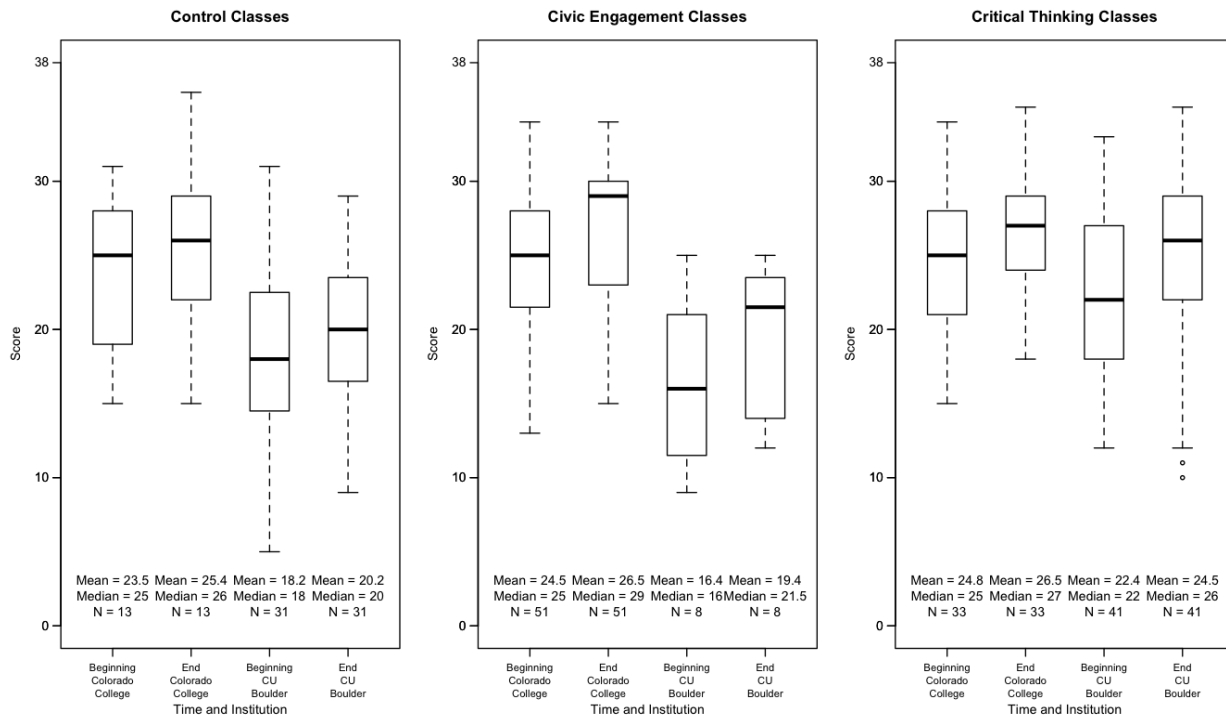
Figure 1. CAT Overall Raw Scores for Beginning Versus End of Term for All Students



We show, in Figure 1, the pattern of results over time for the highest, most inclusive level of analysis. Clearly, this large sample of students demonstrates significant improvement in a single term ($p < 0.002$) in critical thinking as assessed by average scores on the CAT instrument. Altogether, students showed a median improvement of 13% in critical thinking skills over the course of a term. This result stands in contrast to the widely publicized results from *Academically Adrift* (Arum & Roksa, 20011 which primarily relied on the CLA to reach the conclusion that students are typically making little academic progress during their first two years at college. The increases in critical thinking skills documented here for a single course over a single term (3.5 or 15 weeks, respectively) demonstrate a very positive level of improvement in those skills.

These data, in Figure 2 as standard box and whisker plots, are disaggregated into their separate combinations, which qualitatively show critical thinking improvements (or not) in each comparison for all class types and at both institutions.

Figure 2. CAT Overall Raw Scores for Beginning Versus End of Term for All Students by Class Type and Institution.



We were unable to reasonably control for confounding factors of students being enrolled in previous or concurrent classes of these types at UCB. The CC protocol of one course per term precluded this potentially confounding effect. We also examined, at UCB, possible confounding effects of academic preparation levels as indicated by SAT scores, ACT scores, and number of credit hours. The consistency of these measures across all three categories, with or without these confounding factor adjustments, indicate they do not significantly bias our primary conclusions.

There are three important items of context to keep in mind when examining Figure 2, above: (1) the entering level of academic preparation for students as indicated by SAT/ACT scores, at CC (median composite scores SAT = 2010, ACT = 31), substantially exceed those of students at UCB (median composite scores SAT = 1180, ACT = 26), (2) these data indicate potentially much more room for improvement at UCB than at CC, and (3) CC terms are 3.5 weeks; UCB terms are 15 weeks.

With those important caveats in mind, we see evidence that CC students generally began at a higher level of critical thinking skills, compared to UCB students, but still showed significant improvement for all class types. UCB students began at lower levels of critical thinking skills relative to CC students, but showed comparatively more improvement over the course of the term for all class types, particularly for Control classes.

Students in all three types of class designations made demonstrable progress in critical thinking skills over the course of a single term while those enrolled in Civic Engagement or Critical Thinking courses typically showed greater improvement as measured by changes in median scores than those students enrolled in the Control classes

We follow-up these qualitative results with formal analyses.

Table 1. Means and Cell sizes for Two Factor Design

Type of Class	Colorado College	Univ. of Colorado Boulder
Critical Thinking	25.93; SE=0.885; n=70	23.69; SE=0.674; n=84
Civic Engagement	25.50; SE=0.499; n=112	17.88; SE=1.38; n=16
Control	24.74; SE 1.010; n=31	19.19; SE=0.714; n=62

Because of dramatic imbalance in the cell sizes, we have elected to employ the so-called "Type III" sums of squares approach for the initial two factor design on Total Score. We recognize that the literature remains somewhat unsettled regarding how 'best' to analyze unbalanced factorial designs. We chose to follow the common strategy of using the so-called "Type III" formal testing and the unweighted cell means above (De Rosario Martinez, 2015) and (<https://www.r-bloggers.com/r-tutorial-series-two-way-anova-with-unequal-sample-sizes/>).

Table 2. ANOVA Table for Type III Tests

	Sum of Sq	Df	F-value	Pr > F
Intercept	18977.1	1	620.8	.000***
Class Type	30.5	2	0.5	0.6
University	636.2	1	20.8	.000***
Class Type x University	348.8	2	5.7	0.004**
Residuals	11280.0	369		

This analysis demonstrates highly significant interaction between class type and university and so the main effects p values are not reliable. Given our primary focus on the potential difference in critical thinking enhancement among class types, not on the differences between CC and UCB students, we chose to treat the analysis of total scores in each of the three class types separately for each institution via simple one-way anova models.

First, we show the CAT score analysis for students at UCB for all three class types:

Table 3. ANOVA Table for UCB By Class Type

	Df	Sum of Sq	Mean Sq	F-value	Pr > F
Class Type	2	941.3	470.7	13.5	.000***
Residuals	159	5559.4	35.0		

This analysis shows strong differences in mean performance at UCB for at least one pair of the three class types. We followed this analysis with the conservative "Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test Adjusted for Unequal Variances and Unequal Sample Sizes" to formally test, via *a posteriori* confidence intervals, all three pairwise comparisons, without assuming equal cell sizes or homoscedasticity. (<https://cran.r-project.org/web/packages/DTK/DTK.pdf>)

Table 4. UCB Total Scores, Dunnett-Tukey-Kramer Pairwise Comparison Confidence Intervals

Class Type	Diff	Lower CI	Upper CI
Civic vs Control	-1.32	-5.29	2.65
Critical vs Control	4.50	1.99	7.01
Critical vs Civic	5.82	2.14	9.49

These conservative 95% confidence intervals identify significantly higher scores for the UCB critical thinking classes than Control or Civic Engagement classes. No significant differences were detected between Control and the Civic Engagement classes as the interval includes zero.

We employ the same technique for CC:

Table 5. ANOVA Table for CC By Class Type

	Df	Sum of Sq	Mean Sq	F-value	Pr > F
Class Type	2	30.5	15.2	0.6	0.6
Residuals	210	5720.6	27.2		

We found no detectable differences among class types at CC for Total Score on the CAT test. Thus, no further testing was needed.

Beginning vs Ending Scores

We move to focus on changes in student scores over a single term as the variable of interest. We use the Total Score at the end of term minus the Total Score at the beginning of the term to quantify changes (=Delta). As before, we have an unbalanced, two factor, pure Model I design so we employ the same strategy as described above.

Table 6. ANOVA Table for Type III Tests

	Sum of Sq	Df	F-value	Pr > F
Intercept	44.3	1	2.2	0.1
University	0.1	1	0.0	0.1
Class Type	0.3	2	0.0	0.1
University x Class Type	6.8	2	0.2	0.9
Residuals	3479.5	170		

We see no evidence of any differences in mean change between class types or between universities, nor do we see interaction. Whatever changes may have taken place, none showed up as differences, in aggregate, for either of the two experimental factors: institution and type of class. Note that this result does not mean changes did not occur, as we show in the next section.

A main question of this project aimed to determine if students improved their critical thinking skills significantly over the course of the two different types of terms. We tested the difference between students' CAT scores at the end of the term and those students' CAT scores at the beginning of the term. The null hypothesis here is that there would be no change over the course of the term.

We used a one-sample t-test with the null hypothesis of Delta = 0 for each class at each institution separately. The Delta variable showed no marked graphical deviation from a normal distribution.

Table 7. One Sample t-test for Beginning vs End of Term by Institution and Class Type

	Control	Civic	Critical Thinking
Colorado College	1.9	1.7***	1.8***
University of Colorado	1.9*	3.0	2.1*

*p<0.05, ***, p<0.001

All combinations showed positive change from beginning to end of the term with four of the six cells showing statistically significant improvement in their critical thinking scores. These students showed demonstrable improvement in critical thinking skills as measured by CAT over the course of a single term which, in the case of Colorado College, means improvement over a single course studied for only 3.5 weeks.

Analysis of Student Performance on Individual CAT Questions

In this section we shift the focus to our analyses of student performance on several individual questions of the Critical-thinking Assessment Test. Each question of the CAT probes a particular aspect of critical thinking. We first present an overview of the changes in student performance for all students on individual questions. Then we present an analysis of student performance on individual questions by class type from the beginning to the end of the term.

Table 8. Change in Mean Score as Percentage of Maximum Possible Points From Beginning to End of Term on Each Question for All Students. (Rounded to integers)

	Question Number														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
% change	18	22	6	4	1	-1	-8	13	3	11	16	5	8	18	7

Table 8 demonstrates improved critical thinking change in 13 of the 15 questions for all students that completed the Critical-thinking Assessment Test at CC and UCB.

Figure 3. Comparison of Mean Scores as Percentages of Maximum Possible Points per Question by Class Type for All Students.

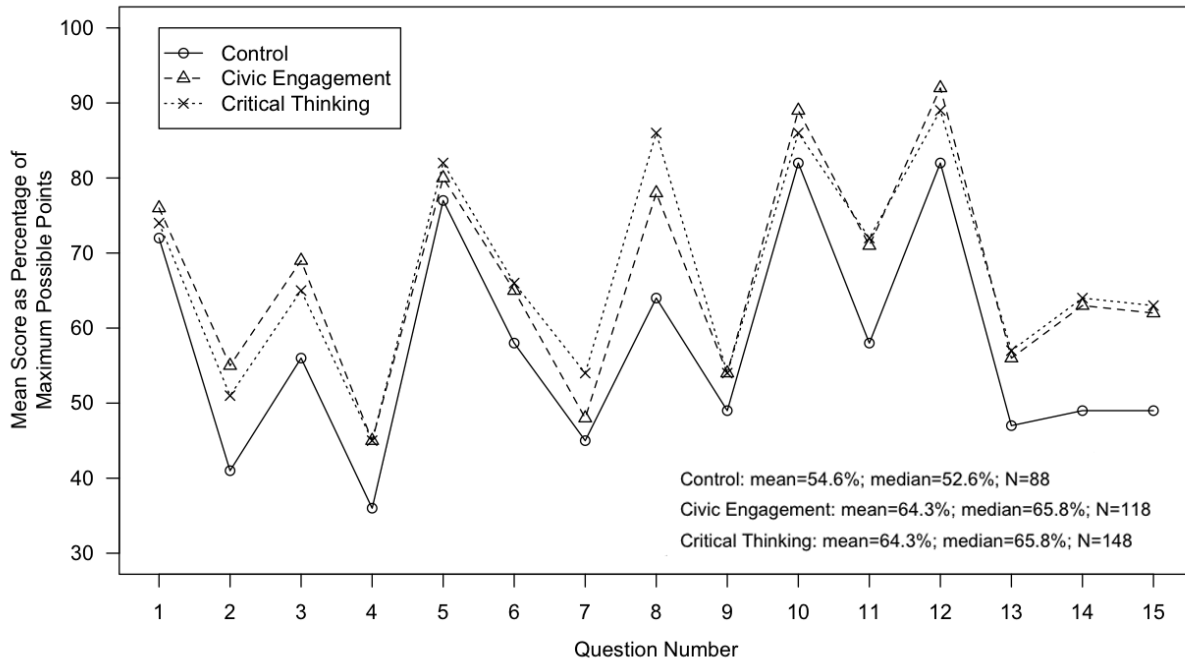


Figure 3 emphasizes the high degree of similarity in performance on each question across class types as well as the modestly but consistently lower scores for students in Control classes as compared to the Critical Thinking and Civic Engagement classes. The data used here were not separated by beginning and end of the term as the question of interest was comparison of the class types; these data demonstrate the overall pattern of similarity in performance across the three class types.

Table 9. Averaged Mean Percentage Changes for Each Individual Question as Percentage of Maximum Possible Points by Class Type from Beginning to End of Term.

Measure	Class Type		
	Control (N=44)	Civic Engagement (N=59)	Critical Thinking (N=74)
Mean of Individual Question Means % Change	+9.4	+9.4	+7.3
Median of Individual Question	+8.3	+8.9	+7.8

Table 9 presents the mean and median values for the percentage changes for all questions over the course of the term. The mean of each question was calculated individually at the beginning and end of the term. Those values were then used to calculate the percentage change in score for each question over the course of the term. Those percentage changes on all individual questions were then all averaged to produce row 1. Row 2 presents the median of the average differences of percentage change for all questions. We emphasize that there was overall improvement across all three class types as demonstrated by the increase in averaged mean percent changes of individual questions for each class type.

The overall performance increase, as measured by percentage change across all class types, is approximately 9%. Students in the Control classes (+9.4%), as well as in those classes designated by instructors as Civic Engagement (+8.9%) showed slightly more improvement than those classes that were designated as focused on Critical Thinking (+7.8%).

Greatest Increase or Decrease in Percentage Change for Specific Questions

The greatest percentage change increase on particular questions over the course of the term occurred on questions 1 (+18.3%), 2 (+22.4%), 8 (+13.2%), 11 (+15.7%), and 14 (+18.5%), all positive.

Question 1 asks students to summarize the pattern of results in a graph without making inappropriate references. Question 2 focuses on students' ability to evaluate how strongly correlational-type data supports an hypothesis. Question 8 asks students to determine whether an invited inference in an advertisement is supported by the information presented. Question 11 requires students to analyze and integrate information from separate sources to solve a real-world problem. Question 14 asks students to identify and explain the best solution for a real-world problem using relevant information.

Questions 1, 2, and 8 also focus on how information supports a particular inference or hypothesis. Questions 11 and 14 focus on developing and communicating solutions to real-world problems. We note there were valuable improvements both in theoretical example scenarios as well as in real-world scenarios.

As a single group, students performed most poorly on questions 6 (-1.2%) and 7 (-8.1%) at the end of the term in comparison to the beginning. Question 6 is designed to elicit from students alternative explanations for spurious relationships. Question 7 asks students to identify additional information necessary to evaluate an hypothesis or interpretation. These results pointedly demonstrate that the students' skills in these areas could substantially benefit from explicit faculty attention.

DISCUSSION

Our results show that students at both the University of Colorado Boulder and Colorado College significantly improved their critical thinking skills—generally agreed upon as the most important skill for undergraduates by university faculty—over the course of a single term whether that term was 3.5 or 15 weeks long. The enhancement effect was demonstrably stronger for students enrolled in courses that emphasized critical thinking *per se* or employed civic engagement/service learning pedagogical strategies when compared to students enrolled in courses (Controls) where neither of these was identified as a major focus. However, we also point out that students in the Control courses also improved their critical thinking skills during the same time periods. Clearly, these students were, indeed, gaining significant educational ground during these terms on this key dimension of goals for undergraduates.

We found a rather surprising degree of consistency in the data between these two institutions that utilize different term schedules and exhibit different student academic profiles. Consistency showed in the general patterns of improvement and was most visible in the patterns of scores on individual questions. The jagged line plots of student scores over the entire question set were surprisingly similar, indicating that the CAT instrument and its scoring are broadly applicable to students at two dramatically different types of higher education institutions. At the same time, we did see some distinctive differences between them.

We suggest that one important reason why students from CC generally demonstrated higher CAT scores at the beginning of each term as well as higher CAT scores at the end of each term in comparison to students at UCB, relates to the different academic preparation selectivity levels for these institutions: CC's median SAT and ACT scores are typically about 2010 and 31 whereas median SAT and ACT scores at UCB are typically about 1200 and 25, respectively. However, we saw no evidence of a 'ceiling effect' (Whang, Zhang, McArdle, & Salthouse, 2009) in any of our data for either institution.

Students at UCB generally showed a greater enhancement and gain (added value) in critical thinking CAT scores over their single term experiences than did students at CC. We again note that UCB terms span 15 weeks and UCB students began with CAT scores lower than those of students at CC. The gains in the one course, 3.5-week term at CC probably represents results with the fewest possible confounding factors to provide compelling results. Students at both institutions made demonstrable, significant progress in enhancing and developing their critical thinking skills while in these undergraduate programs.

Our comparison of three different class types supports the increasingly widely held view that interactive and focused course goals (e.g. practical context problem solving, critical thinking) can be quite effective. We interpret these results to mean that different pedagogical goals and strategies of those class types produced measureable differences in students' enhancement of their critical thinking skills. Given that critical thinking stands as arguably the single most important undergraduate skill to be developed in baccalaureate programs, faculty would be well-advised to explicitly incorporate these strategies into their own course work. Several possible strategies can be found: (Kivunja, 2015; Mayweg-Paus, et al., 2016; Whiley, et al., (2017), Bensley & Murtaugh, (2012). Students in the control group classes showed improvement but typically less than students in either of the other two types of classes, pointing to the potential efficacy of intentional critical thinking instruction by teaching faculty. Assuming, as many faculty do, that the nature of their discipline and ordinary course structure will automatically require students to improve their critical thinking skills is likely an incorrect assumption.

The variation in improvement across class types on questions 2, 8, and 14, together with the lack of variation in improvement across class types on questions 3, 10, and 12, strongly suggest that class type, and presumably current pedagogical emphasis, rank as important in terms of improving the particular aspects of critical thinking addressed by these particular questions.

Together, the overall trends on the individual questions suggest an interesting dichotomy. On one hand, students improved notably in being able to sort through information provided, decide relevancy, and apply that information to devise a solution to a real-world problem. In contrast, students' performances declined when they were not presented with specific options to choose from, or when they were required to imagine what other information might be needed without constraint. We see this pattern as indicating a need for more intentional pedagogical focus on giving students insight into how to think about solving problems that are unconstrained by the information at hand.

Faculty participation and critiques of the students' CAT work showed consistent, if yet only anecdotal, evidence that collectively and simultaneously scoring the CAT test generated important faculty self-reflection of course design and pedagogical methods. In particular, every faculty participant (n>50) who expressed a view on this topic said they expected to make significant changes in their own courses with particular emphasis on identifying and studying critical thinking as a specific, named, and intentional course topic, regardless of faculty discipline. The scoring process where faculty examined dozens of student responses to the same well-crafted instrument highlighted weaknesses and flaws in students' critical thinking skills in a highly visible manner. Importantly, faculty articulated a change in view from the common 'my discipline requires critical thinking automatically' to the intentional 'my students need formal guidance and assistance as they learn to be critical thinkers in my discipline'. We assert this is one of the most

important results of this study of critical thinking skills in undergraduate students at these two institutions. This strategy, and the CAT instrument, can likely be applied effectively across a wide range of institutions of higher learning. We encourage use of the CAT instrument.

ACKNOWLEDGEMENTS

This work was funded by The Teagle Foundation, 570 Lexington Avenue, 38th Floor New York, NY 10022, US, 212-373-1972 and benefitted substantially from the generous advice and counsel of Barry Stein, Ada Haynes, M. Shane Grant, Alex Cruz, and Janis Antonovics.

REFERENCES

- Abrami, P., Bernard, R., Borokhovski, E., Wade, A., Surkes, M., Tamim, R., & Dai Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*, 1102–34.
- Abrami, P., Bernard, R., Borokhovski, E., Waddington, D., Wade, C., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research, 85*, 274–314.
- Arum, R. & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: The University of Chicago Press.
- Bailin, S. & Siegel, H. (2003). Critical thinking. In N. Blake, P. Smeyers, R. Smith, & P. Standish (Eds.), *The Blackwell Guide to the Philosophy of Education* (pp. 181–193). Oxford, UK: Blackwell.
- Bensley, D., & Murtagh, M. (2012). Guidelines for a scientific approach to critical thinking assessment. *Teaching of Psychology, 39*, 5–16.
- Bok, D. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton, NJ: Princeton University Press.
- De Rosario Martinez, H. (2015). *Analysing interactions in fitted models*. https://www.researchgate.net/publication/265067919_Analysing_interactions_of_fitted_models
- Eagan, M., Stolzenberg, E., Berdan Lozano, J., Aragon, M., Suchard, M., & Hurtado, S. (2014). *Undergraduate teaching faculty: The 2013–2014 HERI faculty survey*. Los Angeles: Higher Education Research Institute, UCLA.
- Ennis, R. (2008). Nationwide testing of critical thinking for higher education. *Teaching Philosophy, 31*, 1–26.
- Franco, A. (2016). What do *Ode to Joy*, the Nobel Peace Prize, umbrellas, and cartoons have in common? Why critical thinking matters and how higher education moulds. *Higher Education for the Future, 3*, 108–24.
- Frisch, J., Jackson, P., & Murray, M. (2013). WikiED: Using Web 2.0 tools to teach content and critical thinking. *Journal of College Science and Teaching, 43*, 70–80.
- Gasper, B., Minchella, D., Weaver, G., Csonka, L., & Gardner, S. (2012). Adapting to osmotic stress and the process of science. *Science, 335*(6076), 1590–1.
- Gasper, B., & Gardner, S. (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, 14*, 25–34.
- Gottesman, A., & Hoskins, S. (2013). CREATE Cornerstone: Introduction to scientific thinking, a new course for STEM-interested freshmen, demystifies scientific thinking through analysis of scientific literature. *Life Sciences Education, 12*, 1.
- Hatcher, D. (2011). Which test? Whose scores? Comparing standardized critical thinking tests. *New Directions for Institutional Research, 149*, 29–39.
- Kivunja, K. (2015). Exploring the pedagogical meaning and implications of the 4Cs “Super Skills” for the 21st Century through Bruner’s SE Lenses of Knowledge Construction to Improve Pedagogies of the New Learning Paradigm. *Creative Education, 6*, 224–239.
- Lindholm, J., Szelenyi, K., Hurtado, S., & Korn, W. (2005). *The American college teacher: National norms for the 2004–2005 faculty survey*. Higher Education Research Institute, UCLA.
- Mayweg-Paus, E., Thiebach, M., & Jucks, R. (2016). Let me critically question this! Insights from a training study on the role of questioning on argumentative discourse. *International Journal of Educational Research 79*, 195–210.
- Moon, J. (2008). *Critical thinking: An analysis of theory and practice*. Routledge, N.Y.
- Mulnix, J. (2012). Thinking critically about critical thinking. *Education Philosophy and Theory, 44*, 464–79.
- Pascarella, E. & Terenzini, P. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.
- _____. (2005). *How college affects students: A third decade of research*. San Francisco: Jossey-Bass.
- Possin, K. (2008). A field guide to critical-thinking assessment. *Teaching Philosophy, 31*, 201–28.
- _____. (2013). A serious flaw in the collegiate learning assessment [CLA] Test. *Informal Logic, 33*, 390–405.
- Rowe, M., Gillespie, B., Harris, K., Koether, S., Shannon, L., & Rose, L. (2015) *CBE–Life Sciences Education, 14*, 1–12.

- Shavelson, R. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford, CA: Stanford University Press.
- Stein, B., Haynes, A., & Redding, M. (2006). *Proceedings of the National STEM Assessment Conference: Project CAT: Assessing critical thinking skills*. <http://www.tntech.edu/files/cat/reports>
- Stein, B., Haynes, A., Redding, M., Ennis, T., & Cecil, M. (2007). Assessing critical thinking in STEM and beyond (pp. 79-82). In M. Iskander (Ed.), *Innovations in E-learning, Instruction Technology, Assessment, and Engineering Education*. Dordrecht, The Netherlands: Springer.
- Stein, B., Haynes, A., Redding, M., Harris, K., Tylka, M., & Lisic, E. (2010). *Proceedings of the 2009 International Joint Conferences on Computer, Information, and System Sciences, and Engineering: Faculty driven assessment of critical thinking: National dissemination of the CAT instrument*. <http://www2.tntech.edu/cat/presentations>
- Stein, B., & Haynes, A. (2011). Engaging faculty in the assessment and improvement of students' critical thinking using the critical thinking assessment test. *Change: The Magazine of Higher Learning, 43*, 44-49.
- Whang, L., Zhang, Z., McArdle, J., & Salthouse, T.A. (2009). Investigating ceiling effects in longitudinal analysis. *Multivariate Behavioral Research, 43*, 476-96.
- Whiley, D., Witt, B., Colvin R., Arrue, R., & Kotir, J. (2017). Enhancing critical thinking skills in first-year environmental management students: A tale of curriculum design, application, and reflection. *Journal of Geography in Higher Education, 41*, 166-81.