Science Journal of Education
2026, Vol. 14, No. 4, pp. 98–110
https://doi.org/10.11648/j.sjedu.20261404.11

Research Article

Learning How to Measure STEM Student Critical Thinking Through Undergraduate Research Engagement Across a Three-Institution Consortium

Donna Chamely-Wiik[1,*], Daniel Eduardo Meeroff[2], Evelyn Frazier[3], William Reed Kwochka[4], Kloo Hansen[5], Michael Aldarondo-Jeffries[6], Patrice Lancey[7], Uday Nair[7], Kimberly Schneider[8]

[1]Department of Chemistry and Biochemistry, Florida Atlantic University, Boca Raton, United States
[2]Department of Civil, Environmental & Geomatics Engineering, Florida Atlantic University, Boca Raton, United States
[3]Department of Biological Sciences, Florida Atlantic University, Boca Raton, United States
[4]Department of Chemistry and Physics, Western Carolina University, Cullowhee, United States
[5]Office of Research Administration, Western Carolina University, Cullowhee, United States
[6]Divison of Student Success and Well-Being, University of Central Florida, Orlando, United States
[7]Operational Excellence and Assessment Support, University of Central Florida, Orlando, United States
[8]Undergraduate Studies, University of Central Florida, Orlando, United States

Abstract

Undergraduate research is a high-impact practice associated with positive student outcomes, including improved retention, graduation rates, and post-graduation success. It is also linked to gains in intellectual skills, such as critical thinking; however, these gains are typically inferred from student self-reports and faculty observations rather than direct measurement. A key challenge is the lack of robust methods to directly assess critical thinking development through undergraduate research experiences in Science, Technology, Engineering, and Math (STEM). To address this gap, the Learning Environment and Academic Research Network® (LEARN) Consortium was established across three institutions. The consortium implemented a year-long, faculty-mentored research experience through two parallel programs: one for underrepresented first-time-in-college (FTIC) students and another on Associate of Arts (AA) or Associate of Science (AS) transfer students. To directly measure changes in critical thinking the consortium employed a pre-post assessment design using two instruments: 1. the Critical Thinking Assessment Tool (CAT©) and 2. STEM writing assignments evaluated with a specifically designed critical thinking rubric. Results from the CAT© indicated gains in critical thinking for both FTIC and transfer cohorts, with statistically significant gains for the FTIC students at two institutions in two of the three years. Similarly, rubric-based evaluations of STEM writing, showed gains for both cohorts, though these were not statistically significant. This article examines the challenges of directly assessing critical thinking, including issues of rubric development, and shares lessons learned from implementing and evaluating this transferable undergraduate research model across three different institutions.

[*]Correspondence: Donna Chamely-Wiik (dchamely@fau.edu)

Received: 11 June 2026; **Accepted:** 3 July 2026; **Published:** 24 July 2026

Keywords

Undergraduate Research, Critical Thinking, Assessment, First-time-in-college, Transfer Students

1. Introduction

Critical thinking skills are becoming more relevant across academic fields and professional career pathways. These skills are embedded into undergraduate curricula in engineering and computer science [1], physical sciences [2], and other Science, Technology, Engineering, and Mathematics (STEM) fields. Critical thinking skills are also increasingly becoming a differentiator for informing hiring practices [3-8]. Surveys of hiring managers across the country suggest that employers increasingly consider process skills, such as critical thinking, more important than content knowledge when making hiring decisions [3, 9] and as core competencies required in the workforce [5, 6].

Critical thinking involves analysis and evaluation of evidence, consideration of different perspectives, identification of underlying assumptions, and challenging those assumptions to draw well-reasoned and defensible conclusions, followed by reflection [10]. Skills like critical thinking can be taught by modeling structured approaches, such as the scientific method or through constructive peer review feedback [11, 12]. Critical thinking skills are often gained through undergraduate research in STEM fields which involves hands-on, relevant, authentic, and transformative experiences that develop curiosity, information literacy, and problem solving [13-16].

Some of the most common approaches for teaching and gaining critical thinking include Course-Based Undergraduate Research Experiences (CUREs) and through student engagement in faculty-mentored research experiences [17-19]. Typically, these research experiences are scaffolded, which develop critical thinking skills over time, usually with hands-on active participation at its core [20]. Iterative written assignments in STEM, such as an analysis of scientific literature, can also promote the development of students' critical thinking skills by incorporating feedback, reflecting on their thought process, and reviewing and refining their ideas [21, 22].

It has been widely acknowledged that defining and measuring critical thinking skills is challenging, due to the complexity of objectively evaluating learning [10], because the skill is complex, individuals develop critical thinking skills at different paces and ages, and assessors themselves are fallible and can be inconsistent. However, measuring gains in critical thinking skills can be accomplished. Several tools already exist for evaluating critical thinking skills [23]. For example, standardized tests have been employed for assessing critical thinking skills in academic settings including the Watson-Glaser Critical Thinking Appraisal (WGCTA) [24], the California Critical Thinking Skills Test (CCTST) [25], and the Critical Thinking Assessment Test (CAT©) [26].

Rubrics have also been employed to directly assess critical thinking skills through written assignments [27-29]. For example, Harris et al. [30] assembled data from over 200 institutions and 33,000 student responses, indicating a 26% gain in critical thinking over four years of study of the college experience (i.e., first year through senior year), with the largest gains occurring in years 3-4, junior and senior years. Their findings also suggest that gains in critical thinking were not related to the students' academic preparedness as measured by standardized test scores. Furthermore, they also noted that at the single-course level, performance often resulted in little to no gains in critical thinking, although they suggested that the gains may manifest longitudinally over a longer-term, multi-year experience. Ideally, to establish meaningful evaluation of students' critical thinking skills, a combination of assessment tools would be needed to evaluate student critical thinking gains more accurately.

This paper focuses on using multiple measures to determine if critical thinking gains can be documented from a two semester STEM undergraduate research learning community program for first-time-in-college (FTIC) freshmen and transfer students at three different institutions, utilizing coordinated programming, academic assignments, and immersion into academic research. The guiding hypothesis was that gains in critical thinking could be directly measured within the program in a pre-post fashion using 1. A critical thinking assessment diagnostic tool and 2. rubrics for assessing writing assignments within two sequential Introduction to Research courses. Recognizing that direct measures of student learning are challenging across diverse programs, an additional objective was to establish best practices and lessons learned to inform the continuous improvement process for the program.

2. Materials and Methods

Data from three academic years were collected and evaluated from three different universities. This work was approved by each university's Institutional Review Board (IRB): 15-11494, 767795-3, and 959817-4. Two programs were established for the two student populations with the primary objective of increasing retention and student success for 1) incoming first-time in college and 2) transfer students in STEM by providing research experiences the first semester they enter 4-

year universities [31-34]. Through funding from a collaborative grant from the National Science Foundation (NSF), the LEARN Consortium was established consisting of University of Central Florida (UCF), Florida Atlantic University (FAU) and Western Carolina University (WCU). The FTIC freshmen program was initially created at UCF, entitled Freshmen Learning Environment and Academic Research Network® (FLEARN) [34]. The Consortium adapted FLEARN and expanded implementation to the other two institutions within the Consortium [31]. Additionally, the Consortium developed, implemented, and assessed a similar program (TLEARN) modified specifically for STEM transfer students from a 2-year college with an AA or AS degree. Program details and comparison of FLEARN and TLEARN are outlined previously in Chamely-Wiik et al. [32]. The TLEARN program utilized the same three pillars of FLEARN, 1) Academics and Research: combining two Introduction to Research courses (fall and spring semesters) with a research apprenticeship 2) Mentoring: employing a tiered mentoring approach of faculty, program directors, and peer mentors, and 3) Community Building: social programming and other non-curricular events and activities, specifically adapting the model to meet the needs of transfer students [32, 33].

2.1. Course Experience

The LEARN approach consisted of providing undergraduates with mentored research experiences in STEM research laboratories while developing students' research skills in two consecutive Introduction to Research courses in Fall and Spring where transferable scientific research skills were developed and critical thinking skills were assessed. The purpose was to align academic research principles with the curriculum in scaffolded stages. The first Introduction to Research course in the fall semester focused on introducing students to the nature of research, learning comprehension of research literature, and creating components of a research proposal including how to develop a research objective and rationale. By the end of the first term, students in the program had identified a research mentor. The second course in the Spring focused on conducting a research literature review, establishing a testable research question/hypothesis, and delivering a classroom research poster presentation. The courses were designed so that students were simultaneously applying and honing their critical thinking skills to communicate the rationale for their research objective. Efforts were made to standardize the delivery and the implementation timeline of assignments at all three institutions [31] to ensure we could compare and contrast the model effectiveness.

To assess the impact of engaging FTIC and transfer students in undergraduate research experiences early in their academic career on the development of their critical thinking skills, each LEARN Consortium institution used two different assessment tools.

2.1.1. Critical Assessment of Thinking (CAT©)

Each institution administered the Critical Thinking Assessment Tool (CAT) to their students using a pre-post design. The CAT© exam was designed to assess the ability to evaluate information, creative thinking, learning and problem solving, and communication [26, 30]. The instrument, developed and validated, has been extensively tested and refined with a broad range of institutions, faculty, and students across the United States. Most questions require short answer essay responses, and a detailed scoring guide helps ensure scoring reliability, good internal consistency, and established validity with no gender or cultural bias [26]. The CAT instrument was selected because multiple measures aligned closely with the student learning outcomes within the writing assignment rubric.

The pre-test was completed at the beginning of the Fall term, and the post-test was completed at the end of the Spring term. The exams were evaluated externally by the CAT© team at Tennessee Tech, and total scores at each institution were analyzed for statistically significant developmental gains using t-tests. These results were compared between institutions and cohorts.

2.1.2. Research Literature Analysis Writing Assignments

Two research writing assignments were designed by the LEARN Consortium. These pre/post writing assignments were designed to teach the skills of deconstructing and summarizing a research article from primary literature (Fall pre-assignment) and synthesizing information from multiple research articles centered around a singular research question leading to a project hypothesis (Spring post-assignment). These two writing assignments were then evaluated using a critical thinking assessment rubric developed by the LEARN Consortium. The assignments targeted several student learning outcomes including one on critical thinking for comparison with the CAT© results.

Pre-research literature analysis assignment: During the fall semester, students were asked to complete literature review summaries of two related primary research articles within one research topic and critically analyze connections between the articles. FTIC students were assigned specific papers, while transfer students selected their papers with guidance from their research mentors and course instructors. Students were given two weeks to complete the first draft. Feedback was provided by the instructors and through in-class peer-review, after which a final draft was submitted near the end of the semester.

Post-research literature analysis assignment: During the spring term, students completed a similar assignment to the pre-assignment within the same timeframe. The main difference with the post-assignment was the number of primary references being summarized, namely 5 references for FTIC and 10 references for transfer students.

2.1.3. Critical Thinking Assessment Rubric

The initial critical thinking assessment rubric, which used a three-point scale, was adapted from a Student Learning Outcome (SLO) rubric for student engagement in undergraduate research which was developed at Florida Atlantic University for their 2013 Quality Enhancement Plan (QEP) [35]. The SLO rubric included 6 learning outcomes (content knowledge, formulate question, plan of action, critical thinking, communication, ethics), which were inspired by the Association of American Colleges and Universities (AAC&U) Valid Assessment of Learning in Undergraduate Education (VALUE) Rubrics [36]. Florida Atlantic University previously deployed the rubric in a variety of courses at the institution for research-based scaffolded assignments [20].

To acquire more accurate data, after two cohort cycles, a second, more sensitive rubric (Rubric 2) was developed and implemented using a four-point scale. The transition from a three-point (Exemplary, Competent, Developing) to a four-point rubric (Exemplary, Outstanding, Competent, Developing) in the 2019/2020 academic year, however, introduces a potential confounding factor in the interpretation of results. The addition of an intermediate performance category increased scoring granularity but limits direct comparability across cohorts assessed with different rubrics. This effect is discussed in the Reflections section later. Table 1 summarizes the 4-point rubric, which encompasses both iterations of rubric development. Although the four-point rubric enhances assessment resolution, comparisons across rubric versions should be interpreted carefully.

Table 1. *Pre-Post critical thinking assessment rubric.*

Category	Pre rubric	Post rubric
Exemplary	Defines accurate and complete assembly and analysis of information, data, and observations to support the hypothesis and conclusions found Relates the data and findings to the research question(s) being pursued Discusses broader impacts and significance of the research clearly Establishes and aligns a new research question with prior work and within the broader impact context and significance of the research	Accurately relates the data and findings in reference to the research question(s) being pursued Thoroughly evaluates information while acknowledging limits and synthesizing competing points of view Clearly discusses broader impacts and significance of the research
Outstanding	Defines mostly accurate and complete assembly and analysis of information, data, and observations to support the hypothesis and conclusions found, minor revisions needed for clarity Relates most data and findings to the research question(s) being pursued with minor omissions Mentions but does not discuss broader impacts and significance of the research Establishes and aligns a new research question with prior work but not the broader impact context and significance of the research	Relates most data and findings to the research question(s) being pursued; minimal questions remain for the reader Mostly evaluates key information, while acknowledging most limits or synthesizing competing points of view, but minor revisions are needed for clarity Mostly discusses broader impacts and significance of the research, with some revisions needed for further clarity
Competent	Defines mostly accurate (with some minor errors) and mostly complete (with only a few minor omissions) assembly and analysis of information, data, and observations to support the hypothesis and conclusions found Relates some data and findings to the research question(s) being pursued There is some mention of broader impacts and significance of the research Establishes a new research question but does not align with prior work or within the broader impact context and significance of the research	Relates some data and findings to the research question(s) being pursued; some questions remain for the reader Sufficiently evaluates information, while acknowledging some limits or synthesizing competing points of view, but some questions remain for the reader Minimally discusses broader impacts and significance of the research
Developing	Does not: Define assembly and analysis of information, data, and observations or work is largely inaccurate or incomplete Relate data and findings to the research question(s) being pursued Mention broader impacts and significance of the research Establish a new research question, or question is not aligned within the broader context and significance of the research	Does not: Relate data and findings to the research question(s) being pursued; uses generalizations or inaccuracies Evaluate information, while not acknowledging limits or competing points of view Mention broader impacts or significance of the research

2.2. Sampling Methodology

Written student works were collected by each institution and rated by independent evaluators who underwent a formal norming process. In some instances, the student with the selected pre-assignment artifact did not complete a post-assignment in the subsequent term, resulting in a missing pre-post pair. To allow for such losses of samples, the pre-assignments were initially sampled at a slightly higher rate than 50% to the best of our ability. Only students who had completed both the pre and the post assignments were included in this study. Between 50 to 55 FTIC students and 37 to 48 transfer students participated in this assignment for each one of the three years studied, with 22 to 24 FTIC students and 17 to 22 transfer student assignments randomly selected for scoring (approximately 50%).

2.3. Analysis Methodology

2.3.1. Inter-Rater Reliability

Raters underwent a standardized norming procedure and were provided access to a centralized database portal to share and upload documents securely. The norming process was modeled after Crisp [37], involving task review, examination of the assignment prompt, clarifying questions, rubric interpretation/clarification, read/score, discussion/debrief, and analysis of results. To improve inter-rater reliability, the raters held a pre-meeting to calibrate scoring. Each rater applied the rubrics to the assignments individually, recorded their ratings, and then discussed their rating methodology for the critical thinking assessment rubric with each other. Any divergence in the ratings at the individual Student Learning Outcome level was discussed to develop a shared understanding of the rubric application.

For all cohort years, multiple raters applied the critical thinking assessment rubrics to all selected artifacts using the methodology described in Schoepp et al. [38]. *Note:* The 2019-20 student work samples were evaluated by different raters, as the original reviewers were no longer available. However, the original raters provided standardized protocols on the review process and assisted in training the additional raters. After applying the rubric, inter-rater reliability was evaluated using the method of adjacent agreement (within one score indicator) between raters. According to Stemler [39], agreement among raters should be above 70%, and the consensus is that a low percent agreement (below 60%) would suggest significant disagreement between raters such that agreement would likely only be by chance. A summary of the inter-rater reliability using percent adjacent agreement in this study (Table 2) showed values were generally high, varying from 85-98%, which is similar to findings of Jonsson and Svingby [40], and are indicative of consistent general agreement among raters during norming and scoring with the rubric.

Table 2. *Summary of inter-rater reliability for the critical thinking assessment rubric using percent adjacent agreement.*

Year	Rubric type	Pre-	Post-
2017-2018	3-point	85%	97%
2018-2019	3-point	96%	98%
2019-2020*	4-point	98%	91%

* reflects the new 4-point rubric utilized for 2019-2020

For all years, the critical thinking assessment rubric scores ranged from 3.00 to 12.00. After the norming process, each rater individually applied the rubric to their assigned pre- and post-assessments from their assigned sample set.

2.3.2. Data Analysis

Data for all populations was collected through the centralized Institutional Research offices at each of the three institutions. Differences in student demographics, academic preparation, institution, student type, and year were treated as random effects. Acknowledging that the sample size for each year at each institution was relatively small, the total sample size was 109. This is sufficient for most inferential statistics. The hypothesis is that participants in the LEARN program will demonstrate gains in critical thinking as evidenced by a standardized test (CAT© exam) and a pre-post written assignment methodology evaluated with a rubric.

The CAT exam results calculated overall means, standard deviations, and 2-tailed t-test p-values, evaluated by three levels: * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Mean difference divided by pooled group standard deviation. Effect size is reported when significance is observed (0.1 - 0.3 = small effect; 0.3 - 0.5 = moderate effect; >0.5 = large effect). This information was provided in a CAT institutional report authored by Tennessee Tech University Center for Assessment and Improvement of Learning.

Critical thinking assessment rubric scores were aggregated to calculate a growth score using the arithmetic difference between the post- and pre-scores. Significance was evaluated using a two-tailed t-test.

3. Results and Discussion

We analyzed the pre-post assessment of critical thinking using both the CAT exam and written assignments scored with a rubric.

3.1 CAT© Results

For the FTIC CAT© results, Florida Atlantic University (FAU) and Western Carolina University (WCU) data were compared to a previous study conducted by the University of

Central Florida (UCF) in 2015 to evaluate the replicability of this model at two different institutions [34], which is also included in Table 3 for comparison.

Table 3. FTIC CAT© test results for Pre and Post implementations for 3 different cohorts of FTIC students at FAU and WCU in comparison to published results for FTIC cohorts at UCF. Measures include number of students evaluated, mean and standard deviation of CAT test scores.

Year		N FAU	FAU (Starting in 2017)	N WCU	WCU (Starting in 2016)	N UCF	FTIC UCF (Starting in 2011; Schneider et al. 2015)
Year 1	Pre	10	11.90 (±3.84)	19	14.16 (±4.51)	NR***	17.79
	Post	10	15.40 (±4.58)* *p<0.05 effect size +0.83	19	14.68 (±4.01)	NR	20.00
Year 2	Pre	12	18.17 (±5.83)	10	16.40 (±4.93)	20	15.70 (±4.13)
	Post	12	22.08 (±5.90)* *p<0.05 effect size +0.67	10	16.60 (±6.64)	20	19.00 (±4.38)**
Year 3	Pre	9	16.89 (±5.06)	7	14.43 (±5.29)	22	16.36 (±3.47)
	Post	9	18.33 (±6.00)	7	15.29 (±4.23)	22	19.50 (±4.92)* *p<0.05

*p<0.05
**p<0.01
*** Not reported

Overall, the findings demonstrated CAT score gains from the pre- to the post- test at all three institutions for all three years for our FTIC students, with statistical significance observed at two of the three institutions (FAU and UCF) for two consecutive years (Table 3). Note for FTIC LEARN at FAU: Year 3 for FAU was when institutions were impacted by the Covid-19 pandemic and had to convert to remote learning midway through the spring semester. For LEARN however, the program was almost completed by the time institutions converted to virtual (end of March), and all the components of LEARN offered to the participants were preserved for the rest of the semester, so we believe the impact was minimal. Although the increasing pre-post trend is similar to previous years, the pandemic may have impacted the results of the post-exam for our students, which was implemented in April.

Additionally, we evaluated the CAT exam results of the transfer students at all three institutions (Table 4).

Table 4. Transfer CAT test results for pre and post implementations, across 3 institutions and for 3 different cohorts of transfer students. Measures include number of students evaluated, mean and standard deviation of CAT test scores.

Year		N FAU	FAU (starting in 2016)	N WCU	WCU (Starting in 2017)	N UCF	UCF (starting in 2016)
Year 1	Pre	14	14.64 (±4.25)	9	15.56 (±3.97)	11	16.73 (±4.78)
	Post	14	16.21 (± 4.15)	9	16.44 (±3.68)	11	18.36 (±5.87)
	Gain		+1.57		+0.88		+1.63
Year 2	Pre	10	14.50 (±4.70)	11	20.73 (±6.08)	15	16.60 (±4.60)
	Post	10	16.80 (±4.44)	11	19.18 (±4.40)	15	18.73 (±6.02)
	Gain		+2.30		-1.55		+2.13
Year 3	Pre	7	18.86 (±6.15)	13	17.00 (±3.83)	15	16.87 (±5.58)
	Post	7	20.43 (±4.79)	13	20.08 (±3.52)	15	18.40 (±5.41)

Year	N FAU	FAU (starting in 2016)	N WCU	WCU (Starting in 2017)	N UCF	UCF (starting in 2016)
Gain		+1.57		+3.08		+1.53

In summary, for transfer students overall, CAT exam scores increased from pre- to post (Table 4). However, no statistical significance was observed at any of the three institutions for this student population. Perhaps significance was found with the FTIC cohort because their prior knowledge or level of college preparedness may have been lower than the incoming transfer students who already had attained an AA degree. This factor could not be controlled for by institution or cohort year. Additionally, for both FTIC and transfer student cohorts, the numbers of students evaluated were smaller (*n* = 9-15 for transfers compared to 7-22 for FTIC), which may also have impacted our ability to measure statistical significance for transfer students.

3.2. Critical Thinking Assessment Rubric Results

Table 5 summarizes the critical thinking assessment rubric results for the critical thinking SLO for FTIC students, and Table 6 presents the results for transfer students. Pre-post analysis of the written assignment assessed with a rubric was conducted for three consecutive years. In Yrs 1-2, we administered the 3-point rubric, while in Yr 3, the 4-point rubric was administered. After the analysis was complete for cohort years 2017/18 (Cohort Yr 1) and 2018/19 (Cohort Yr 2) using the 3-point critical thinking assessment rubric with no significance determined, the consortium recalibrated the scoring scale and corresponding indicators to be more sensitive to the extent of growth students experienced through the critical thinking assessment rubric assignment. An additional rating labeled "Outstanding" was included in 2019-2020 (Cohort Yr 3) to detect the fine detail between perfect ("Exemplary") and acceptable ("Competent").

In summary, as seen in Table 5 (FTIC), and Table 6 (transfer), once the 4-point rubric was administered for the critical thinking assessment rubric assignments, we were able to demonstrate gains (but not significance) across all three institutions, both for FTIC and for transfer students. Again, statistical significance was not obtained likely because of the relatively small sample size.

Table 5. *Results for pre- and post- critical thinking assessment rubric grading for different FTIC student cohorts across the 3 institutions. Measures include number of students evaluated, mean and standard deviation of rubric scores.*

Year	Measures	N FAU	FAU FTIC	N WCU	WCU FTIC	N UCF	UCF FTIC
Year 1	Pre	5	5.40 (± 2.03)	5	4.45 (± 0.69)	12	5.80 (± 2.14)
	Post	5	4.20 (± 1.68)	5	5.20 (± 1.48)	12	5.96 (± 2.56)
	Gain		-1.20		+0.75		+0.16
Year 2	Pre	6	7.75 (± 1.53)	6	5.63 (± 1.69)	16	6.25 (± 2.49)
	Post	6	6.75 (± 1.41)	6	4.92 (± 0.97)	16	5.44 (± 1.39)
	Gain		-1.00		-0.71		-0.81
Year 3 (with 4-point rubric)	Pre	6	8.50 (± 2.15)	2	4.50 (± 1.73)	16	5.63 (± 1.98)
	Post	6	8.75 (± 2.99)	2	6.00 (± 2.51)	16	6.37 (± 2.61)
	Gain		+0.25		+1.50		+0.74

Table 6. *Results for pre and post critical thinking assessment rubric grading for different transfer student cohorts across the 3 institutions. Measures include number of students evaluated, mean and standard deviation of rubric scores.*

Year	Measures	N FAU	FAU TRANSFER	N WCU	WCU TRANSFER	N UCF	UCF TRANSFER
Year 1	Pre	5	7.35 (± 3.23)	5	6.00 (± 2.44)	7	5.25 (± 1.87)
	Post	5	5.70 (± 3.58)	5	8.40 (± 3.45)	12	9.00 (± 2.71)

Year	Measures	N FAU	FAU TRANSFER	N WCU	WCU TRANSFER	N UCF	UCF TRANSFER
	Gain		-1.65		+2.40		+3.75
Year 2	Pre	7	8.09 (± 2.54)	5	7.35 (± 1.32)	10	5.06 (± 3.54)
	Post	7	6.79 (± 2.97)	5	5.30 (± 2.20)	10	5.90 (± 1.59)
	Gain		-1.30		-2.05		+0.84
Year 3 (with 4-point rubric)	Pre	6	7.50 (± 2.02)	5	6.60 (± 1.26)	11	6.68 (± 2.44)
	Post	6	9.30 (± 2.38)	5	8.40 (± 1.90)	11	8.05 (± 2.52)
	Gain		+1.80		+2.20		+1.38

3.3. Critical Thinking Assessment Results: 4-Point Rubric

To determine if statistical significance could be attained by aggregating student artifact analyses from all institutions and both programs, we conducted the following analysis within the critical thinking assessment rubric study for Year 3 (4-point rubric) only. The findings were presented in Table 7.

Table 7. *Critical thinking assessment 4-point rubric grading for different FTIC and transfer student cohorts grouped across the 3 institutions. Student type reflects the two student cohorts FTIC and Transfer pooled across 3 institutions. Measures include mean of rubric scores, total student number and learning gain.*

Year	Institution			Student Type		Overall
	FAU	WCU	UCF	FTIC	TRANSFER	
Year 3 Pre	8.000 (n=12)	6.000 (n=7)	6.056 (n=27)	6.250 (n=24)	6.886 (n=22)	6.554 (n=46)
Year 3 Post	9.000 (n=12)	7.714 (n=7)	7.056 (n=27)	6.938 (n=24)	8.455 (n=22)	7.663 (n=46)
Gain	+1.000	+1.714	+1.000	+0.688	+1.569	+1.109

The 4-point rubric did improve the ability to distinguish the fine detail differences between acceptable quality and near perfect quality student work (Table 7). In summary, for the critical thinking assessment rubric with the 4-point rubric in Year 3, students, regardless of institution or student type (FTIC or transfer), showed growth in critical thinking skills from pre to post. We also conducted t-tests on the results of the writing assignment rubric by comparing the pre versus post assessment rubric scores (Table 8). This modification may have redistributed student ratings. For example, students previously classified as "Exemplary" may have been reassigned to "Outstanding," however, the increased sensitivity of the four-point scale may capture incremental gains not detectable under the original rubric.

Table 8. *Summary of pre vs post t-test data to assess critical thinking assessment rubric on critical thinking skills by way of assessment rubrics, listed years, institution, and program type.*

Year	Institution	LEARN Cohort	P value for t-test
2017-2018	FAU	FTIC	0.339
		TRANSFER	0.466
	WCU	FTIC	0.348
		TRANSFER	0.244
	UCF	FTIC	0.874

Year	Institution	LEARN Cohort	P value for t-test
		TRANSFER	0.012*
	FAU	FTIC	0.265
		TRANSFER	0.395
2018-2019	WCU	FTIC	0.400
		TRANSFER	0.119
	UCF	FTIC	0.268
		TRANSFER	0.249
	FAU	FTIC	0.843
		TRANSFER	0.016*
2019-2020	WCU	FTIC	0.500
		TRANSFER	0.067
	UCF	FTIC	0.263
		TRANSFER	0.134

* Statistical significance is indicated when the value is less than 0.05.

Overall, the t-test did not indicate significant statistical difference in critical thinking gains between the pre- and post-scores of students in either FTIC or transfer programs, across 3 years. The only exceptions were for transfer students at UCF in the cohort of 2017-2018 and again for transfer students at FAU for the cohort of 2019-2020.

Since inter-rater reliability met the criteria and no differences were noted by institution, the independent raters' scores were averaged for both pre and post scores by student type (FTIC vs. transfer), and a t-test was performed to determine if gains were significant. As can be seen in Table 9, only Yr 3 transfer scores indicated statistically significant gains.

Table 9. *Summary of t-test results of aggregated FTIC and transfer students from 3 institutions of pre-post critical thinking learning gains by year.*

All Institutions	FTIC	TRANSFER
2017	0.976	0.091
2018	0.104	0.462
2019	0.285	0.003*

* statistical significance is indicated when the p-value is less than 0.05.

Since learning is iterative, and deep learning happens when students regularly engage in the cycle or practice of coherently connecting concepts and then attempting to relate those concepts to new ideas and applications [41, 42], the LEARN Consortium's efforts to recalibrate to the 4-point scale was indicative of the need to emphasize the iterative process of student learning and growth. The 4-point scale enabled instructors and students to be more engaged in providing and receiving supportive feedback, setting and applying realistic goals to improve writing quality, and drawing connections among the scientific process, research, and the LEARN course sequence.

4. Reflections, Lessons Learned and Recommendations

Research has identified many challenges associated with assessment of critical thinking [43], and yet the need to continue efforts to assess this key component of student learning remains important. The lessons learned along the way can provide guidance to other institutions who are interested in collaborative assessment of student learning across multiple institutions and to different student audiences. Below we present lessons learned from this study, categorized by topic, to facilitate others interested in adapting one or more aspects of the consortium's efforts within this program.

4.1. General Reflections, Implications, and Recommendations

The findings in this study, offer several implications for the assessment of students involved in undergraduate research, particularly at community colleges and institutions serving underrepresented student populations. First, the use of multiple assessment instruments proved essential for capturing gains in critical thinking, across diverse groups, institutions, and years. This suggests that reliance on a single measure may underestimate program impact while using more than one assessment tool approach may yield a more robust and comprehensive evaluation of student learning outcomes. We recommend that institutions, particularly community colleges and those that serve underrepresented student populations, seeking to assess critical thinking development, consider adopting multiple complimentary assessment tools.

Second, although increasing sample size would strengthen the statistical power and improve the ability to detect gains, doing so may require additional resources in support of larger numbers of participants. The observed statistically significant gains among transfer students when data were pooled indicate that meaningful gains can be detected within resource-constrained settings. Conversely, the absence of statistically significant gains for FTIC students suggests that there may be differences among student types, perhaps related to differences in starting points, prior preparation, or responsiveness to research-based interventions.

To address Harris et al. 2014's finding that single-course interventions often yield little to no improvement in critical thinking, we implemented a two-semester, sequence-based experience. This approach appears to support measurable gains over a relatively short timeframe. We recommend utilizing a pre/post/final design to measure longitudinal growth (or regression) of critical thinking over the 4-year academic program. Considerations for adding a third or final touchpoint later in their academic programs may have demonstrated gains that take a longer time to manifest. This additional third touchpoint could be a culminating research experience or capstone within an academic program. Prior literature indicates that the real demonstrated growth in critical thinking can/should be expected in their junior/senior year after multiple opportunities to apply this skill with feedback loops are offered. It is, however, promising that we did observe gains using both instruments after only two semesters.

Finally, the results highlight the need for early and sustained engagement in undergraduate research, particularly for transfer students. Implementing a bridge program with feeder institutions may initiate earlier involvement, promote continuity, and ease transitions into research activities within a university setting. In all, these findings suggest that a structured, multi-semester research experience, combined with comprehensive assessment strategies and intentional support of transfer pathways, can enhance critical thinking and encourage a smooth transition into undergraduate research.

4.2. CAT Exam

At the time of the study, we were interested in determining whether gains in critical thinking for FTIC students previously demonstrated at University of Central Florida with CAT (Schneider et al, 2015) could be replicated at two other institutions. We recommend that all participating institutions conduct the FTIC student CAT analyses at the same time, to isolate differences in prior knowledge of different cohorts. One way to potentially achieve this would have been to have a matched control group of non-participants at each institution for each cohort year and student type taking the CAT test, if funds are available.

4.3. Critical Thinking Assessment Rubric

Writing based assessments involve assignment design, design of rubric, and inter-rater reliability. Lessons learned are presented as follows.

Assignment Design: In our efforts to capture scaffolded learning of critical thinking between the two touch points (pre-post), we made a decision to make the post-assignment slightly different than the pre-assignment. In doing so, we introduced an additional variable which posed a challenge from an assessment standpoint. Ideally, the pre-post assignments should be the same, with considerations for still ensuring the usefulness of the assignments to the students.

Rubric design: In this study the pre-post rubrics were different, which introduced another variable for the analysis. It would have been more ideal to use the same rubric for pre- and post-assessments and include both novice-level and advanced-level critical thinking indicators in both the pre and post rubric to better evaluate growth. Additionally, there was a switch from a 3-point to a 4-point rubric in year 3, when we discovered the challenges of differentiating between competent and exemplary within the 3-point rubric. Adding in the additional category of "outstanding" allowed for greater sensitivity to capture the differences between a perfect score and an acceptable score. Ideally, it would have been beneficial to use the more sensitive 4-point rubric from the beginning of the study.

Norming: Since the adjacent agreement inter-rater reliability was evaluated post assessment, it is preferred to evaluate this value prior to concluding the norming process and conducting the assessment. As indicated, raters changed within the study timeframe, it is preferable to have the same raters complete the assessments for all years without introducing potential variability or bias by introducing new raters. For future years, there are now Artificial Intelligence agents that are becoming more readily available to score rubrics instead of using human raters. This practice can potentially reduce rater bias.

Maximizing sample size: For our study, initially the students selected to be evaluated for the pre-assessments were evaluated at a different touchpoint than the post-assessments. In doing so, we were faced with non-completer students who were selected during the pre-assessment but never completed the post assignment. This practice artificially limited sample size, so we recommend the dataset be sorted by students with completed pre and post assignments prior to randomizing and assessing. This would maximize the n value from the available sample.

Working within a multi-institution consortium:

The LEARN Consortium used the backward design framework to standardize both the course design and the implementation, leading to consistency across the institutions, with a shared syllabus, course schedules, and course materials. All curricular materials were also shared in an open-source format (e.g., Google Drive). The Consortium engaged in bi-weekly meetings, including regular debriefs and reflections on progress and continued a norming process for rubric use at every level of implementation. We encourage other cross-institutional teams to utilize these successful approaches, including encouraging the instructors to regularly participate in the norming process for each cohort, as echoed by Belanger et al. [43].

5. Conclusions

The findings demonstrated increased CAT scores from the pre- to the post- test at all three institutions for all three years for our FTIC freshmen students, with statistical significance observed at two of the three institutions (UCF and FAU) for two consecutive years. For transfer students, we observed increases in CAT© scores from pre-to-post. Although no statistical significance was observed at any of the three institutions for the transfer student population, we are encouraged by the trend of increasing gains.

The findings for the critical thinking assessment rubric analysis demonstrated no statistically significant gains for cohort years 2017/18 (Cohort Yr 1) and 2018/19 (Cohort Yr 2) using the 3-point critical thinking assessment rubric. However, once the 4-point rubric was administered, gains were observed across all three institutions, both for FTIC (without statistical significance) and for transfer students (with statistical significance). Additionally, we demonstrated significant gains in critical thinking for transfer students when we pooled all students from the Consortium. In terms of lessons learned, we identified best practices in sample size and selection; rubric design, the norming process, and deployment; and assignment design and implementation across multiple faculty, institutions and years. Finally, using multiple measures of critical thinking assessment allowed us to demonstrate that providing the opportunity for mentored research experiences in the early years of undergraduate study demonstrated gains in critical thinking for all groups, all institutions within the Consortium, and all years.

Abbreviations

STEM	Science, Technology, Engineering, and Math
LEARN	Learning Environment and Academic Research Network®
FTIC	First time in college
AA	Associate of Arts
AS	Associate of Science
CAT©	Critical Thinking Assessment Tool
CURE	Course-Based Undergraduate Research Experience
WGCTA	Watson-Glaser Critical Thinking Appraisal
CCTST	California Critical Thinking Skills Test
IRB	Institutional Review Board
NSF	National Science Foundation
UCF	University of Central Florida
FAU	Florida Atlantic University
WCU	Western Carolina University
FLEARN	Freshmen Learning Environment and Academic Research Network®
TLEARN	Transfer Learning Environment and Academic Research Network®
SLO	Student Learning Outcome
QEP	Quality Enhancement Plan
AAC&U	Association of American Colleges and Universities
VALUE	Valid Assessment of Learning in Undergraduate Education

Acknowledgments

Funding for this project was provided by the NSF's IUSE program, grants 1524601, 1524666, and 1524607. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. The authors would like to acknowledge the valuable contributions of Jodiene Johnson, Dr. Jordan Merritt, Colleen M. Smith, and Dr. Mary Tripp for supporting the critical thinking assessment rubric efforts. Their insights and efforts greatly enhanced the quality and scope of this work.

Author Contributions

Donna Chamely-Wiik: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing

Daniel Eduardo Meeroff: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing

Evelyn Frazier: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Visualization, Writing – original draft, Writing – review & editing

William Reed Kwochka: Funding acquisition, Project administration, Supervision, Writing – review & editing

Kloo Hansen: Supervision, Validation, Writing – review & editing

Michael Aldarondo-Jeffries: Supervision, Writing – review & editing

Patrice Lancey: Validation, Writing – review & editing

Uday Nair: Validation, Writing – review & editing

Kimberly Schneider: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – review & editing

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] ABET. (2025). Criteria for accrediting engineering programs, 2026–2027. ABET. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2026-2027/

[2] Undergraduate professional education in chemistry: ACS guidelines and evaluation procedures for bachelor's degree programs. (2015). American Chemical Society.

[3] Gray, K., & Koncz, A. (2017). The Key Attributes Employers Seek on Students' Resumes (Publication). National Association of Colleges and Employers Website. http://www.naceweb.org/about-us/press/2017/the-key-attributes-employers-seek-on-students-resumes/

[4] Pearl, A. O., Rayner, G. M., Larson, I., & Orlando, L. (2019). Thinking about critical thinking: An industry perspective. *Industry and Higher Education*, 33(2), 116-126. https://doi.org/10.1177/0950422218796099

[5] Retnawati, H., Djidu, H., Apino, E., & Anazifa, R. D. (2018). Teachers' Knowledge about Higher-Order Thinking Skills and Its Learning Strategy. *Problems of Education in the 21st Century*, 76(2), 215-230. https://doi.org/10.33225/pec/18.76.215

[6] Chonkaew, P., Sukhummek, B., & Faikhamta, C. (2016). Development of analytical thinking ability and attitudes towards science learning of grade-11 students through science technology engineering and mathematics (STEM education) in the study of stoichiometry. *Chemistry Education Research and Practice*, *17*(4), 842-861. https://doi.org/10.1039/x0xx00000x

[7] National Association of Colleges and Employers, NACE, B. C. R. S. A. (2022).

[8] Stefán, C. I. (2023). The world economic Forum. In The Palgrave handbook of non-state actors in East-West relations (pp. 1-13). Cham: Springer International Publishing.

[9] Economics, D. A. (2018). Deloitte Access Economics. Canberra (AUST).

[10] Butler, H. A. (2024). Predicting Everyday Critical Thinking: A Review of Critical Thinking Assessments. *Journal of Intelligence*, 12(2), 16-28. https://doi.org/10.3390/jintelligence12020016

[11] Sari, D. S., Widiyawati, Y., Nurwahidah, I., & Setiawan, T. (2023). STEM critical thinking assessment for measuring students' critical thinking skills in the automotive chemistry course. *Jurnal Penelitian Pendidikan IPA*, 9(7), 5289–5295. https://doi.org/10.29303/jppipa.v9i7.4750

[12] Sari, R. M., Sumarmi, S., Astina, I. K., Utomo, D. H., & Ridhwan, R. (2021). Increasing students critical thinking skills and learning motivation using inquiry mind map. *International Journal of Emerging Technologies in Learning (iJET)*, 16(03), 4-19. https://doi.org/10.3991/ijet.v16i03.16515

[13] Ashcroft, J., Blatti, J., & Jaramillo, V. (2020). Early career undergraduate research as a meaningful academic experience in which students develop professional workforce skills: A community college perspective. In Integrating professional skills into undergraduate chemistry curricula (pp. 281-299). American Chemical Society. https://doi.org/10.1021/bk-2020-1365.ch016

[14] Chadha, D., & Nicholls, G. (2006). Teaching transferable skills to undergraduate engineering students: Recognizing the value of embedded and bolt-on approaches. *International Journal of Engineering Education*, 22(1), 116-122.

[15] Chamely-Wiik, D., Ambrosio, A., Baker, T., Ghannes, A., & Soberon, J. (2023). Impact of undergraduate research experience intensity on measures of student success. *Journal of the Scholarship of Teaching and Learning*, *23*(1), 14-30. https://doi.org/10.14434/josotl.v23i1.32675n

[16] Hernandez, N. V., Fuentes, A., & Crown, S. (2018, October). Effectively transforming students through first year engineering student experiences. In 2018 IEEE Frontiers in Education Conference (FIE) (pp. 1-5). https://doi.org/10.1109/FIE.2018.8658752

[17] Koh, G. C., Khoo, H. E., Wong, M. L., & Koh, D. (2008). The effects of problem-based learning during medical school on physician competency: a systematic review. *Canadian Medical Association Journal*, *178*(1), 34–41. https://doi.org/10.1503/cmaj.070565

[18] Marhamah, M., Fatmawati, B., Sarwati, S., & Ariandani, N. (2024). How to Train Critical Thinking Skills?: Application of Problem-Based Learning Model Lesson Study pattern. *Jurnal Kependidikan Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan Pengajaran Dan Pembelajaran*, 10(3), 1283. https://doi.org/10.33394/jk.v10i3.12751

[19] Meir, E., Pope, D., Abraham, J. K., Kim, K. J., Maruca, S., & Palacio, J. (2024). Designing activities to teach higher-order skills: How feedback and constraint affect learning of experimental design. *CBE—Life Sciences Education*, *23*(1), ar 1-20. https://doi.org/10.1187/cbe.22-08-0158

[20] Chamely-Wiik, D., Dunn, K., Kirsch, P., Holman, M., Meeroff, D., & Peluso, J. (2014). Scaffolding the development of student research skills for Capstone Experiences: a multi-disciplinary approach. *Council on Undergraduate Research Quarterly*, 34(4), 18-25.

[21] Hunter, A., Laursen, S. L., & Seymour, E. (2006). Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education*, 91(1), 36–74. https://doi.org/10.1002/sce.20173

[22] Lopatto, D. (2007). Undergraduate research experiences support science career decisions and active learning. CBE—Life Sciences Education, 6(4), 297–306. https://doi.org/10.1187/cbe.07-06-0039

[23] Landrum, R. E., & McCarthy, M. A. (2015). *Chapter 7: Measuring Critical Thinking Skills*. A compendium of scales for use in the scholarship of teaching and learning, 74-86.

[24] Watson, G. (1980). *Watson-Glaser critical thinking appraisal* (Vol. 3). San Antonio, TX: Psychological Corporation.

[25] Facione, P. A. (1990). The California Critical Thinking Skills Test - college level. Technical report #1. Experimental validation and content validity. California Academic Press.

[26] Stein, B., Haynes, A., Redding, M., Ennis, T., Cecil, M. (2007). Assessing Critical Thinking in STEM and Beyond. In: Iskander, M. (eds) Innovations in E-learning, Instruction Technology, Assessment, and Engineering Education. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6262-9_14

[27] Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *Journal of Effective Teaching*, *7*(1), 3-14.

[28] Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to assess critical thinking and information processing in undergraduate STEM courses. *International Journal of STEM Education*, *7*(1). https://doi.org/10.1186/s40594-020-00208-5

[29] Suhairom, N., Saipudin, N. A., Bakar, Z. A., Ibrahim, H., Talib, R., & Haladin, N. B. (2024). Communication Skills Assessments and Rubrics In The Higher Education: A Meta-Analysis. *Sains Humanika*, *16*(3), 37–47. https://doi.org/10.11113/sh.v16n3.2159

[30] Harris, K., Stein, B., Haynes, A., Lisic, E., & Leming, K. (2014). Identifying courses that improve students' critical thinking skills using the CAT instrument: a case study. In *Proceedings of the 10th Annual International Joint Conferences on Computer, Information, System Sciences, and Engineering* (Vol. 10, pp. 1-4).

[31] Meeroff, D., Chamely-Wiik, D., Kwochka, W., Frazier, E., Merritt, J., Aldarondo-Jeffries, M., Morrison-Shetlar, A., & Schneider, K. (2020). Work in Progress: A Transferable Model to Improve Retention and Student Success in STEM through Undergraduate Research (NSF LEARN Consortium). In *2019 ASEE Annual Conference & Exposition.* https://doi.org/10.18260/1-2--3358

[32] Chamely-Wiik, D., Frazier, E., Meeroff, D., Merritt, J., Johnson, J., Kwochka, W. R., Morrison-Shetlar, A. I., Aldarondo-Jeffries, M., Schneider, K. R. (2021). Undergraduate Research Communities for Transfer Students: A retention model based on factors that most influence student success. *Journal of the Scholarship of Teaching and Learning*, 21(1), 193–224. https://doi.org/10.14434/josotl.v21i1.30273

[33] Schneider, K., Aldarondo Jeffries, M., Smith, C., Chamely-Wiik, D., Kwochka W., Meeroff, D., & Frazier, E. (2024). Undergraduate Research Programs for STEM Transfer Success: A Multi-Institutional Approach. *Scholarship and Practice of Undergraduate Research*, 8(1), 54-63. https://doi.org/10.18833/spur/8/1/5

[34] Schneider, K. R., Bickel, A., & Morrison-Shetlar, A. (2015). Planning and Implementing a Comprehensive Student-Centered Research Program for First-Year STEM Undergraduates. *Journal of College Science Teaching*, 44(3), 37–43. https://doi.org/10.2505/4/jcst15_044_03_37

[35] Chamely-Wiik, D. (2013). Distinction through Discovery: Expanding a Culture of Undergraduate Research and Inquiry. Florida Atlantic University, https://www.fau.edu/ouri/documents/quality-enhancement-plan-1-11-13.pdf

[36] Association of American Colleges and Universities. (2009). Valid Assessment of Learning in Undergraduate Education (VALUE). https://www.aacu.org/initiatives/value

[37] Crisp, E. A. (2017). Calibration: Are you seeing what I's seeing? *Intersection, Winter*, 1(3), 7-13.

[38] Schoepp, K., Danaher, M., & Kranov, A. A. (2018). An effective rubric norming process., *Practical Assessment, Research & Evaluation*, 23(11), 1-12. https://portal.issn.org/resource/ISSN/1531-7714

[39] Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, 9(4). 1-11. https://doi.org/10.7275/96jp-xz07

[40] Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

[41] Biggs, J. B., & Collis, K. F. (1982). The psychological structure of creative writing. *Australian Journal of Education*, 26(1), 59-70. https://doi.org/10.1177/000494418202600104

[42] Hodges, L. C., & Harvey, L. C. (2003). Evaluation of student learning in organic chemistry using the SOLO taxonomy. *Journal of Chemical Education*, 80(7), 785-787. https://doi.org/10.1021/ed080p785

[43] Belanger, J., Zou, N., Mills, J. R., Holmes, C., & Oakleaf, M. (2015). Project RAILS: Lessons Learned about Rubric Assessment of Information Literacy Skills. *Portal Libraries and the Academy*, *15*(4), 623–644. https://doi.org/10.1353/pla.2015.0050