Journal of the Scholarship of Teaching and Learning, Vol. 22, No. 3, September 2022, pp.34-54.
doi: 10.14434/josotl.v22i3.31801

Intending to Teach Critical Thinking: A Study of the Learning Impacts over One Semester of Embedded Critical Thinking Learning Objects

Tom Lilly[1], Pratima Darr[2], Matthew Schmolesky[1], Todd Lindley[1], Patrick Ludolph[1], Marieke Schilpzand[1], Young Shim[1], Rebecca Higgins[1], Aurelie Weinstein[1], Lior M. Burko[3], and Daniel von Deutsch[1]

[1]Georgia Gwinnett College
[2]Independent Scholar
[3]Theiss Research

Abstract: Numerous studies from recent years and going back decades suggest that post-secondary students are failing to sufficiently improve their critical thinking (CT) skills during their undergraduate years (Abrami et al., 2015; Arum & Roksa, 2011; Huber & Kuncel, 2016). Meanwhile, institutions have increasingly embraced CT as a core competency and educational outcome. Several studies have demonstrated measurable within-semester increases in CT, but most often without a meaningful control group for comparison (Cargas et al. 2017; Grant & Smith 2018; Styers et al. 2018). This study asks if an intervention of embedding content-driven critical thinking exercises within courses would cause a measurable impact on critical thinking outcomes within one semester. All participating courses were paired with an instructor teaching a control section alongside an experimental section. All sections were exposed to pre- and post-assessments, using the Critical Thinking Assessment test. Pre-post results indicated statistically significant gains for experimental groups compared with control groups.

Keywords: critical thinking, higher education, pedagogy

Educators, policy makers, politicians, and employers have increasingly emphasized the importance of critical thinking (CT) as an essential skill (Facione et al., 1995; Huber & Kuncel, 2016). The origin of the term can be traced to the 1910s and philosopher John Dewey's notion of 'reflective thinking,' defined as "active, persistent, and careful consideration of any belief or supposed form of knowledge in light of the grounds that support it, and the further conclusions to which it tends" (Dewey, 1933, p.9). Colleges and universities routinely include the specific term as a core initiative and a stated learning objective, and/or integrated educational outcome. Georgia Gwinnett College (GGC), the home institution of the study, likewise posits critical thinking as one of its core institutional learning outcomes. More specific and complex meanings of the term have been covered in journals of education, philosophy, psychology, and the social sciences (e.g., Behar-Horenstein & Niu, 2011; Moore, 2013; Dwyer et al., 2014).

Critical thinking scholarship in the 1990s often centered around pedagogy, methodology, and teaching strategies (Angelo, 1995; King, 1995; McDade, 1995; Robertson & Rane-Szostak, 1996). Definitions of CT, strategies for improvement, and pedagogical best practices have been challenged by inconsistency in terminology and by prevailing faith-based attitudes that CT would happen organically (Atkinson, 1997; Fox, 1994; Moore, 2013). The 2000s offered a clarification of terms (Moore, 2013), along with a battery of CT instruments, providing somewhat more reliable methods and consistent measurement tools for scholars, institutions, and disciplines. A list of recognized assessment measures includes eight standardized tests for CT (Hitchcock, 2018): the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser 1994), the Ennis-Weir Critical Thinking Essay Test

(Ennis & Weir, 1985), the California Critical Thinking Skills Test (Facione et al., 1995), the Halpern Critical Thinking Assessment (Halpern, 2013), the Collegiate Learning Assessment (Shermis, 2008), and the Critical Thinking Assessment Test (CAT) (Center for Assessment & Improvement of Learning, 2017).

Scholarship on the post-2000s era calls for more action in increasing students CT skills and mindsets (Shim & Walczak, 2012) – some calling for a paradigm shift to fulfill the 'moral purpose of education' (Kivunja, 2014), others searching for the best predictors of high CT skills (Perry et al.,2014), and many offering pedagogical approaches to improve student CT aptitudes (Bean, 2011; King, 1995; Rowe et al., 2015). Prevailing empirical findings strongly suggest that classroom intervention is required to achieve gains in CT among students (Stein & Haynes 2011; Stieha et al., 2016; Grant & Smith, 2018). However, inconsistency in approaches, assumptions, and methods have yielded mixed results, failing to provide clarity on the question of classroom and teaching impact upon CT outcomes (Hathcoat et al., 2016; Liu et al., 2014). For example, little consistency emerged from the empirical research as to specific instructional techniques that effectively enhance students' abilities to think critically (McMillan, 1987; Tsui, 2000, 2002; Shaarawy, 2014). Behar-Horenstein and Niu (2011) have argued that results of interventions vary with implementation and sample size (Niu et al., 2013). Others argue that critical thinking cannot really be 'taught' at all (Willingham, 2008). One of the most serious critiques of CT efforts is a report that 45% of college students in a large-scale study demonstrated no significant gains in the development of critical thinking, complex reasoning, and writing during their first three semesters (Roksa & Arum, 2011).

More recent studies offer concrete empirical evidence to support the hypothesis that teaching and focused effort does, indeed, increase student critical thinking skills as measured by normed, nationally recognized CT tests (Stein & Haynes, 2011; Carson, 2015; Evans et al., 2018). Success in increasing student CT skills has been demonstrated across multiple kinds of tests/instruments. For example, 'overt teaching of critical thinking' led to significantly higher improvements in students (n = 50) than that of 'inquiry-based' methods in biotechnology courses as measured by the University of Florida Critical Thinking Skills test (Friedel et al., 2008). Utilizing a battery of their own pre/post-tests, researchers in the Netherlands demonstrated significant improvements in the performance of students (n = 141) in economics classes, noting "the combination of instruction with practice is crucial and has a large effect on both the immediate and the delayed post-test improvements" (Heijltjes et al. 2014, p.38). Another study at a Belgian university, using yet another metric, found that first-year students (n = 752) across disciplines improve in CT during their first year; however, the increase varied by previous educational background, and the study found that those on a professional track saw greater gains than others (Evens et al., 2013).

Research on semester-scale interventions provides even greater reason for optimism, but several facets of previous experimental models posed challenges to the interpretation of their results due to potential confounds. Styers et al. (2018) noted increased CAT scores resulting from targeted active-learning practices in a "Foundations of Science" course but compared students across different courses and did not maintain separate control and experimental groups. Similarly, Cargas et al. (2017) found marginal and inconsistent gains across the multiple courses in their study and had no control population against which to compare results. Using the CAT, Grant and Smith (2018) found significant improvements across two institutions in multiple courses with learning outcomes and designed interventions intended to foster critical thinking. However, control groups in that study were in different courses that were selected specifically because learning outcomes, instructors, and course activities explicitly did not stress critical thinking, preventing a valid comparison between the two groups.

Study Questions and Study Design

The present study pursued the following research question:

> Would the classroom deployment of intentionally designed critical thinking learning activities result in observable gains in the quality of students' CT skills over the course of one semester?

For its CT intervention, the study team developed a set of learning experiences that the team calls Critical Thinking Learning Objects (CTLOs), classroom activities intentionally designed to synthesize critical thinking concepts and course-specific content and then integrated into the normal semesterly learning plan of a course. The study model compared different sections of the same course taught by the same instructor in the same semester: one section integrated CTLOs into the routine learning of the course, the other did not. The study team hypothesized that students experiencing these activities in one semester would show significant improvement in their CT competence, when compared with students in the other section of the same course who did not experience these activities.

Four foundational cross-disciplinary considerations informed the study design: adherence to a standard definition of CT; administration of a uniform, reliable CT instrument; consistent implementation of its comparative study model to reduce variability between non-CTLO (control) and CTLO (experimental) groups; and use of a uniform template for development of course-specific CTLOs. Regarding the first consideration of a standard operational definition of CT, the present study defined CT consistent with that of the Center for Assessment and Improvement of Learning (CAIL) at Tennessee Technological University (TTU), the designers of the CAT: CT is the analytical, interpretive, and creative thinking competencies identified across academic disciplines as crucial for skills development in evaluating information, problem solving, and communication (Al-Mazroa, 2017; Center for Assessment and Improvement of Learning, 2017).

To address the second consideration, reliable CT assessment, the study used the CAT. The CAT is a validated instrument for assessing critical thinking across disciplines (Center for Assessment and Improvement of Learning, 2016). Unlike most other CT instruments, the CAT relies entirely upon a set of free response questions rather than Likert scale surveys, an improvement over other tests (Ku, 2009; Haynes et al., 2015). Scores on the 15-question CAT instrument range from 0 to 38; in one multi-institution study the actual student score average was 21.02 (SD = 6.19) with a range of 6.0 to 36.3 (Stein et al., 2007). Its scoring methodology involves extensive college faculty training and participation, minimizing faculty scorer recognition of individual students, randomizing faculty scorer encounters with each test question, and norming faculty scores question-by-question. Each individual answer is independently scored by two separate faculty members and, if those scores match, they are recorded; if they do not match, a third faculty member independently scores the item, and the average of the three scores is recorded. Furthermore, all scored test sheets from institutions are authenticated by CAIL by audit, where the CAIL test designers score a sample of CAT tests already administered and scored at an institution and compare auditors' scoring accuracy with that of the institution. Alongside the reliability and accuracy of CAT results, the faculty-centered approach to administering and scoring the CAT is associated with better faculty engagement with the process and with promoting faculty-led CT interventions (Center for Assessment & Improvement of Learning, 2017; Haynes et al., 2016; Lisic, 2015).

Another reason the study team used the CAT is GGC's longstanding assessment relationship with CAIL from TTU. GGC has administered the CAT for collegewide core curriculum assessment since August 2015. Collegewide assessment results for first-time freshman (FTF), sophomores and juniors completing their General Education curriculum (GEN ED) and seniors about to graduate

(SENIOR), are represented in Table 1. See Appendix 1 for a description of the home institution's methodology for administering the CAT for college-wide assessment.

Table 1. College-wide CAT assessment results, Fall 2015 – Fall 2020.

	N	MEAN	MEDIAN	SD
OVERALL	2127	12.33	12	5.05
FTF	773	11.33	11	4.51
GEN ED	484	12.12	12	4.96
SENIOR	512	13.30	13	5.41

Assessment results for GGC's core curriculum are analyzed by interdisciplinary faculty committees at GGC. Because critical thinking is one of the college's core curriculum outcomes, numerous GGC faculty have received training from CAIL on using the CAT for assessment, scoring the CAT, and enhancing CT learning via CT activities in their classes. The work of the faculty committee responsible for oversight of collegewide assessment of critical thinking led to this study.

The third consideration, study design, turned upon the question of minimizing variability of students' classroom experiences among students in control and experimental groups. To do so, participating faculty taught two sections of the same course in the same semester, designating one as the experimental section, the other as the control section. CAT pretests and posttests were scheduled in both experimental and control sections at the beginning and end of the semester. The experimental section differed in that it used CTLOs, the first following the CAT pretest by 1-2 weeks, the second preceding the CAT posttest by 2-4 weeks. The team then aggregated data and compared pre-post performance of the experimental and control groups to ascertain any significant differences. The disciplinary range of the current study is represented in Table 2, which also indicates the number of paired sections that participated by semester.

Table 2. Participating Courses by Semester.

Semester	**Course #**	**Course Name**	**Division**	**# Paired Sections**
Fall 2018	BIOL 1107K	Principles of Biology with Lab	Natural Sciences	3
	PSYC 1102	Introduction to Psychology	Social Sciences	1
Spring 2019	PSCI 1101K	Physical Sciences 1 with Lab	Natural Sciences	1
	GEOG 1101	Introduction to Human Geography	Social Sciences	1
Fall 2019	PSYC 1102	Introduction to Psychology	Social Sciences	2
	MGMT 3400	Ethics and Corporate Social Responsibility	Social Sciences	1

Faculty participants for this study were selected either because of their previous experience in assessing critical thinking at GGC or by recommendation of other members of the research team.

Determining the baseline critical thinking potential of the courses of this study was difficult to ascertain. To clarify at least faculty participants' perceptions of their own courses' intentional learning relationship to critical thinking outcomes, faculty participants were asked to rate their courses' relationship to the above three learning outcomes and to list regular classroom activities aligned with

them. The result provided a coherent, qualitative impression of the critical thinking conditions and support in the course without CTLO intervention, in other words, the control condition of the course. Table 3 conveys faculty participants' impression of critical thinking intentionality in the control sections of the courses of this study.

Table 3. Faculty Participant Impressions of Critical Thinking Competency Conditions in Control Sections.

	Evaluate how strongly information supports an idea or interpretation		**Provide alternative interpretations for information or observations that have several interpretations**		**Identify additional information to evaluate alternative interpretations**	
	Degree Outcome 1 taught in control	**Classroom activities aligned with Outcome 1 in control**	**Degree Outcome 2 taught in control**	**Classroom activities aligned with Outcome 2 in control**	**Degree Outcome 3 taught in control**	**Classroom activities aligned with Outcome 3 in control**
BIOL 1107K	Moderate	Graphing data from laboratory experiments followed by group presentations or lab reports	Little	Lab reports require a section on alternative explanations for any unexpected results, often accounted for as 'human error' in laboratory procedure.	Little	Brief class discussion of alternative interpretations of lab data that could be attributed to anything beyond "human error."
PSCI 1101K	A lot	Class discussion of the novel Contact by C. Sagan, evaluating the evolution of understanding based on available evidence.	A lot	Class discussion of possible interpretations based on information available at the time, and how the interpretations change as more information becomes available.	Moderate	Some class discussions involved the suggestions of what kind of information would be needed to support or refute an interpretation.
GEOG 1101	Little	Class discussion and lecture using graphs and maps.	A lot	Map interpretation exercises to provide possible explanations, Landscape interpretation exercises to brainstorm possible alternative explanations, Homework quiz and exam questions to interpret graphs and maps.	Little	Class discussion and lecture.
PSYC 1102	Little	List real life examples of each psychology topic.	Moderate	Clinical case interpretation with alternative options.	Little	Provide additional information about clinical case interpretation.
MGMT 3400	Little	Ethical decision-making simulation, Video assignment write-up.	Little	Ethical decision-making simulation, Video assignment write-up.	None	

The final consideration was to embed some uniformity in the design and development of CTLOs, which would necessarily differ, as elements of any CTLOs are course-specific. One step the study team took was to apply a uniform intention to the design and development of a CTLO. CTLOs were required to align with course content and outcomes while simultaneously designed to intentionally challenge students to develop three critical thinking competencies:

- evaluate how strongly information supports an idea or interpretation,
- provide alternative interpretations for information or observations that have several interpretations, and
- identify additional information to evaluate alternative interpretations.

These outcomes correspond to the design of the CAT and align with the first of two CT "Skillsets" mapped to CAT questions by CAIL (CAIL, "About the CAT"). A second step was to use a uniform set of design tools. For that, the team relied on training and pedagogical templates from CAIL for embedding specific course content and knowledge within an activity aligned with those three learning outcomes. CAIL's training materials guided the team on creating content-specific CT activities and activity-specific CT evaluation rubrics. With further guidance from CAIL trainers and other members of the research team, participating faculty developed the scope, focus, and content of the CTLOs, as well as targeted evaluation rubrics for each CT outcome, by crosswalking their course outcomes, content areas, and semesterly schedule. The study team implemented an additional design standard outside of the CAIL pedagogical toolkit, a reflective review of the activity involving student blind scoring of another student's CTLO, using the rubric the instructor developed to evaluate student CTLO performance. For quality assurance, prior to classroom use of any CTLO, other members of the research team reviewed CTLOs for clarity, adherence to CAIL's design templates, and alignment with the three CT outcomes listed above. Complete, reviewed CTLOs were then permitted for classroom use and stored within a library for future reference and use. See Appendix 2 for a sample CTLO; the rubric has been omitted because it adheres to CAIL's proprietary formatting and content.

Methodology

Once CTLOs were approved for classroom use by the study team, participating faculty deployed them in one of two sections of the same course they were teaching during the same semester. Two CTLOs were used in each experimental section, based on a predicted impact on critical thinking ability and likely variability in the frequency of students' experiencing them due to absence. The study team also recognized that the fact that students were likely to experience different numbers of CTLOs would allow for an additional comparison of the impact of one CTLO versus two CTLOs on end-of-semester CAT performance. Faculty participants determined the order of CTLO deployment, which depended on when in the semesterly schedule the course content of the CTLO was being taught.

The classroom administration of each CTLO in its entirety took place over two consecutive class periods. The first involved the completion of the activity, the second a review of student responses and reflection on the activity, which included student peer evaluation of the CTLO using the rubric developed for that CTLO. On the first day, the CTLO was completed over approximately 30 minutes of the class period. The exercise was introduced briefly over the first five minutes of the class. Students did not receive intentional coaching on what CTLOs are or what constitutes critical thinking so that the CTLO experience alone could be the focus. The response sheets with the students' college ID number as an identifier were collected by the instructor and graded using the pre-established rubrics. In the subsequent class period, students received the response sheet of an anonymous classmate and were then guided through use of the scoring rubrics so they could score a classmate's work, by both assigning a numerical score in pencil and writing in the margins any specific comments. Students were instructed to be as precise as possible and to avoid personal comments. This peer grading process was applied to all questions featured in the CTLO, with the instructor first introducing and explaining the relevant rubric before asking students to grade; the instructor answered any questions students had while the grading took place. This peer grading process as applied spanned

the entire length of the 50-75 minutes class period, to enable students to perform this task without being unduly rushed. Scores from peer graded response sheets were recorded in the same spreadsheet used to record instructor scoring.

In the parallel control section, students performed independent coursework so that they did not receive any additional lecture or any specific instruction that the students in the experimental section would not receive.

Consent was solicited from all students in every section. In all, 345 students consented to participate, 163 in the control sections and 182 in the experimental sections. An incentive of one extra credit point was offered for each class period in which an activity tied to the study took place. The study was approved by the GGC's Institutional Review Board and funded for two years by internal seed grants. As indicated in Table 4, implementation in each experimental section required six class periods throughout the semester (about 7.5 hours in total).

Table 4. Semesterly workflow schedule of CTLO deployment in experimental sections.

Week	Activity	Description
1	Student consent secured	Students reviewed and signed the study's informed consent form
2-3	Pre-administration of CAT test	Paper test over entire 50–75-minute class
4-7	First CTLO experience	Students work on CTLO in 1st class; peer-grading in 2nd class period.
	Grading of CAT pre-tests	Faculty score CAT tests
8-12	Second CTLO experience	Students work on CTLO in 1st class; peer-grading in 2nd class period.
Weeks 13-14	Post-administration of CAT test	Paper test over entire 50–75-minute class
Weeks 15-17	Grading of CAT post-tests tests and compilation of peer-grading outcomes.	Faculty score CAT tests and analyze data from peer grading of CTLO experiences in all courses.

For pre-post testing, the paper version of the CAT test was administered over the duration of one entire class period during the intervals noted in Table 4. Students were introduced to the test using the same script, which instructors read aloud in class before administering pretests. All students began the test at the same time. Instructors recorded the test completion time for each student so that any relationship between test completion time and quality of performance could be explored. Pretests were graded within 3-4 weeks of completion and posttests were graded immediately after finals week each semester. Consistent with CAIL's CAT scoring protocols, CAT scoring was blind and involved faculty from across the college not affiliated with the study. Scorers did not know which tests were from control or experimental groups and could not identify any student whose test they scored. Per test scoring protocols designed and required by CAIL, any tests that had more than two unanswered items were removed from consideration before grading commenced.

Results

Results are presented in Table 5. Overall, 345 students participated in the study, 182 in the experimental sections, 163 in the control sections. For analysis, the study considered only students who took both pre- and posttests because completing both tests provided the most complete impression of critical thinking development through the semester. The study team was aware that the

study design could have permitted further comparisons for experimental subgroups experiencing zero or one CTLO. However, the small number of study participants in the experimental group who either experienced no CTLOs (n = 2) or one CTLO (n = 15), made analysis based on the number of CTLOs moot. Subsequent analysis was conducted only on the students in the control group who completed both pre- and posttests (n = 98) and the number of students in the experimental group who completed both pre- and posttests and who experienced two CTLOs (n = 110).

Table 5. Overall Study Results.

	n	Mean Pretest Score	Mean Posttest Score	Difference Pre-Post	Two-tailed T-test
All	208	13.17	14.88	1.72	0.0000003
Experimental	110	13.13	15.51	2.38	0.0000004
Control	98	13.21	14.18	0.98	0.0411622

Results by course knowledge domain (Natural Sciences, Social Sciences) and by self-reported class rank (freshman, sophomore, junior, and senior), are presented in Tables 6 and 7, respectively. Of the overall study population, 21 students did not self-report class rank. Of students who completed both pre- and posttests, two students in the experimental group and two students in the control group did not self-report class rank.

Table 6. Study Results by Course Knowledge Domain.

	Course Knowledge Domain	*n*	Mean Pretest Score	Mean Posttest Score	Difference Pre-Post	Two-tailed T-test
Control						
	Natural Science (Biology, Physical Science)	22	16.32	16.91	0.59	0.63058627
	Social Science (Human Geography, Management, Psychology)	76	12.31	13.39	1.09	0.03326717
Experimental						
	Natural Science (Biology, Physical Science)	37	14.66	17.69	3.04	0.00064283
	Social Science (Human Geography, Management, Psychology)	73	12.36	14.4	2.05	0.00020221

Table 7. Study Results by Self-Reported Class Rank.

	Self-reported class rank	*n*	Mean Pretest Score	Mean Posttest Score	Difference Pre-Post	Two-tailed T-test
Control						
	Freshman	15	13.6	13	-0.6	0.6717544
	Sophomore	47	13.34	15.03	1.69	0.0130101
	Junior	23	12.06	13.43	1.38	0.1499009
	Senior	11	15.18	14.24	-0.94	0.5207975
Experimental						
	Freshman	24	13.58	16.49	2.9	0.0081127
	Sophomore	42	13.53	14.98	1.44	0.0229522
	Junior	26	10.73	14.97	4.24	0.0000860
	Senior	16	15.31	16.19	0.88	0.5207975

Which students show CAT score gains?

The proportion and scope of gains pre-to-post are illustrated in Figures 1 and 2.

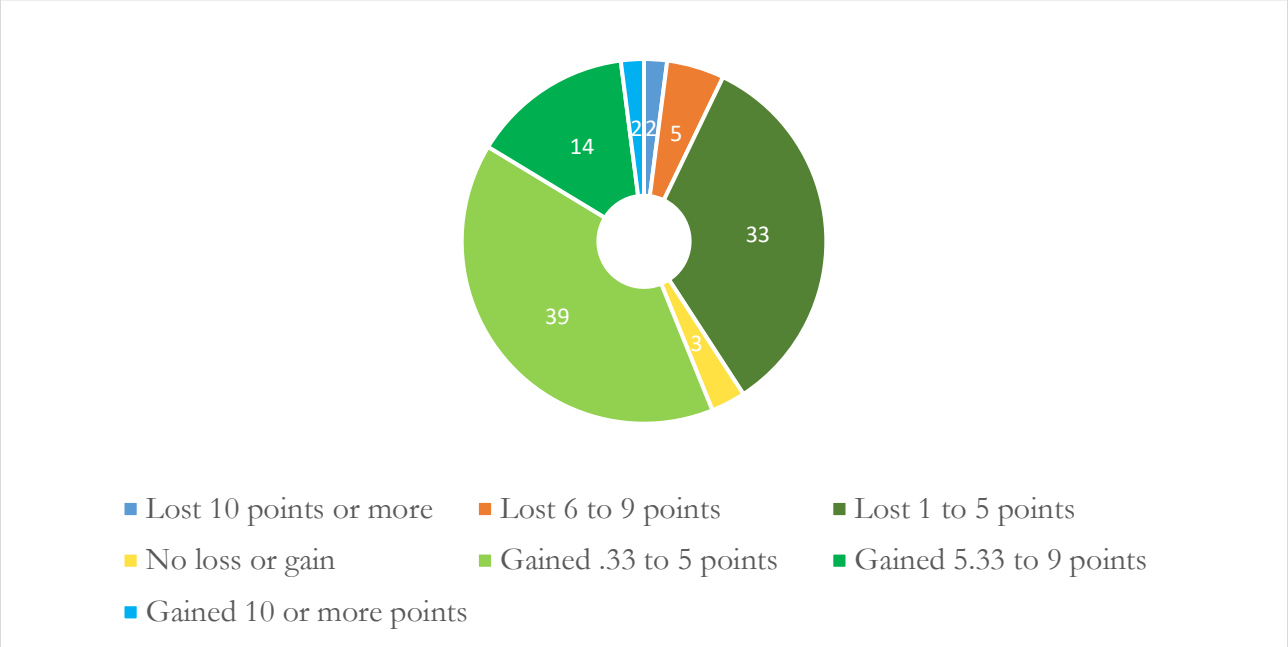

Figure 1. Proportion of pretest-posttest results, control group completing both pre- and posttests.

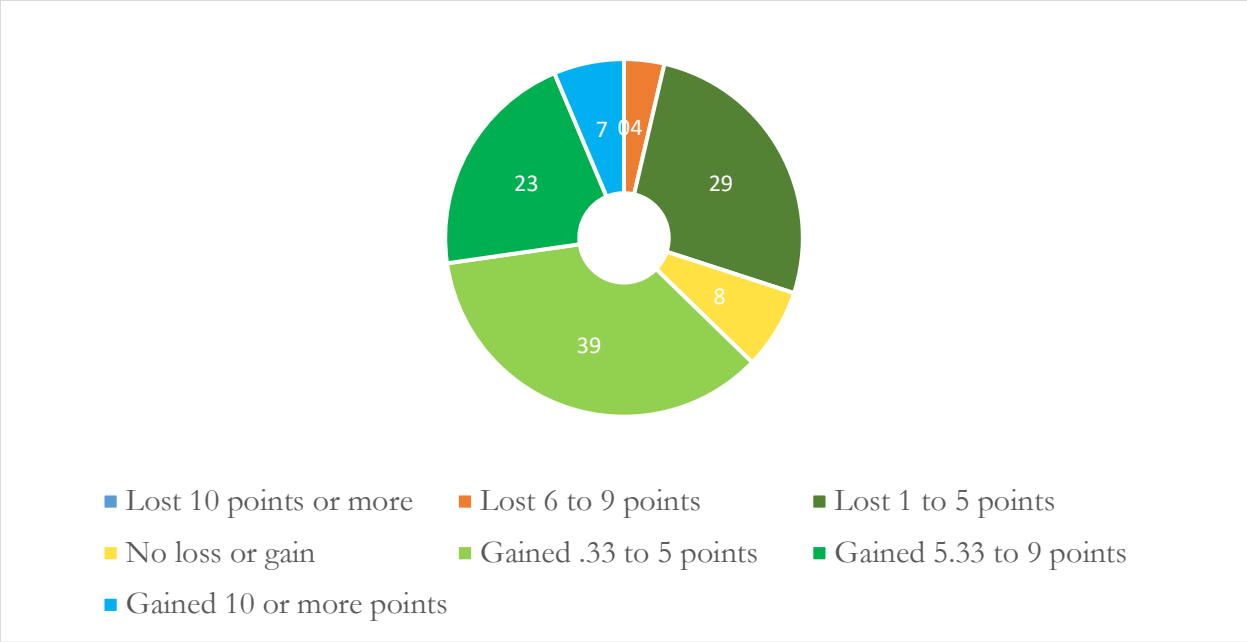

Figure 2. Proportion of pretest-posttest results, experimental group completing both pre- and posttests and both CTLOs.

Of the 110 students in the experimental condition, 33 showed a decrease in CAT score from pre to post (M = -3.10, SD = 2.03, min = -8.0, max = -1.0), 8 showed no change, and 69 showed an improvement (M = 5.28, SD = 2.95, min = 1.0, max = 13.0), demonstrating that ~63% of students experienced a gain. In contrast, of the 98 students in the control condition, 40 showed a decrease in CAT score from pre to post (M = -3.56, SD = 2.59, min = -12.0, max = -1.0), 3 showed no change, and 55 showed an improvement (M = 4.33, SD = 2.76, min = 0.33, max = 12.0), demonstrating that approximately 56% of students experienced a gain.

Do CTLOs significantly impact CAT scores?

A two-way mixed model ANOVA revealed a main effect for Time (pre vs. post) ($F(1, 206) = 26.99$, $p < 0.05$, $\eta^2 = 0.12$), no main effect for Condition ($p > 0.05$, $\eta^2 = 0.004$), and a significant Time by Condition interaction ($F(1, 206) = 4.72$, $p < 0.05$, $\eta^2 = 0.012$). Thus, there was a statistically significant difference comparing posttest results (M = 14.88, SD = 5.56) to pretest results (M = 13.17, SD = 5.46) when collapsing across condition. There was no overall difference comparing the control (M = 13.70, SD = 5.35) and experimental (M = 14.32, SD = 5.73) groups when collapsing across time. While the Time by Condition interaction was a weak effect, on average the exposure to the critical thinking exercises in the experimental group provided more than double the gain in performance on the CAT test (M = 2.38, SD = 4.63) compared to that observed in the control condition (M = 0.98, SD = 4.67) (see Figure 3; error bars represent 95% confidence intervals). Post-hoc independent t-tests demonstrated that difference between the control and experimental conditions for the pre-test was not significant ($p > 0.05$), but was for the post-test ($t(206) = 1.73$, $p = 0.04$, $d = 0.24$). Paired sample tests revealed that the pre vs. post CAT score gain was statistically significant for both the control condition ($t(97) = 2.07$, $p = 0.04$, $d = 0.21$), and the experimental condition ($t(109) = 5.39$, $p<0.01$, $d = 0.51$). The 2.38 point average gain in the experimental condition represents an 18.1% improvement over the baseline pre-test score in that condition while the 0.98 point average gain in the control condition represents a 7.3% improvement over baseline in that condition.

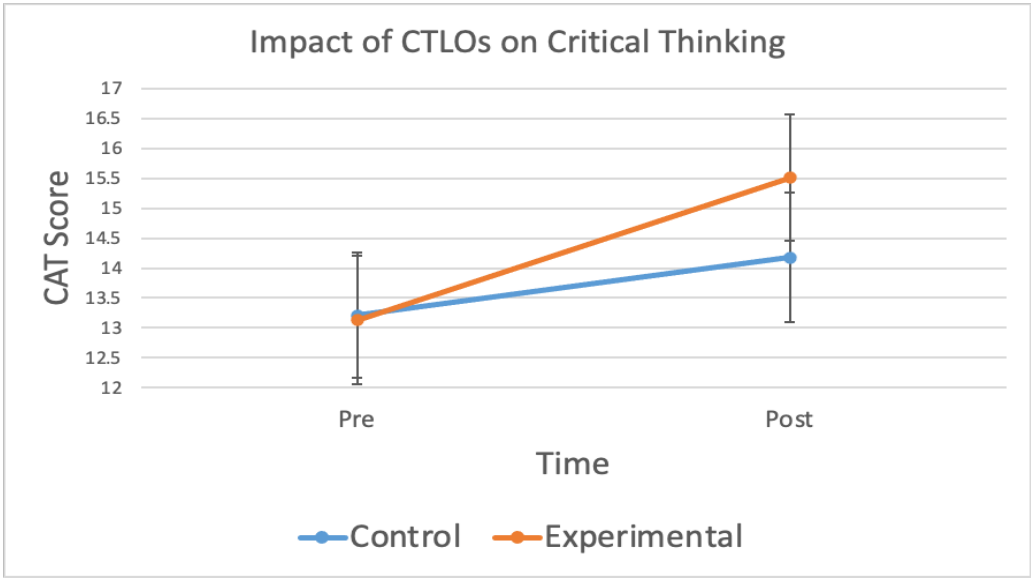

Figure 3. Pre-post CAT gains, control and experimental groups.

We next examined the CTLO scores themselves and their relationship to CAT scores. The maximum score possible for different CTLOs varied from 5 to 10. All CTLO scores were converted to percentages before further analysis was conducted. There was no statistically significant difference between CTLO 1 scores (M = 34.90, SD = 2.42, min = 0.0, max = 90.0) and CTLO 2 scores (M = 32.03, SD = 2.21, min = 0.0, max = 100.0), $p > 0.05$. CTLO 1 was significantly correlated to the pre CAT (r = +.33) and post CAT (r = +.24) but not to CTLO 2 (r = +.12) or to the pre-post CAT difference score (r = -.11). CTLO 2 was not significantly correlated to the pre CAT (r = +.15), post CAT (r = +.08), or pre-post CAT difference score (r = -.09). Thus, students who engaged in both CTLOs showed a significant improvement in CAT scores relative to control (see Figure 3), but actual scores on the CTLOs did not predict improvement in the CAT.

There was no significant difference in the duration of time (in minutes) to complete the pre CAT (M = 43.66, SD = 9.99, min = 23, max = 83) compared to the post CAT (M = 39.34, SD = 10.04, min = 18, max = 64) for the experimental group, $p > 0.05$. The pre CAT time was significantly correlated to the pre CAT score (r = +.27) but not the pre-post CAT difference score (r = -.08). The post CAT time was significantly correlated to the post CAT score (r = +.23) but was also not significantly correlated to the pre-post CAT difference score (r = +.06). It is also worth highlighting that that the control condition pre CAT time (M = 43.93, SD = 9.12, min = 23, max = 66) and post CAT time (M = 38.55, SD = 11.00, min = 20, max = 59) were not statistically significantly different from one another, or from the pre CAT and post CAT times for the experimental condition ($p > 0.05$).

We next examined the impact of student academic rank on CAT scores. As would be expected from the collegewide assessment data, a one-way ANOVA comparing pretest CAT scores by class rank revealed a significant difference ($F(3, 200) = 3.36$, $p < 0.05$). Seniors in the current study scored numerically higher on average ($M = 15.26$, $SD = 4.72$, n = 27) than did freshmen ($M = 13.59$, $SD = 4.55$, n = 39), sophomores ($M = 13.43$, $SD = 5.58$, n = 89), or juniors ($M = 11.35$, $SD = 5.96$, n = 49), though planned pairwise comparisons were only marginally significant for two of these comparisons (seniors vs. freshmen $p = 0.077$; seniors vs. sophomores $p = 0.062$; seniors vs. juniors $p < 0.05$). A one-way ANOVA on pre-post changes scores for the experimental condition also revealed a significant difference by class rank ($F(3, 104) = 2.72$, $p < 0.05$). Seniors in the current study showed less improvement on average ($M = 0.88$, $SD = 5.37$, n = 16) than did juniors ($M = 4.24$, $SD = 4.63$,

n = 26, ($p < 0.05$). While a numeric trend suggests seniors might also have improved less than freshmen ($M = 2.90$, $SD = 4.91$, n = 24) or sophomores ($M = 1.44$, $SD = 3.96$, n = 42), these differences were not statistically significant ($p > 0.05$). A one-way ANOVA on pre-post changes scores by class rank for the control condition was not statistically significant (F (3, 88) = 1.93, $p > 0.05$).

Because seniors demonstrated higher baseline CAT scores compared to freshmen in both our collegewide assessment data (see Table 1) and in the current study, it was important to evaluate the proportion of academic ranks in the control and experimental conditions to ensure equivalency. The percentage of students for each academic rank for each condition were: freshmen (18.4% control vs. 20.6% experimental), sophomore (41.9% vs. 47.1%), junior (23.1% vs. 25.9%), and senior (12.7% vs. 14.3%). A chi-square analysis demonstrated that there was no significant relationship between the study condition (experimental vs. control) and academic rank (X^2 (3, N = 204) = 2.77, p = .43).

Taking a different approach to examine if some individuals benefitted more from the CTLOs than others, a correlational analysis was conducted on the pre-test CAT scores versus the pre-post CAT change scores for participants in the experimental group. While the result (r (108) = -.42, $p <$ 0.05) suggests that those who do worse in the pre-test improve the most, an essentially identical result was obtained in the control condition (r (96) = -0.40, $p < 0.05$) suggesting this effect is due to regression to the mean.

Discussion

The intent of this study aligns with the growing body of research (Evens et al., 2013; Cargas et al. 2017; Grant & Smith, 2018; Grussendorf & Rogol, 2018; Styers et al. 2018), focused on the possibility of observing CT gains over one year. Our results demonstrated that, on average, both control and experimental groups showed statistically significant improvements in CT scores, but that the single semester CT change for those who experienced the CTLOs and the peer-review processes (~18% gain) was essentially double that observed in students who did not (~7% gain). The CT improvement in the control condition could be due in part to organic improvement in critical thinking over a semester of college coursework and here it is worth noting that while no intentional CT training was offered in the control condition course sections, the students in those sections were also enrolled in courses unrelated to this study. Practice effects may have also played a role since the CAT test given at the end of the semester was identical to that students took at the semester start. It is logical to consider that the actual CT gain seen in the experimental condition is superimposed upon the smaller gains seen in the control condition, leaving a true gain of approximately 11%. While the skills students in the experimental group practiced in the CTLOs and peer-review processes overlap with some of those required in the CAT itself, we argue that this is not a limitation but a feature of intentional CT training, and the results indicate that even the limited training provided here at two time-points in the semester, provides statistically significant benefits.

The statistically significant results that answer our central research question, *would the classroom deployment of CTLOs result in observable gains in the quality of students' critical thinking skills*, likewise align with results of Cargas et al. (2017) and Grant and Smith (2018) and as such contribute to the growing optimism that the teaching of critical thinking may not be as elusive as Willingham (2008) and Roksa and Arum (2011) claim. Another similarity between this study and the work of Cargas et al., Grant and Smith, and Styers et al. (2018) is the use of the CAT to measure overall CT. The validity and interdisciplinary/universal design of the CAT recommend it for future studies. The design of this study notably departs from models of previous studies in two ways. As Styers et al. (2018) commented, "future studies comparing sections of the same course with comparable student demographics, taught by the same faculty using more traditional teaching pedagogies would allow for a more objective analysis" (p.10). This study is novel in that it establishes the baseline conditions for such a comparison

in having the same instructor teach two identical classes at the same time, differing only in the integration of CTLOs in one of the classrooms. This allowed the study to draw objective, comparative observations of the impact of a single type of pedagogical intervention, the CTLO and peer-review learning process, given at two time points. This is the first study the authors are aware of that observes a significant impact on CT learning due to a specific learning activity in comparison with matched control sections across diverse disciplines.

Before considering implications, it is worth remarking on a limitation of this study, sample size. The size of the study sample permitted analysis of aggregate results but not of disaggregated results. The study therefore was prevented from considering demographic or other notable variables in student participants' experience that may have impacted learning of CT competencies. For example, results rule out the possibility that disproportionate class rank numbers influenced comparisons of experimental and control sections, but the results also show no more than marginally significant differences between students of different class ranks. More participants are needed to draw those comparisons. Similarly, the nature and content of the knowledge of the classes could correlate with differential impacts, but grouping by broader knowledge domain could minimize the significance of the course-specific design and integration of a CTLO. Furthermore, any broader claims derived from the current data set regarding the CT interdisciplinarity of the results would need to note that no Humanities courses provided data. Course-level analysis is similarly affected. *N*s for any one course were too small or varied to permit generalizable conclusions regarding course-specific performance. To consider those questions would involve greater costs, and more faculty and student participants, than the current study could manage. The execution of this study was time and resource intensive. The research team devoted extensive hours to receiving training on the CAT, developing CTLOs, planning on integrating those activities into course content over multiple classroom days in a semester, reviewing materials from other members of the team, evaluating student results on CTLOs, and scoring CAT tests. In particular, the scoring of CAT tests is extremely time intensive because of the CAT's written-answer design and scoring procedure. Multiple-choice CT instruments would have taken less time to score, but the CAT is uniquely reliable and accurate, as well as applicable to diverse learning experiences and learner populations (CAIL, "About the CAT"). Scaling up this study to explore more fully the demographics of CT learning would require an even more upscaled investment of time, resources, and people, as well as a restructuring of the faculty training model to accommodate a greater diversity of content knowledge expertise and CT awareness.

A second limitation of this study is its understanding of the experience of the CTLO itself. The study was designed to gauge students' CT development primarily in one way, through instructor scoring of each CTLO. Pretest and posttest CAT scores provided reliable information on the CT learning after both CTLOs. It was assumed that scheduling CTLOs in sequence was sufficient to improve critical thinking. However, we observed that individual CTLO performance did not necessarily improve sequentially. Predictably, high instructor-derived scores on CTLOs 1 and 2 were significantly positively correlated with CAT posttest scores. However, students did no better on CTLO 2 than on CTLO 1. There was no significant difference between CTLO 1 scores ($M = 34.90$, $SD = 2.42$, min = 0.0, max = 90%) and CTLO 2 scores ($M = 32.03$, SD 2.21, min = 0.0, max = 100%, $p > .05$). There was no significant correlation between differences from CTLO 1 to CTLO 2 and pre/post CAT differences. Several attitudinal or experiential reasons that affect student persistence could have also affected students' performance on the CTLOs and CAT. The study did not seek student attitudes about the experience. Knowing more about how students perceived these experiences could shed light on why mean CTLO student performance declined as the semester progressed and suggest ways of making CTLO experiences more engaging or effective. Also, worth asking is whether faculty CTLO scoring was, itself, consistent. Faculty scored their own students' CTLOs without participation from other members of the study team. A multiple instructor scoring

model could have assured greater CTLO score consistency across courses and managed potential biases that could have affected any one instructor's evaluation of student CTLO engagement. Questions about CTLO scoring consistency treated an individual CTLO as a standalone problem set. These activities are more than the problems themselves: they are constructed as learning events that take place over multiple class periods, involving case studies analysis, data interpretation, situational role-playing and decision-making, and student peer review, and reflection on the experience.

Despite these limitations, the study team is confident that refining its processes and understanding the nature of these activities will only further benefit students. The prospect of overcoming those limitations in fact gestures towards interesting opportunities to expand the directions and implications of this study model. One opportunity pertains to conducting comparative analysis by demography, and one important demographic variable is learning by discipline area or knowledge domain. Broadly considered, the study was conducted in courses in the Social Sciences (Psychology, Management, Geography) and Natural Sciences (Physical Science, Biology). A comparison of impact by knowledge domain could shed light on whether content learning in some areas affects critical thinking development comparatively more than content learning in other areas does. That comparison would also benefit from development and deployment in Humanities courses; the current study was prevented from gathering information about Humanities courses because coronavirus closed GGC the semester that the study was testing CTLO interventions in its first Humanities course, History 1122, Survey of Western Civilization 2. Such analysis could also qualify the understanding of the CAT as an instrument designed to observe interdisciplinary or universal CT. The results presented in Table 7 are suggestive of this. A broader study with greater student participation could clarify questions in that direction as well as pose new questions about CT learning specific to disciplines and knowledge domains.

Given that comparative analyses would hinge upon scaling up the entire study, it would offer further lines of inquiry besides comparing across populations. Because critical thinking grew not necessarily in sequence but certainly by regular exposure to intentionally wrought CT learning activities integrated into routinized learning, providing more students with more frequent CTLO-based experiences is likely to yield greater effects. The current study found that two CTLOs correlated to CT growth over one semester, even though CTLO performance did not incrementally improve over that semester. This suggests multiple exposure to the experience could have greater impacts, perhaps even retained over a greater length of time. An expanded study could involve incorporating this model into learning communities or pathways of courses where multiple CTLOs could grow and reinforce critical thinking for more students over longer spans of their overall college experience.

A third opportunity for further study pertains to a factor intrinsic, indeed central to the design of the entire study - namely, a close examination of whether components of the CTLO impact critical thinking differently, in what ways, and to what effect. The totality of the CTLO experience was measured but not the various steps or phases of each experience. This warrants consideration, as it could be argued that expressly because students in the experimental group were provided both with exercises modeled on CAT skills and a reflection period including blind scoring of those exercises, they were expressly prepared to succeed on the CAT. A closer examination of the design of the CTLOs in relation to CAT skills and questions suggests that students in the experimental group were not unduly prepared to excel at specific CAT questions just because they practiced *and* reviewed experiences resembling them. CTLOs were based on CAIL templates aligned to Skillset 1 of CAT-aligned CT competencies. CAT questions aligned with Skillset 1 account for 20 points of the overall CAT score of 38. Results on Skillset 1 questions are provided in Table 8.

Table 8. Study Results, Skillset 1.

	N	Mean Pretest Score	Mean Posttest Score	Difference Pre-Post	Two-tailed T-test
All	208	5.64	6.50	0.86	0.0000884
Experimental	110	5.65	6.91	1.25	0.0000512
Control	98	5.63	6.03	0.41	0.1813454

Gains of the experimental cohort on questions aligned with Skillset 1 negligibly differed from gains of the experimental cohort on the entire CAT. As indicated in Table 5, proportional improvement of the experimental group on the entire CAT was 6.26% (2.38/38). The proportional improvement of the experimental group on questions aligned with Skillset 1 and with CAIL's pedagogical templates was slightly less, 6.25% (1.25/20). The proportional difference of the control cohort was slightly larger, 2.58% (.98/38) on the entire CAT compared with 2.05% (.41/20) on questions aligned with Skillset 1. If the classroom reflection/scoring sessions after the CTLOs were merely preparing students to excel at Skillset 1 questions, results would indicate disproportionate improvement on those questions. We observed no such targeted improvement, suggesting that students retained critical thinking abilities not only to answer specific questions built around specific CT skills but also to consider how to answer questions involving different CT skills than what they practiced.

Understanding better the relationship of the features of these learning activities could illuminate further innovative practices with more targeted and impactful CT teaching, integratable in diverse ways within a course or among different courses. The opportunities to expand and innovate are, ultimately, what inspire our continued commitment and optimism. Now that we can observe an impact of our intention to teach critical thinking, we can understand those intentions more clearly to refine our practice for the benefit of our students. We, as teachers, can also learn to intentionally diversify our practices to help more students learn critical thinking in different ways towards, perhaps, even more long-lasting and impactful outcomes.

Acknowledgments

The authors wish to acknowledge the following, without whom this study would not have been possible:

- *The late Dr. Tom Gluick, Assistant Professor of Chemistry, Georgia Gwinnett College.* A member of the research team whose insights gave shape to the scope of the study's design, Tom passed away before the drafting of this manuscript. We miss you, Tom.
- *The Office of the Senior Vice President of Academic and Student Affairs and Provost, Georgia Gwinnett College*. Seed grants from the Office of the Provost funded CAT testing for this study for two years.
- *The Office of Academic Assessment, Georgia Gwinnett College*. The Office of Academic Assessment subsidized some CAT testing costs, led CAT scorer training and college faculty scoring sessions, and organized collegewide CAT/CT trainings offered by CAIL.
- *The Center for Assessment and Improvement of Learning, Tennessee Technological University*. The developers of the CAT, CAIL provided valuable advice and guidance throughout this study

Appendix

Appendix 1. Georgia Gwinnett College's for Collegewide CT Assessment using the CAT.

GGC has used the CAT for college-wide assessment since Fall 2015 semester.

Collegewide assessment captures an annual cross section of critical thinking abilities of representative groups of students by status at the college: first-time freshman (FTF) and first-year students with fewer than 15 earned credit hours; students completing their general education (GE) curriculum (sophomores and juniors with more than 30 and fewer than 75 earned credit hours); and near-graduates (seniors with more than 90 earned credit hours). The test is administered during two different periods each academic year. First-year students are assessed within the first four weeks of each fall semester, and the general education and senior groups are assessed within the last four weeks of each spring semester. Students are selected to participate in college-wide critical thinking assessment based on whether they are enrolled in a course that meets criteria for appropriate level of the curriculum (General Education courses for first-year students and GE groups, upper-division courses for senior group) and if registration indicates a high percentage (greater than 80%) of students who fall into the appropriate credit-hour range. Results are analyzed by faculty committees charged with recommending changes to teaching and curriculum that could result in gains in learning.

Appendix 2. Sample CTLO, Introduction to Human Geography, "US Manufacturing in the 21st Century."

Manufacturing was the main driver of economic growth in the United States for much of the 20th century. The U.S. was the world's leading producer of cars, steel, construction equipment, and thousands of other items for several decades beginning in the 1930s. The country ranked number one in total industrial output every year until 2010, when China became the world's leading manufacturer. Politicians, economic advisers, and media reports have commented on the rise of China and the decline of manufacturing in the U.S. Some American manufacturers relocated operations to Mexico, China, India, and to other corners of the world to save money, which may signal the demise of American manufacturing in the 21st century and beyond. Geographers are interested in understanding the relationship between globalization, economic restructuring, and political decision-making and the effects of such phenomena on unemployment, social development, and general well-being. As such, historical data and graphs are helpful to better understand current, past, and future trends in American manufacturing. Some have argued that American manufacturing has been weakened because of competition from other countries and that output will continue to decline without governmental protection from competitors. Examine the graph below showing the number of employees working in manufacturing in the U.S. from 1940-2015 to answer the following questions. (**Source: US Bureau of Labor Statistics)**

1. How strongly do the data support the idea that the American manufacturing output declined from 1985 to 2015? Please comment.

2. If American manufacturing output did not decline from 1985-2015, what are two (2) other possible interpretations of the data presented in this graph?

3. What other specific piece of information might be helpful to determine whether manufacturing output increased in the U.S. from 1985-2015?

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, *85*(2), 275-314. https://doi.org/10.3102/2F0034654314551063

Al-Mazroa, S. (2017). Assessment of critical thinking skills in undergraduate animal science students and curriculum.

Angelo, T. A. (1995). Classroom Assessment for Critical Thinking. *Teaching of Psychology*, *22*(1), 6-7. https://doi.org/10.1207/s15328023top2201_1

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Atkinson, D. (1997). A critical approach to critical thinking in TESOL. *TESOL quarterly*, *31*(1), 71-94. https://doi.org/10.2307/3587975

Bean, J. C. (2011). *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom*. John Wiley & Sons.

Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching & Learning (TLC)*, *8*(2). https://doi.org/10.19030/tlc.v8i2.3554

Cargas, S., Williams, S., & Rosenberg, M. (2017). An approach to teaching critical thinking across disciplines using performance tasks with a common rubric. *Thinking Skills and Creativity*, *26*, 24-37. https://doi.org/10.1016/j.tsc.2017.05.005

Carson, S. (2015). Targeting critical thinking skills in a first-year undergraduate research course. *Journal of Microbiology & Biology Education*, *16*(2), 148. https://doi.org/10.1128/jmbe.v16i2.935

Center for Assessment & Improvement of Learning, Tennessee Tech University (CAIL). (2017). *About the CAT*. Retrieved from: https://www.tntech.edu/cat/about.php.

Center for Assessment & Improvement of Learning, Tennessee Tech University (CAIL). (2016). *CAT Instrument Technical Information*. Retrieved from: https://www.tntech.edu/cat/pdf/reports/CAT_Technical_Information_V8.pdf

Dewey, J. (1933). *How we think. Lexington, MA*.

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, *12*, 43-52. https://doi.org/10.1016/j.tsc.2013.12.004

Ennis, R. H., & Weir, E. (1985). *The Ennis-Weir Critical Thinking Essay Test. Pacific Grove, CA*.

Evans, C., Kandiko Howson, C., & Forsythe, A. (2018). Making sense of learning gain in higher education. *Higher Education Pedagogies*, *3*(1), 1-45. https:doi.org/10.1080/23752696.2018.1508360

Evens, M., Verburgh, A., & Elen, J. (2013). Critical thinking in college freshmen: The impact of secondary and higher education. *International Journal of Higher Education*, *2*(3), 139-151. https://doi.org/10.5430/ijhe.v2n3p139

Facione, P. A., Sanchez, C. A., Facione, N. C., & Gainen, J. (1995). The disposition toward critical thinking. *The Journal of General Education*, *44*(1), 1-25.

Fox, H. (1994). *Listening to the World: Cultural Issues in Academic Writing*. National Council of Teachers of English, 1111 W. Kenyon Road, Urbana, IL 61801-1096.

Friedel, C., Irani, T., Rudd, R., Gallo, M., Eckhardt, E., & Ricketts, J. (2008). Overtly teaching critical thinking and inquiry-based learning: A comparison of two undergraduate biotechnology classes. *Journal of Agricultural Education*, *49*(1), 72-84. http://dx.doi.org/10.5032/jae.2008.01072

Grant, M., & Smith, M. (2018). Quantifying assessment of undergraduate critical thinking. *Journal of College Teaching & Learning (TLC)*, *15*(1), 27-38. https://doi.org/10.19030/tlc.v15i1.10199

Grussendorf, J., and Rogol, N. (2018). Reflections on Critical Thinking: Lessons on a Quasi-Experimental Study. *Journal of Political Science Education 14(2),* 151-166. https://doi.org/10.1080/15512169.2017.1381613

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449. https://psycnet.apa.org/doi/10.1037/0003-066X.53.4.449

Halpern, D. F. (2013). The Halpern critical thinking assessment: A response to the reviewers. *Inquiry: Critical Thinking Across the Disciplines*, 28(3), 28-39. https://doi.org/10.5840/inquiryct201328317

Hathcoat, J. D., Penn, J. D., Barnes, L. L., & Comer, J. C. (2016). A second dystopia in education: Validity issues in authentic assessment practices. *Research in Higher Education*, *57*(7), 892-912. https://doi.org/10.1007/s11162-016-9407-1

Haynes, A., Lisic, E., Harris, K., Leming, K., Shanks, K., & Stein, B. (2015). Using the Critical Thinking Assessment Test (CAT) as a model for designing within-course assessments: Changing how faculty assess student learning. *Inquiry: Critical Thinking Across the Disciplines*, *30*(3), 38-48. http://dx.doi.org/10.5840/inquiryct201530316

Haynes, A., Lisic, E., Goltz, M., Stein, B., & Harris, K. (2016). Moving beyond assessment to improving students' critical thinking skills: A model for implementing change. *Journal of the Scholarship of Teaching and Learning*, *16*(4), 44-61. https://doi.org/10.14434/josotl.v16i4.19407

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction*, *29*, 31-42.

Hitchcock D. (2018). Assessment: As Supplement to Critical Thinking, The Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/entries/critical-thinking/assessment.html.

Huber, C. R., & Kuncel, N. R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, *86*(2), 431-468. https://doi.org/10.3102%2F0034654315605917

Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. *Quantitative Analysis of Culture Using Millions of Digitized Books*. **Science** (Published online ahead of print: 12/16/2010) (retrieved: https://books.google.com/ngrams/graph?content=%22critical+thinking%22&year_start=1920&year_end=2019&corpus=26&smoothing=3#)

King, A. (1995). Designing the instructional process to enhance critical thinking across the curriculum. *Teaching of Psychology*, *22*(1), 13-17. https://doi.org/10.1207%2Fs15328023top2201_5

Kivunja, C. (2014). Do You Want Your Students to Be Job-Ready with 21st Century Skills? Change Pedagogies: A Pedagogical Paradigm Shift from Vygotskyian Social Constructivism to Critical Thinking, Problem Solving and Siemens' Digital Connectivism. *International Journal of Higher Education*, *3*(3), 81-91. https://doi.org/10.5430/ijhe.v3n3p81

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*(1), 70-76. https://doi.org/10.1016/j.tsc.2009.02.001

Lisic, E. S. (2015). *Creating change: Implementing the critical thinking assessment test (CAT) as faculty development to improve instruction* (Order No. 3722968). Available from ProQuest Central.

(1728162696). https://www.proquest.com/dissertations-theses/creating-change-implementing-critical-thinking/docview/1728162696/se-2?accountid=11244

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, *2014*(1), 1-23.

McDade, S. A. (1995). Case study pedagogy to advance critical thinking. *Teaching of Psychology*, *22*(1), 9-10. https://doi.org/10.1207%2Fs15328023top2201_3

McMillan, J. H. (1987). Enhancing college students' critical thinking: A review of studies. *Research in Higher Education*, *26*(1), 3-29. https://doi.org/10.1007/BF00991931

Moore, T. (2013). Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education*, *38*(4), 506-522. https://doi.org/10.1080/03075079.2011.586995

Niu, L., Behar-Horenstein, L. S., & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational Research Review*, *9*, 114-128. https://doi.org/10.1016/j.edurev.2012.12.002

Perry, D. K., Retallick, M. S., & Paulsen, T. H. (2014). A Critical Thinking Benchmark for a Department of Agricultural Education and Studies. *Journal of Agricultural Education*, *55*(5), 207-221. http://dx.doi.org/10.5032/jae.2014.05207

Robertson, J. F., & Rane-Szostak, D. (1996). Using dialogues to develop critical thinking skills: A practical approach. *Journal of Adolescent & Adult Literacy*, *39*(7), 552-556.

Roksa, J., & Arum, R. (2011). The state of undergraduate learning. *Change: The Magazine of Higher Learning*, *43*(2), 35-38. https://doi.org/10.1080/00091383.2011.556992

Rowe, M. P., Gillespie, B. M., Harris, K. R., Koether, S. D., Shannon, L. J. Y., & Rose, L. A. (2015). Redesigning a general education science course to promote critical thinking. *CBE—Life Sciences Education*, *14*(3), ar30. https://doi.org/10.1187/cbe.15-02-0032

Shaarawy, H. Y. (2014). The Effect of Journal Writing on Students' Cognitive Critical Thinking Skills: A Quasi-Experimental Research on an English as a Foreign Language (EFL) Undergraduate Classroom in Egypt. *International Journal of Higher Education*, *3*(4), 120-128. https://doi.org/10.5430/ijhe.v3n4p120

Shermis, Mark D. (2008). The collegiate learning assessment: A critical perspective. *Assessment Update* 20(2), 10-12. http://dx.doi.org/10.1002/au.202

Shim, W. J., & Walczak, K. (2012). The Impact of Faculty Teaching Practices on the Development of Students' Critical Thinking Skills. *International Journal of Teaching and Learning in Higher Education*, *24*(1), 16-30.

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*(2), 44-49. https://doi.org/10.1080/00091383.2011.550254

Stieha, V., Shadle, S. E., & Paterson, S. (2017). Stirring the pot: Supporting and challenging general education science, technology, engineering, and mathematics faculty to change teaching and assessment practice. *The Journal of General Education*, *65*(2), 85-109. https://doi.org/10.5325/jgeneeduc.65.2.85

Styers, M. L., Van Zandt, P. A., & Hayden, K. L. (2018). Active learning in flipped life science courses promotes development of critical thinking skills. *CBE—Life Sciences Education*, *17*(3), ar39. https://doi.org/10.1187/cbe.16-11-0332

Tsui, L. (2002). Fostering critical thinking through effective pedagogy: Evidence from four institutional case studies. *The Journal of Higher Education*, *73*(6), 740-763. http://dx.doi.org/10.1353/jhe.2002.0056

Tsui, L. (2000). Fostering critical thinking in college students: A mixed-methods study of influences inside and outside of the classroom. University of California, Los Angeles.

Watson, G., & Glaser, E.M. (1994). Watson-Glaser Critical Thinking Appraisal, Form S manual. San Antonio, TX: The Psychological Corporation.

Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, *109*(4), 21-32. https://doi.org/10.3200/AEPR.109.4.21-32