Using Case Studies to Improve the Critical Thinking Skills of Undergraduate Conservation Biology Students

Ana L. Porzecanski 1, Adriana Bravo 2,3, Martha J. Groom 4, Liliana M. Dávalos 5,6, Nora Bynum 1, Barbara J. Abraham 7, John A. Cigliano 8,9, Carole Griffiths 10, David L. Stokes 4, Michelle Cawthorn 11, Denny S. Fernandez 12, Laurie Freeman 13, Timothy Leslie 10, Theresa Theodeose 14, Donna Vogler 15, and Eleanor J. Sterling 1

1 Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY, USA
2 Department of Natural History, Royal Ontario Museum, Toronto, Ontario, Canada
3 Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada
4 School of Interdisciplinary Arts & Sciences, University of Washington Bothell, WA, USA
5 Department of Ecology and Evolution, Stony Brook University, New York, NY, USA
6 Consortium for Inter-Disciplinary Environmental Research, Stony Brook University, New York, NY, USA
7 Department of Organismal and Environmental Biology, Christopher Newport University, Newport News, VA, USA
8 Biological Sciences, Environmental Conservation Program, Cedar Crest College, Allentown, PA, USA
9 Schoodic Institute at Acadia National Park, Winter Harbor, ME, USA
10 Department of Biology, Long Island University, Brooklyn, NY, USA
11 Department of Biology, Georgia Southern University, Statesboro, GA, USA
12 Department of Biology, University of Puerto Rico, Humacao, Puerto Rico
13 Science Division at Fulton-Montgomery Community College, Johnstown, NY, USA
14 Department of Biology, University of Southern Maine, Portland, ME, USA
15 Department of Biology, State University of New York Oneonta, Oneonta, NY, USA

Email: alporze@amnh.org

ABSTRACT Critical thinking (CT) underpins the analytical and systems-thinking capacities needed for effective conservation in the 21st century but is seldom adequately fostered in most postsecondary courses and programs. Many instructors fear that devoting time to process skills will detract from content gains and struggle to define CT skills in ways relevant for classroom practice. We tested an approach to develop and assess CT in undergraduate conservation biology courses using case studies to address both challenges. We developed case studies with exercises to support content learning goals and assessment rubrics to evaluate student learning of both content and CT skills. We also developed a midterm intervention to enhance student metacognitive abilities at a light and intensive level and asked whether the level of the intervention impacted student learning. Data from over 200 students from five institutions showed an increase in students’ CT performance over a single term, under both light and intensive interventions, as well as variation depending on the students’ initial performance and on rubric dimension. Our results demonstrate adaptable and scalable means for instructors to improve CT process skills among undergraduate students through the use of case studies and associated exercises, aligned rubrics, and supported reflection on their CT performance.

KEYWORDS rubric, assessment, case studies, conservation, critical thinking, biology

INTRODUCTION

Educating the next generation of professionals to address complex conservation and environmental challenges involves more than teaching disciplinary principles, concepts, and content—it also requires cultivating core competencies in critical thinking (CT), collaboration, and communication [1, 2]. CT skills are key desired outcomes of college and university education [3] and can strongly...
influence how students make life decisions [4]. Unfortunately, college graduates in the United States appear to lack strong CT skills despite several years of instruction [5-7]. This may be due to overreliance on teaching and assessment approaches that emphasize mastery of large volumes of content and offer few opportunities to think critically while acquiring and using knowledge [8, 9]. Courses that take an alternative approach, using active, collaborative, or inquiry-based approaches to learning could contribute to both long-term retention of knowledge and CT skills [10-12].

Using case studies to support active, inquiry-based approaches can be especially effective [13, 14]. Case study pedagogies are well suited to supporting the development of CT skills because of their sustained focus on a theme with applications in a specific setting and the opportunity to emphasize distinct steps in the processes of understanding and analyzing issues that comprise essential CT skills. Creating exercises that foster CT using a case study approach combines strengths from both inquiry-based and case study–based best practices.

While definitions vary, CT is broadly recognized as “a habit of mind characterized by the comprehensive exploration of issues and evidence before accepting or formulating an opinion or conclusion” [15]. CT involves higher-order thinking skills, as well as a suite of concrete capacities, including the ability to select, analyze, infer, interpret, evaluate, and explain information, as well as draw conclusions based on varied and conflicting evidence [15-17]. Not confined to specific analytical tasks, strong CT skills support the ability to think in a complex manner and to process and assess diverse inputs in a constantly changing environment [17]. This capacity is essential to effective decision-making, problem solving, and adaptive management in conservation research and practice, particularly in addressing the tradeoffs and multiplicity of perspectives at the core of environmental concerns.

CT has been a focus of K–12 educational and cognitive researchers, who show that explicit instruction can enhance learning of CT skills [18-21]. Unfortunately, adoption of these ideas and practices has been slower in tertiary STEM education [22]. Educators in undergraduate science classrooms rarely prioritize explicit direct instruction in CT skills or their assessment, fearing compromising the time available for “coverage” of content [6, 23]. Thus, many instructors rely on teaching and assessing core content, assuming the CT skills will automatically develop along with deeper disciplinary knowledge [17, 24]. Further, educators typically lack training on CT instruction [25]. Perhaps not surprisingly, on average, very small or empirically nonexistent gains in CT or complex reasoning skills have been found in a large proportion of students over the course of 4-year college programs [6, 7, 26].

Besides the potential to enhance learning outcomes, an emphasis on CT skills through more active and collaborative learning can also promote equal opportunity in STEM and boost completion rates. It has been shown that these approaches to teaching and learning can enhance learning for underrepresented groups in science [27] and could also boost performance and hence retention in the field in the midst of current high attrition rates in STEM [28].

Our experience running faculty professional development programs in diverse contexts over several years [29] has shown that faculty in diverse learning contexts seek and welcome evidence-based guidance on teaching and assessment practices that promote CT. Given only informal preparation on building CT skills and a curricular focus on essential disciplinary concepts, instructors often search for guidelines on how to incorporate these practices while supporting the simultaneous learning of concepts. Case studies can provide a particularly strong way to support the enhancement of CT skills that are adaptable for individual instructors.

To better understand the investment in time and effort needed for conservation students to learn process skills and for faculty to develop efficient teaching tools, we designed a multi-institutional study on three fundamental process skills: oral communication [30], data analysis [31], and CT. These different skills were selected to match the diverse interests of the participating faculty and were targeted by different faculty in different “arms” of the study (in different institutions, courses, and groups of students) to allow for comparison among results. Here we report on the results for CT. A key component of this portion of the study was the use of case studies to foster both content and skill development.

Our study design built on evidence showing that case study exercises help reinforce concept knowledge, as well as cognitive skills, and further, that repetition and reflection [32] support development of higher-order thinking skills. We investigated three questions: (1) Does instructor emphasis on CT skills—providing metacognitive support...
for reflection on their performance at light and intensive levels—influence the magnitude of individual CT skill gains? (2) Do students show similar responses for the different dimensions of CT learning, or are any of them more challenging than others? and (3) How does our intervention influence students at different initial achievement levels?

To address these questions, we first created and validated instructional materials in the form of case study exercises and assessment rubrics designed to develop and assess four main dimensions of CT skills (see below) and piloted these materials in diverse classroom settings across five institutions. We assessed student learning using a common rubric to score CT performance on two case study–based exercises, and using an independent assessment of CT skills (The Critical Thinking Assessment Test (CAT) [33]), applied at the start and end of each course. To address the frequent concern of instructors regarding trade-off with content learning, we investigated these questions while also measuring content gains. A key aim of this study was to develop and use approaches for active teaching using case studies that instructors can readily adopt as part of their regular teaching practices.

METHODOLOGY

Developing, Validating, and Implementing Assessment Tools

Between April and July 2011, we created and validated a set of instructional materials based on case studies designed to develop CT skills (Instructional Unit for CT skills). The Instructional Unit consisted of (1) Case Study Exercise 1 on amphibian declines, with a solution file, (2) Case Study Exercise 2 on invasive species, with a solution file, (3) a pre/post content knowledge assessment for each exercise, (4) a student’s pre/post self-assessment of their CT skills, (5) our CT Rubric, and (6) the files associated with the intensive versus light Teaching Intervention, including a third brief Case Study on climate change used in the intensive intervention, with a solution file. The complete Instructional Unit as used in the study, as well as updated versions of the case studies, can be downloaded from the website of the Network of Conservation Educators and Practitioners (NCEP).1

DEVELOPMENT AND VALIDATION OF THE CT RUBRIC AND CASE STUDY EXERCISES. To evaluate student CT performance, we developed a rubric based on elements found in existing and available rubrics (e.g., Washington State University’s Guide to Rating Critical & Integrative Thinking from 2006, and Northeastern Illinois University CT Rubric from 2006) and the VALUE Rubric for CT [34]. The resulting rubric included descriptions of four performance levels (from 1 to 4) for four dimensions of CT: (1) explanation of issues or problems, (2) selection and use of information, (3) evaluation of the influence of context and assumptions, and (4) reaching positions or drawing conclusions. The final rubric drew on broadly validated rubrics and was adapted by a core group of eight participating project faculty at a workshop in 2011. Using a collaborative and participatory approach to rubric development, we sought to validate rubric content, ensure familiarity of faculty participants with the rubric, and minimize scoring differences among project participants.

We then developed two exercises based on real-world case studies, as recommended by the Vision and Change Report [1]. Case study topics were selected to correspond to core topics that could be incorporated into all courses with minimal syllabus disruption. We developed Case Study Exercise 1 with a focus on threats to biodiversity, specifically on understanding the causes of amphibian declines. We adapted Case Study Exercise 2 on the topic of invasive species, specifically on rusty crayfish in the Eastern United States, from a version previously published by the NCEPs (http://ncep.amnh.org). Each case study exercise contained three main parts: (1) a short introduction and instructions to the exercise, (2) the case study, and (3) a section with questions designed to prompt students’ CT skills in relation to the case. Each case study exercise was designed to teach conservation biology content in alignment with the CT skills assessed in the rubric; it included questions and tasks intended to elicit student performance in each of the four CT dimensions described in the rubric.

IMPLEMENTATION OF THE CASE STUDY EXERCISES AND CT INTERVENTIONS. Between August 2011 and August 2013, we implemented the Instructional Unit following

1. Both versions can be downloaded by registering as an educator on http://ncep.amnh.org. To find the original versions used in the study, which also include all instructions given to participating faculty, see “NSF CCLI / TUES Instructional Unit: Critical Thinking.” For classroom-ready, updated versions of the case studies, see “Applying Critical Thinking to an Invasive Species Problem,” and “Applying Critical Thinking to the Amphibian Decline Problem.” These cite more current literature and have been edited for clarity.

Improving Critical Thinking With Case Studies
Within a term, do students gain:
• Skills?
• Content knowledge?
• Confidence?

Figure 1. Experimental design and main questions within and across terms. The discontinuous arrow between Light and Intensive Teaching Interventions (TI) indicates an interchangeable order. Abbreviations are as follows: CAT = Critical Thinking Assessment Test; CE Ex = case study exercise. Not all students completed all components, so total sample size differs in our analyses of these data. N ranged from 52 to 82 for completion of the pre-/postcase study exercise content assessments for the specific case studies, with \( N = 113 \) total for the light teaching intervention and \( N = 103 \) total for the intensive teaching intervention. For CT skill gains across the case study exercises, \( N = 216 \). For the pre- and postcourse student-self assessments, \( N = 76 \) for the light and 79 for the intensive teaching intervention. \( N = 78 \) for the light and 71 for the intensive teaching interventions for the pre- and postcourse CAT tests.

the experimental design shown in figure 1 in upper-level conservation biology courses given at five U.S. higher education institutions (table 1). Case Study Exercise 1 was administered within the first 2 weeks of class as a preassessment and Case Study Exercise 2 was administered within the last 2 weeks of class in the term as a postassessment (see figure 1). To guide and facilitate data collection, we provided each professor with a scoring guide to assign points to answers to each question in the case study exercise and a spreadsheet to enter points. Scores from specific questions were assigned to one of the four CT dimensions. Professors then reported these final scores on each dimension of the rubric to the students. Scores from both case study exercises contributed toward students’ grades.

We evaluated whether students gained CT skills, content knowledge, and self-confidence in their skills in courses that used the Instructional Unit with one of two levels of teaching intervention: light and intensive. The intervention differed in the amount of time spent in class and the level of reflection required from the students. In the light intervention, students were only given the CT rubric and their scores from the first exercise, while in the intensive intervention, students received the same and also worked in groups around the CT rubric on an additional case over a single class period, followed by individual reflection on how to improve their performance in CT. Using both interventions in the same course, but during different academic terms, we investigated whether the intensity of emphasis on CT in a course influences students’ overall CT gains.

In addition, we conducted an independent assessment of CT gains under the two interventions. At the beginning and end of each course, we administered the Critical Thinking Assessment Test (CAT), a published, validated
TABLE 1. Institution Type, Student Level, Class Size, and Term When the Instructional Unit With the Intensive (ITI) and/or Light (LTI) Teaching Intervention Was Used for Each Participating Course.

<table>
<thead>
<tr>
<th>Institution Typea</th>
<th>Course</th>
<th>Student Level</th>
<th>Class Sizeb</th>
<th>ITI</th>
<th>LTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baccalaureate College—Diverse Fields</td>
<td>Biodiversity &amp; conservation biology</td>
<td>Junior and senior</td>
<td>~10 students</td>
<td>Spring 13</td>
<td>Spring 12</td>
</tr>
<tr>
<td>Master’s College and University 1</td>
<td>Conservation biology</td>
<td>Junior and senior</td>
<td>~15 students</td>
<td>Fall 12</td>
<td>Fall 11</td>
</tr>
<tr>
<td>Master’s College and University 2</td>
<td>Conservation biology</td>
<td>Junior and senior</td>
<td>~40 students</td>
<td>Spring 12</td>
<td>Spring 13</td>
</tr>
<tr>
<td>Research University 1</td>
<td>Conservation biology</td>
<td>Junior and senior</td>
<td>~60 students</td>
<td>Fall 12</td>
<td>Fall 11</td>
</tr>
<tr>
<td>Research University 2</td>
<td>Conservation biology</td>
<td>Junior and senior</td>
<td>~15 students</td>
<td>Spring 13</td>
<td>Spring 12</td>
</tr>
</tbody>
</table>

a Following the Carnegie Classification of Institutions of Higher Education http://classifications.carnegiefoundation.org/.
b Class size = average number of students enrolled in ITI and LTI sections of the course.

instrument developed by the Center for Assessment & Improvement of Learning at Tennessee Tech University (CAIL at TTU [33]; see figure 1). The CAT is a 1-h written test consisting of 15 questions that assesses student performance in evaluation and interpretation of graphical and written information, problem solving, identifying logical fallacies or needs for information to evaluate a claim, understanding the limitations of correlational data, and developing alternative explanations for a claim. These CT dimensions were comparable to those we evaluated in our rubric, particularly those under Evidence, Influence of context and assumptions, and Conclusions.

Further, to examine whether explicit instruction in CT skills was more influential than explicit instruction in other skills, the CAT assessments were also given in the other two arms of the study that evaluated interventions designed to improve data analysis [31] and oral communication skills [30]. Unfortunately, only one instructor in the oral communication study applied the CAT instrument, so we restricted comparison to the data analysis study, where four instructors applied the CAT in their courses.

We scored batches of completed CAT tests in nine full-day scoring sessions, including only tests for which we have both a pre- and postcourse test from the same student (N = 290 total; CT study, N = 1.49; data analysis study, N = 1.41). In each session, we scored a sample of tests from across multiple institutions, study arms, and intervention levels, and each test was assigned a numerical code so that all scoring was blind. Following CAT procedures, scoring rigorously adhered to CAT scoring rubrics and was discussed by the scoring group as needed to ensure inter-scorer reliability. The CAT tests and scores were then sent to Tennessee Tech University for independent assessment, cross-validation, and analysis. For 2 of the 15 questions, the CAT scoring performed by our team was more generous than norms for these assessments performed nationally, but otherwise scores fell within those norms (results not included; analysis performed by CAIL at TTU). However, this did not affect the use of this tool as an independent assessment of CT skill gains because the scoring sessions were internally consistent and apply to both pre- and postcourse scores.

The project received an exemption from the AMNH Institutional Review Board (IRB 09-24-2010) and the Stony Brook University IRB (265333-1), and the other institutions operated under these exemptions.

Analysis
A total of 217 students from five upper-level Conservation Biology courses completed both case study exercises over one term. We excluded one student who obtained the maximum score on both exercises while using the light intervention because no improvement was possible, leaving us with N = 216 students in this study. To assess CT skills, content knowledge, and self-confidence, we calculated changes in student performance using normalized change values ($\hat{c}$) [35] and compared pre- and postassessments with paired Wilcoxon signed-rank tests [36]. The two teaching intervention groups (light and intensive) were assessed independently. Changes in the proportions of students scoring in a given quartile before and after the interventions were analyzed using $\chi^2$ tests. We tested for the effect of instructional emphasis using the light versus intensive intervention with a linear mixed-effects model. Online Appendix I has additional description of these analyses.
Because we found no differences among courses given at the different institutions, and CAT test samples were homoscedastic, a repeated-measures ANOVA was used on data pooled across institutions. This ANOVA tested overall differences across teaching interventions, across instructional units, and effects on gains for specific skills measured in the CAT. All calculations and statistical analyses were performed in R [37].

**RESULTS**

Gains in CT Skills as Measured by Performance Over the Instructional Unit

Most students gained CT skills in each term, as measured by their relative CT performance on the two case study exercises (figure 2). In terms where a light intervention was used ($N = 113$ students), $81$ students ($72\%$) gained CT skills (positive $\epsilon$ value), improving their performance, on average, by $34\%$. With the intensive intervention ($N = 103$ students), $79$ students ($77\%$) gained in skills, improving by $37\%$ (table 2).

Significant shifts in performance between the first and second case study exercises were notable in both the light and intensive teaching interventions (TI). Asterisks indicate significant differences ($p < .001$) tested with the paired Wilcoxon signed rank test. In addition, differences in the percent scores were not equally distributed across quartiles in both the light ($N = 113; X^2 = 23.415, p = .0005$) and intensive comparisons ($N = 103; X^2 = 31.893, p = .0005$), and the contingency tables indicated shifts in frequency from the bottom quartile before the intervention to the highest quartile after the intervention.

![Figure 2. Percent scores for Case Study Exercises 1 and 2 used as pre- and postassessment of critical thinking skills, respectively, under the light and intensive teaching interventions (TI). Asterisks indicate significant differences ($p < .001$) tested with the paired Wilcoxon signed rank test. In addition, differences in the percent scores were not equally distributed across quartiles in both the light ($N = 113; X^2 = 23.415, p = .0005$) and intensive comparisons ($N = 103; X^2 = 31.893, p = .0005$), and the contingency tables indicated shifts in frequency from the bottom quartile before the intervention to the highest quartile after the intervention.](http://online.ucpress.edu/cse/article-pdf/5/1/1536396/484400/cse.2021.1536396.pdf)
Under the intensive intervention, 48 students scored below the median score of 64.4%, and 55 scored equal to or above the median score in Case Study Exercise 1. Students below the median improved by an average of 44% with 90% of them showing gains, while students equal to or above the median improved their CT skills on average by 29% with 65% of them having gains (table 2).

A detailed analysis shows students improved their levels of performance in most of the four dimensions of CT defined for this study. However, achievement level varied among dimensions (figure 3). Surprisingly, for Explanation of the issues to be considered critically, students decreased their level of performance under both interventions (\(t = 154.2; p < .0025\), with Bonferroni correction). In the case of Evidence and Influence of context and assumptions, students significantly improved regardless of which intervention was used (\(t = 52.4\) and 39; \(p < .0025\); see figure 3).

Student Content Knowledge, CT Skills, and Self-Confidence
Students gained content knowledge related to the topics of both case study exercises under the light and the intensive intervention, with gains greater than 26% from pre- to postexercise (see table 3). Gains in concept knowledge associated with both case studies were greater than 35% for the light teaching intervention and similarly high for the first case study in the intensive teaching intervention group.

In addition, there was a marginally significant correlation between gains in CT skills and content knowledge (\(N = 136\) students; \(p = .61; p = .06\)). Students who showed greater gains in CT skills also showed greater gains in their content knowledge in the topic areas that were the focus of the case studies.

Based on individual self-assessment questionnaires, we found average gains in student self-confidence with CT skills of 21% regardless of intervention. Increases were statistically significant for some of the self-assessment questions, under the intensive intervention only (figure 4). Our results indicate no correlation between gains in CT skills and self-confidence (\(N = 135\) students; \(p = .049; p = .5\)).

Gains in CT Skills as Measured by the CAT Instrument
We also evaluated differences in CT gains as measured by the CAT instrument, both within the CT study arm described here, and the additional arm of the larger study focused on data analysis skills [31].

Students gained CT skills in both the light and the intensive intervention, with a significant interaction effect of teaching intervention as students had greater gains in the intensive intervention (repeated measures ANOVA: \(F(3, 147) = 4.081, p = .045\); figure 5). Significant gains were seen with the light intervention for two questions related to the ability to summarize the pattern of results in a graph without making inappropriate inferences, and the use of basic mathematical skills to help solve a real-world problem, with effect sizes of 0.28 and 0.35, respectively. Over all 15 questions, CT gains were moderate, with an effect size of 0.19 (table 4). Under the intensive CT intervention, significant gains were seen in five different questions, with effect sizes ranging from 0.32 to 0.38, and overall gains across the 15 questions were large, with an effect size of 0.49 (table 4).
FIGURE 3. Distribution of students’ performance within the four levels of proficiency for critical thinking skills (1 = lowest, 4 = highest) when using the instructional unit with the light ($N = 113$ students) and intensive ($N = 103$ students) teaching interventions. Asterisks indicate significant differences ($p < .0025$) and n.s. indicates no significant differences ($p > .0125$) between the rubric scores for Exercises 1 (preteaching intervention) and 2 (post teaching intervention), tested with the paired Wilcoxon test, Bonferroni corrected.
Students using the CT Instructional Unit showed greater increases in CAT scores than those in the data analysis arm of the study ($F_{1,290} = 9.505, p = .002$), a pattern driven by the results of the intensive teaching intervention. Students in the Intensive Teaching Intervention treatment of the CT arm of the study had significantly higher gains in CAT scores than those in the data analysis arm of the study ($F_{1,148} = 11.861, p < .001$). There was no significant difference in student CAT scores for those in the light teaching intervention of the two arms of the study ($F_{1,142} = 2.540, p = .113$).

### DISCUSSION

Our study adds to recent literature on effective approaches for teaching and learning of CT skills (e.g., [18–21]), an essential outcome of college and university education. Our pedagogical intervention hinges on the use of case studies to foster both content knowledge and CT skills, with support of an assessment rubric. We show that educators can foster measurable gains in CT over the course of a single term or semester by giving students an opportunity to practice these skills through case study exercises at least twice and reflect on their performance midway through the term, using a rubric that provided an operational definition of CT.

We chose a case study approach because real-world problem solving involves making decisions embedded in context [14, 38]. Learning how to think critically about information available in its context and evaluating evidence through identifying assumptions and gaps to arrive at strong inference is better supported through lessons presented in case studies, rather than as abstract principles alone. A key to our process was to help students identify the steps they are taking—enhancing their metacognition—through naming specific skills in formative rubrics. In this way, we specifically targeted enhancing their CT skills while gaining concept knowledge about conservation.

**Does Instructor Emphasis on CT Skill Affect the Magnitude of Individual Skill Gains?**

The light and intensive interventions used in our study differed in level of engagement with a rubric specifically designed to promote and assess CT skills—a type of formative rubric use. Rubrics are generally designed with assessment and grading in mind and developed to fit a specific assignment; however, they have great potential to help with process skill development [39]. In this study, students were given the detailed rubric after completing the first case study exercise and were encouraged to locate their performance on the rubric. In the intensive intervention, students were further tasked with using the rubric to evaluate and improve sample answers to an additional, short case study.

The formative rubric use allowed us to align assignments to the dimensions of a given skill, in line with the principles of backwards design [40] and constructive alignment [41], and to identify the components where students struggle the most, as areas to target. Our results provide support to the benefits of rubric use [39, 42]. Using a rubric to codify and operationalize a complex skill like CT seems to help both educators and learners. Our results are concordant with those of Abrami et al. [18] and Cargas et al. [19], as we show that “corrective feedback on a common rubric” aligned with relevant, authentic tasks supported learning, and that the simple act of sharing a rubric with the students may not be sufficient by itself [43]. We encourage others to make

### TABLE 3. Average Gains in Students’ Content Knowledge Measured as the Average Normalized Change ($c_{ave}$) While Using Exercises 1 and 2 of the Instructional Unit With the Light (LTI) and Intensive Teaching Interventions (ITI).

<table>
<thead>
<tr>
<th>Content Assessment</th>
<th>LTI</th>
<th>ITI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>Gains ($c_{ave}$) ± SE</td>
</tr>
<tr>
<td>Case Study Exercise 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre- versus postscores</td>
<td>52</td>
<td>0.37 ± 0.03</td>
</tr>
<tr>
<td>Case Study Exercise 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre- versus postscores</td>
<td>79</td>
<td>0.36 ± 0.04</td>
</tr>
</tbody>
</table>

Note: $p$ values are for the paired Wilcoxon signed-rank test on the percentages of the pre- and postcontent scores.
FIGURE 4. Frequency distribution of students’ self-assessed confidence levels with their critical thinking skills when using the instructional unit with the light (N = 79 students) and intensive (N = 76 students) teaching interventions. One and two asterisks indicate significant differences between pre- and postassessment scores with p < .0125 and p < .0025, respectively, and n.s. indicates no significant differences (p > .0125) tested with the paired Wilcoxon test, Bonferroni corrected.
use of available collections of rubrics, such as those generated by the VALUE initiative [15].

The rubric allowed us to provide qualitative feedback to students as they practiced—an anchor for student reflection—and to analyze gains quantitatively. Furthermore, the unit as a whole was designed to promote self-reflection, which has been shown to increase students’ ability to monitor their own selection and use of resources and evidence [44]. Self-reflection was also found to increase oral communication performance in a parallel arm of our study [30].

Finally, using case study exercises aligned to the rubric but designed to encompass topics relevant to course content, we were able to assess the learning of content while practicing CT skills. Our results support previous findings [45] that students can experience simultaneous gains in knowledge and skills, even when instructional materials and class time are dedicated to CT skill development. Indeed, we found student content knowledge gains were positively correlated with their CT skill gains, although this was marginally significant. Taken together, our results suggest that cultivation of CT skills may well enhance content knowledge.

Do Students Show Similar Responses for the Different Dimensions of CT Learning?

Formative rubric use provided insights into which dimensions of CT are more challenging to students, providing valuable feedback to educators. Our results indicate that some dimensions of CT are more challenging to improve than others. A finer examination of the CT gains shows that the changes driving our results stem from two

**TABLE 4.** Specific CT Skills Identified in CAT Questions in Which the Students in This Study Showed Significant Gains.

<table>
<thead>
<tr>
<th>CAT Question</th>
<th>Description of the Question</th>
<th>Precourse Mean</th>
<th>Postcourse Mean</th>
<th>Prob. of Difference</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensive intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>Provide alternative explanations for a pattern of results that has many possible causes.</td>
<td>1.17</td>
<td>1.54</td>
<td>p &lt; 0.01</td>
<td>+0.38</td>
</tr>
<tr>
<td>Q4</td>
<td>Identify additional information needed to evaluate a hypothesis.</td>
<td>1.47</td>
<td>1.90</td>
<td>p &lt; .05</td>
<td>+0.32</td>
</tr>
<tr>
<td>Q8</td>
<td>Determine whether an invited inference is supported by specific information.</td>
<td>0.68</td>
<td>0.83</td>
<td>p &lt; .01</td>
<td>+0.36</td>
</tr>
<tr>
<td>Q10</td>
<td>Separate relevant from irrelevant information when solving a real-world problem.</td>
<td>3.14</td>
<td>3.46</td>
<td>p &lt; .01</td>
<td>+0.38</td>
</tr>
<tr>
<td>Q14</td>
<td>Identify and explain the best solution for a real-world problem using relevant information.</td>
<td>2.24</td>
<td>2.96</td>
<td>p &lt; .01</td>
<td>+0.37</td>
</tr>
<tr>
<td>Total score</td>
<td></td>
<td>19.39</td>
<td>22.37</td>
<td>p &lt; .01</td>
<td>+0.49</td>
</tr>
<tr>
<td>Light intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>Summarize the pattern of results in a graph without making inappropriate inferences.</td>
<td>0.67</td>
<td>0.78</td>
<td>p &lt; .05</td>
<td>+0.26</td>
</tr>
<tr>
<td>Q12</td>
<td>Use basic mathematical skills to help solve a real-world problem.</td>
<td>0.81</td>
<td>0.92</td>
<td>p &lt; .01</td>
<td>+0.35</td>
</tr>
<tr>
<td>Total score</td>
<td></td>
<td>20.00</td>
<td>21.22</td>
<td>p &lt; .05</td>
<td>+0.19</td>
</tr>
</tbody>
</table>

Notes: The specific CAT question numbers are given, with a brief description of the CT skill addressed by the question. The pre- and postcourse means, probability of difference, and effect sizes are given only for those cases in which there was a significant difference observed.
dimensions in our rubric: selection and use of evidence and recognition of the influence of context and assumptions (see figure 3). Several aspects of the CT Instructional Unit are likely to have enhanced outcomes in these dimensions, such as the fact that both exercises were based on case studies where students were asked to explain how a change in context would change their course of action or conclusions. Case studies are considered valuable for science teaching because they can reflect the complexity of problems and professional practice in social-environmental systems [14, 38]. Cases present concepts and connections among them in a specific context, therefore highlighting the influence of context and assumptions, and require students to evaluate the information being presented and to select the most useful or relevant evidence for a particular task or decision. Our results support previous studies showing case studies can enhance CT skills and conceptual understanding by design [45–47].

Student performance did not significantly improve in the remaining dimensions of our CT rubric. In the case of ability to clearly and comprehensively explain the issue, overall students showed a loss (figure 3A). This dimension was unique in that students were already high achievers at the outset of the term, and the slight loss may correspond to noise along a dimension in which students were already high achievers at maximum performance levels. Alternatively, exercise structure could have played a role. The instructions for Case Study Exercises 1 and 2 were not identical in the questions relating to this dimension. Case Study Exercise 1 scores were derived from three separate questions (What problem are amphibians facing? Summarize the Climate hypothesis; Summarize the Spread hypothesis), while Case Study Exercise 2 scores rested on a single answer (Write a paragraph for your supervisors describing and explaining the problem Bright Lake is facing and why it is important to address it). Despite the former scores being averaged, having separate questions may have offered more opportunities for achievement in the first exercise and fewer in the second, resulting in an observed loss in this dimension. This was the only rubric dimension for which the number of questions contributing to a dimension’s overall score varied between case study exercises.

Finally, the most challenging dimension for students was the ability to make judgments and reach a position, drawing appropriate conclusions based on the available information, its implications, and consequences. No significant gains and the lowest rates of achievement were observed for this dimension, which maps to higher-order cognitive tasks or higher Bloom’s taxonomy levels, and has also been shown to be the most challenging for students in a broader science context [48, 49]. In a review of student writing in biology courses, Schen [49] observed that students were often adept at formulating simple arguments but showed limited ability to craft alternative explanations or to engage with other more sophisticated uses of available information. Our results mirrored this observation, as students generally only made simple use of information. Becker [50] found similar patterns in student performance and further showed that explicit instruction in constructing arguments based on evidence resulted in students developing more accurate and more sophisticated conclusions. Again, our results spotlight the importance of explicit instruction in CT. Focusing student attention on how to sift among details presented in case studies to draw inferences and conclusions and on expressing their arguments with clear connection to the evidence within case studies may be necessary steps for students to have significant gains in these more advanced aspects of CT.

Similarly, gains in CAT scores were not randomly distributed throughout questions or dimensions of CT. Students significantly improved their CAT scores in questions measuring the ability to evaluate and interpret information, think creatively, and communicate effectively. Conversely, students did not gain in their capacities to use information critically in drawing conclusions (e.g., identify additional information needed to evaluate a hypothesis, use and apply relevant information to evaluate a problem, or explain how changes in a real world problem situation might affect the solution). The results of the CAT and our case study assessments were broadly similar, with many significant gains seen in CT, except in those dimensions that required more sophisticated reasoning. Together, these results suggest that more, and perhaps different, instructional attention is needed to help students achieve certain specific dimensions of CT (see also [11]).

How Does the Intervention Affect Students at Different Achievement Levels?
Students with lower initial performance (i.e., below the median in the first exercise) gained more than those with a higher performance (above the median). These differential CT gains suggest that distinct mechanisms for
improvement may be at play. We hypothesize those students who were initially least proficient in CT were assisted by the combination of repeated practice (two case study exercises) and calling attention to the components of CT through the rubric-driven intervention, along with self-reflection. Using similar instructional activities could enhance performance or retention in science courses in general [51], given links between process skills and risk of failing introductory biology [52]. We further hypothesize that for higher achieving students, the greater emphasis on metacognition in the intensive intervention may be critical to promote gains in performance. Simply prompting students to reflect on their learning may be insufficient [53], as many students need support in implementing metacognitive strategies despite being familiar with them, such as purposeful peer interaction [54]. The combination of repeated practice and reflection through the intervention’s in-class discussion may have helped students engage more effectively with their learning.

Students showed significant gains in CAT scores under both interventions, although significantly higher under the intensive intervention. Importantly, because the students took the CAT at the end of the course, the CAT measured their response to both exercises plus the intervention, which, in the case of the intensive intervention, included practice in improving responses to a short case study exercise in alignment with the CT rubric. This contrasts with the instructional assessment, which measured gains corresponding only to the midterm teaching intervention as measured by improvement in scores for the second case study exercise. Thus, as measured by the CAT, the whole unit improved CT skills among these students over the term in both interventions, while the extensive discussion of CT skills that was part of the intensive intervention improved CT performance even further.

Advancing CT skills has proven to be challenging for many institutions. The CAT test has been used in over 250 institutions around the world [55], but few have observed gains in CT overall (see [11, 56]), although some have found an effect on individual CT dimensions [19, 57]. We consider the inclusion of case study-based exercises to be an important factor in activating student learning and fostering strong CT gains among students in our study.

The CAT assessments were also given in another arm of the overall study that evaluated interventions designed to improve data analysis skills [31], enabling us to compare CT gains when directly targeted (in the CT arm of the study) to when they were not (the data analysis arm of the study). Only students in the CT arm of the study showed notable CT gains under a light intervention, and the gains were greater under the intensive intervention in the CT arm than in the data analysis arm. The intensive intervention was designed particularly to foster student capacity to reflect on their own learning, or metacognition, as this skill has been shown to improve academic performance [53, 58]. Thus, the independent CAT assessment shows that explicit instruction in CT, coupled with repeated practice and reflection, is effective in improving student CT (see also [57]). Importantly, the CAT results imply that by developing CT in a conservation biology context, students are also enhancing their ability to apply that CT skill to other domains of their learning, such as the more general tasks required in the CAT.

Implications for Future Research and Scaling
While the results of this study are promising, our approach could be subjected to further testing. A limitation of the study was the lack of collateral data collection, such as GPA averages or overall course grades, which would have allowed for additional comparison of the student populations in each intervention. Differences in course achievement among classes, however, would not affect our interpretation of the effect of the intervention because the CT gains were observed between exercises in each term and are an internal comparison within the same student population within a given course. Our study did not use a treatment and control design or randomly assign students to the interventions. An approach based on multiple linear regression at the student level [59] could be a helpful alternative.

Adoption of the approach presented here was successful in a variety of contexts and situations. The institutions in this study varied in size and type, class size, and instructor; they included those ranked as R1, MA-granting and undergraduate only, private and public, a Minority Serving Institution, part-time and residential, and with class sizes between 10 and 60 students (see table 1 for details). Despite this variation, in 9 of the 10 classes, we observed an increase in students’ CT performance over a term, under both light and intensive interventions.

CONCLUSIONS
Our study shows educators can foster measurable gains in CT over the course of a single term or semester by giving
students an opportunity to practice at least twice and reflect midway using case study exercises aligned to both course content and a rubric that provides an operational definition of CT. Despite the brevity of the interventions, the study has provided valuable new findings on student performance in different dimensions of CT and shows promising results from instructional approaches that can be easily adapted and integrated into a variety of courses and contexts. Importantly, the study design also allowed us to work together as a team with diverse faculty in the design and application of assessment materials, which served as a professional development for faculty that can help “close the loop” between assessment and future teaching practice.

CT underpins the kind of leadership capacity needed in society today, including “ethical behavior, the ability to work with diverse populations, and the ability to think from a systems perspective” [17]. These skills are essential for conservation biology researchers and professionals because of the multidisciplinary nature of challenges comprising various forms of evidence [60], the potential for consequences to diverse stakeholders, and the high prevalence of trade-offs among alternative scenarios. Encouraged by the results of this study, we urge educators to explore these and other approaches to target CT explicitly in their learning activities and teaching practice.

AUTHOR CONTRIBUTIONS
ALP, AB, NB, and EJS developed the study framework. ALP, AB, MJG, NB, BJA, JAC, CG, MC, TT, DSF, DV, and EJS contributed to development of the instructional units. ALP, MJG, LMD, BJA, JAC, CG, DLS, MC, DSF, LF, TL, and DV implemented the CT study in their classrooms and collected data for the study. AB, LMD, and ALP performed the data analysis. ALP, MJG, and AB led the writing of the manuscript, with contributions from EJS, LMD, and NB. All authors contributed to CAT scoring sessions and to the discussions that supported writing of the manuscript. The study was made possible by an NSF grant to EJS, ALP, and NB.

ACKNOWLEDGMENTS
We are grateful to all study participants, those who helped score the CAT, and K. Douglas and N. Gazit for key assistance. We thank G. Bowser, A. Gómez, S. Hoskins, K. Landrigan, D. Roon, and J. Singleton for their contributions to the initial design and the original authors of the NCEP materials adapted for this study. The Biology Education Research Group at UW provided helpful input in initial discussions.

COMPETING INTERESTS
The authors have declared that no competing interests exist. Martha J. Groom is an editor at Case Studies in the Environment. She was not involved in the review of this manuscript.

FUNDING
This project was supported by the National Science Foundation (NSF) CCLI/TUES Program (DUE-0942789). Opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect NSF views.

SUPPLEMENTAL MATERIAL
Appendix 1. Supplementary Information on Methods.

REFERENCES
9. Momsen JL, Long TM, Wyse SA et al. Just the facts? Introductory undergraduate biology courses focus on low-


