JOURNAL OF RESEARCH IN SCIENCE TEACHING

Translating Expertise Into Effective Instruction: The Impacts of Cognitive Task Analysis (CTA) on Lab Report Quality and Student Retention in the Biological Sciences

David F. Feldon,[1] Briana Crotwell Timmerman,[2,3] Kirk A. Stowe,[2] Richard Showman[2]

[1]*Department of Curriculum, Instruction, and Special Education, University of Virginia, Virginia*
[2]*Department of Biological Sciences, University of South Carolina, Columbia, South Carolina*
[3]*SC Honors College, University of South Carolina, Columbia, South Carolina 29208*

Received 22 August 2009; Accepted 22 February 2010

Abstract: Poor instruction has been cited as a primary cause of attrition from STEM majors and a major obstacle to learning for those who stay [Seymour and Hewitt [1997]. Talking about leaving: Why undergraduates leave the sciences. Boulder, CO: Westview]. Using a double-blind design, this study tests the hypothesis that the lack of explicit instructions in scientific inquiry skills is a major factor in both low STEM retention and academic underperformance. This project delivered supplemental instruction to students in a laboratory-based undergraduate biology course ($n = 314$) that was derived either from cognitive task analyses (CTAs) conducted with expert biologists (treatment) or was authored and delivered by an award-winning biology instructor (control). Students receiving traditional instruction were almost six times more likely to withdraw from the course than students in the treatment condition (8.1% vs. 1.4% of initial enrollment). Of the students who completed the course, those who received the CTA-based instruction demonstrated significantly higher levels of performance in the discussion section of their written laboratory reports. Significantly higher performances were seen specifically in the areas of analyzing data to formulate valid conclusions, considering alternative explanations, consideration for the limitations of the experimental design and implications of the research.

Keywords: cognitive task analysis; direct instruction; inquiry skills; undergraduate education

Nationally, undergraduate majors in the biological sciences have high rates of attrition. The average dropout rate is approximately 50% (Astin & Astin, 1993; Seymour, 2001). Although a broad range of factors influence student retention in the sciences, one that is almost universally cited by exiting students is poor instruction. In a national study of college freshmen who initially declared STEM (Science, Technology, Engineering, Mathematics) majors, 90% of students who switched to non-science majors cited ineffective instruction as a primary reason. Of those who successfully completed degrees in STEM programs, 74% indicated that poor instruction was a major problem (Seymour & Hewitt, 1997). In addition, 55% of students report that they fail to see how the concepts and skills they are taught apply to the problems that they are asked to solve (Seymour, 2001).

Classes in which students learn research methods and scientific inquiry skills (defined here as the logical tools and frameworks used by scientists to collect, analyze, and interpret data within the context of scholarly discourse within their disciplines[1]) are especially problematic for many students. Such courses typically yield great variation in students' skill mastery and self-efficacy at the postsecondary level (Bianchini, Whitney, Breton, & Hilton-Brown, 2001; Onwuegbuzie, Slate, Paterson, Watson, & Schwartz, 2000). Such courses typically present their content through assigned readings and instructors' lectures. In each case, the explanation of how to do research is often dependent upon the ''armchair reflections'' of a researcher

Contract grant sponsor: National Science Foundation; Contract grant number: 0653160 and 0410992.
Correspondence to: B.C. Timmerman; E-mail: timmerman@schc.sc.edu
DOI 10.1002/tea.20382
Published online in Wiley InterScience (www.interscience.wiley.com).

describing his or her own problem-solving practices (Schunn & Anderson, 2001, p. 87) or outmoded linear descriptions of the ''scientific method'' (Hodson, 1996; Lederman, 1998). In other instances, messages about the nature and mechanisms of inquiry are left mostly or entirely implicit (e.g., Delamont & Atkinson, 2001; Leckie, 1996; Palmquist & Finley, 1997). However, even when the instructor provides personalized descriptions of authentic research procedures, research on knowledge elicitation and STEM experts' cognition indicates that self-reports of problem-solving processes are usually incomplete or inaccurate (Chao & Salvendy, 1994; Cooke & Breedin, 1994; Dunbar, 2000; Feldon, 2007, in press).

These inaccuracies occur chiefly through two mechanisms that stem from the development of expert cognition. First, experts develop very sophisticated and extensive schemas in their domains of expertise, which serve as templates for rapidly organizing relevant information (Cooke, 1992). As a result, they are able to assess and approach problems with a greater degree of complexity than a non-expert could manage. These schemas efficiently support information encoding by guiding attention and governing the interpretation of events (Gobet, 1998), but they can do so at the expense of the ability to ''unpack'' the schema and verbalize the relationships between elements with a high level of specificity (Rikers, Schmidt, & Boshuizen, 2000). Robust schemas can also inhibit the recollection of information that does not readily fit the templates that they provide (Wigboldus, Dijksterhuis, & van Knippenberg, 2003). Consequently, if an expert holds an a priori causal theory regarding either the mechanisms of his own problem-solving processes or the nature of the phenomena he is researching, information which conforms to that theory will be attended to while discrepant information is likely to be disregarded or misremembered in a manner similar to the confirmation bias effect (Chinn & Brewer, 1993; Koslowski & Maqueda, 1993). This is true not only of the theoretical frameworks that govern research efforts, but also of scientists' ideas about how they perform their work. Therefore, explanations are highly likely to be based on the generic properties of their schemas rather than the actual basis of their problem solving in a specific instance (cf. Nisbett & Wilson, 1977).

The second element of expert cognition that leads to the inaccuracy of self-report is automaticity (Blessing & Anderson, 1996; Feldon, 2007). Automaticity results from the extensive practice of specific skills and decision-making processes. As individuals acquire expertise, they require significantly less conscious monitoring of the procedures they employ to solve problems in their field. When a given skill has been applied often enough, it can be deployed without conscious effort. Goals and strategy selection can also occur without conscious intent (Aarts & Dijksterhuis, 2000; Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trotschel, 2001). This ''step-skipping behavior'' (Koedinger & Anderson, 1990, p. 511) of experts leads to omissions in the articulation of their problem-solving processes, because they do not consciously select or engage the most familiar aspects of their respective approaches. Consequently, the most frequently employed elements—presumably those of greatest utility within a domain of expertise—are the most difficult to articulate through recall and the least likely to be included in instruction (Feldon, 2007, in press).

Examples of incorrect and incomplete explanations from scientists occur both in laboratory experiments and in authentic environments. Cooke and Breedin (1994) asked expert physicists to predict the trajectories of a variety of thrown objects and provide written explanations of their reasoning processes. However, when the written explanations were used as instructions to replicate the predictions, the results were completely uncorrelated with the experts' original estimates, suggesting that they unintentionally fabricated aspects of their self-reported reasoning processes. Similarly, Dunbar (2000) found extensive omissions by microbiologists in his study of eight prestigious laboratories when describing the reasoning processes that led to important discoveries and solutions to research problems. He recorded and transcribed the weekly meetings in which the results of experiments were discussed and later asked the participants in the meetings to recount how specific conclusions were reached. The participants (professors, postdoctoral fellows, and graduate students) were unable to recall accurately the strategies, collective reasoning processes, or relevant analogies that led to either solutions to mundane problems or groundbreaking insights in the field. ''It appears that the scientists remember the results of their reasoning rather than the small steps that they made [to produce them]'' (Dunbar, 2000, p. 55).

Cognitive Task Analysis

Cognitive task analysis (CTA) is emerging as a valuable tool to overcome this challenge (Clark & Estes, 1996; Schraagen, Chipman, & Shute, 2000). CTA encompasses a range of knowledge elicitation techniques

to help experts accurately and completely articulate their problem-solving processes. Empirical studies suggest that CTA can enhance the completeness of information obtained between 12% (Chao & Salvendy, 1994) and 43% (Crandall & Getchell-Reiter, 1993; see also Clark & Estes, 1996) when compared with unguided knowledge elicitation (i.e., think aloud, free recall, or reflection). When the CTA-elicited knowledge is used as the content basis for instruction, it often yields very large gains over ''current best practice'' comparison conditions (see Clark, Feldon, van Merriënboer, Yates, & Early, 2008 and Feldon, 2007 for reviews of these studies), but CTA has yet to be significantly incorporated into instructional design in academic settings.

The purpose of this study is to test the effectiveness of CTA-based instruction for teaching undergraduate biology students to engage in the scientific process (i.e., conduct scientific observations, frame research questions, generate testable hypotheses, design experiments, and interpret results). In the past, CTA has been applied to instruction in specialized fields that typically deliver training outside the standard university class format. Examples include the training of medical residents in surgical procedures, military officers in radar system troubleshooting, and patent clerks in the process of validating patent applications. Here we report the first application of CTA-based instruction in a traditional university context and the first to address scientific skills.

There are many types of CTA performed by various practitioners. However, most follow a five-stage process (Clark et al., 2008). First, practitioners collect preliminary knowledge to orient them to the general parameters of the target task. Second, they identify major subtasks and the types of knowledge necessary to perform them. Third, they apply focused knowledge elicitation methods that can include combinations of interview techniques, direct observations, and simulations. Multiple experts are generally used as informants in this process, because they are unlikely to each omit the same pieces of information. By synthesizing their separate accounts, a more complete picture of experts' cognitive processes can emerge (Chao & Salvendy, 1994; Lee & Reigeluth, 2003). Fourth, they analyze and verify the data acquired from the elicitation. Lastly, they format the results for the intended application.

Frequently, the formatted outcome consists of a set of action steps and decision rules that represent a viable way to solve a class of authentic problems. The resulting protocol articulates the effective steps and branching decision points that experts navigate during their problem-solving processes, including the relevant cues that guide these decisions, even when experts are not fully aware of which cues influence their decisions.

CTA-based training systems that have explicitly accommodated the tacit nature of experts' knowledge have proven to be significantly more effective than those that have not (e.g., Merrill, 2002; Schaafstal, Schraagen, & van Berlo, 2000; Velmahos et al., 2004). Lee's (2003; see also Clark et al., 2008) meta-analysis found a mean effect size of $d = 1.72$ ($d > 0.8$ is considered large; Cohen, 1988) for CTA-based instructional interventions. Further, there is substantial evidence that gaps in instructional content resulting from the omission of necessary steps in problem-solving procedures induce higher levels of cognitive load in learners, which interferes with learning and can lower motivation (Britt, 2005; Chandler & Sweller, 1991; Kirschner, Sweller, & Clark, 2006; Paas, Tuovinen, van Merriënboer, & Darabi, 2005; Sweller, Chandler, Tierney, & Cooper, 1990; Tuovinen & Sweller, 1999).

Biology as an Ill-Structured Domain

One of the key differences between the domains in which CTA has been used for training and the current application is that the above procedures have definitive outcomes that can be discussed in terms of success and failure. Surgeries can result in patients' recovery, radar systems can function accurately, and the originality and legality of patents can be independently verified. However, scientific endeavors do not have definitive outcomes that can be easily linked to the processes employed. There are nearly an infinite number of ways to approach the study of a scientific problem that can yield valid and productive scientific knowledge.

With highly structured domains, problems can be solved to a large extent using skills that are learned and applied algorithmically. However, scientific problems do not have single correct solutions or standardized processes and each one inherently requires the development of a new solution or configuration to fit the problem's unique parameters. In other words, scientific inquiry is an ill-structured domain that ''possess[es] multiple solutions, solution paths, fewer parameters which are less manipulable, and contain[s] uncertainty

about which concepts, rules, and principles are necessary for the solution or how they are organized and which solution is best'' (Jonassen, 1997, p. 65). This poses a significant instructional problem, because there is little consensus about the most effective means to successfully promote the development of problem-solving skills for domains of this kind (Anderson, Reder & Simon, 1997; Barnett & Ceci, 2002; Bransford & Schwartz, 1999).

The major debate centers on identifying the optimal amount, type, and timing of instructional guidance provided to students to best prepare them to succeed in complex, ill-structured domains (Clark, 2009). One end of the spectrum advocates maximizing the amount of information provided to students prior to problem-solving opportunities (Lederman, 1998). This provides learners with effective problem-solving procedures and opportunities to practice them in increasingly complex ways (e.g., Eilam, 2002; Klahr & Nigam, 2004). Therefore, this type of instruction tends to emphasize procedural knowledge that is explicitly conveyed through the explanation of worked examples followed by opportunities to practice applying the skills to a variety of problems that gradually increase in complexity and/or authenticity (Renkl & Atkinson, 2003).

In contrast, other approaches to instruction provide learners with greater opportunities to discover for themselves the strategies for solving domain problems through problem-based learning or case-based approaches. Information is then disseminated through scaffolding during problem-solving efforts (e.g., Hmelo-Silver, Duncan, & Chinn, 2007; Sadeh & Zion, 2009). In this way, learners will not simply ''execute . . . procedure as dictated—but rather . . . engage in scientific problem solving'' where they are able to practice generating their own solution strategies (Savery & Duffy, 2001, p. 4; see also Roth, 1994). Consequently, theoretical knowledge about the domain is taught prior to problem-solving activities, but specific procedures are not (Schwartz & Martin, 2004).

Although in a general sense CTA-derived content can be incorporated into either instructional strategy (see Feldon & Stowe, 2009 for a discussion of this relationship), it is typically used to provide an explicit procedural approach to problem solving at the outset of instruction that learners are able to apply during problem-solving efforts. Thus, resulting instruction may bear similarity to other highly explicit instructional designs (e.g., Eilam, 2002). However, the content taught is derived directly from the knowledge and strategies of practicing experts rather than through a rational analysis of idealized procedures. To address the impact of CTA-based explanations on undergraduate biology students (treatment) compared to traditional explanations by an award-winning professor (control), the following hypotheses were tested:

(1) CTA-based instruction leads to increased performance in scientific problem solving as measured by the quality of biology laboratory reports.
(2) CTA-based instruction reduces the rate of attrition in an introductory level biology course.

Methods

Course and Student Population

In this study, we examined the effects of CTA-based scientific problem-solving instruction on students' retention and performance in BIOL 101—Biological Principles I. This 1-semester course consists of 3 lecture hours and 3 laboratory hours per week providing an introductory survey of macromolecules, cell structure and function, genetics, and molecular biology. The course primarily serves biology and allied health majors, so the students (mostly freshmen) typically perceive the course material as relevant and necessary for their future goals. Data were collected from all enrolled students in the Spring semester of 2008 ($n = 314$). All participants were blind to experimental conditions and to the existence of the study. Participants did not need to provide consent for data collection, because the study was granted exempt status by the university's institutional review board. Its activities occur as part of normal educational practice using instruments typical of the university classroom environment.

To ensure that the treatment and control populations were equivalent in both, general scientific reasoning ability and motivation Lawson's Test of Scientific Reasoning (Lawson, 1978, 2000) and the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, García, & McKeachie, 1991, 1993) were both administered at the beginning of the course. Lawson's Test for Scientific Reasoning (Lawson, 1978, 2000) is a 24 multiple-choice items assesses participants' abilities to distinguish between discrete sources of variance,

use proportional logic, apply combinational reasoning, and interpret correlations. Satisfactory performance on test items is dependent on participants' abilities to draw correct deductive inferences from presented data and evaluate the effectiveness of strategies to control variables in presented scenarios. The instrument has been validated with high reliability for undergraduates taking science classes (Lawson et al., 2000). The MSLQ is an 81-item, Likert self-report instrument with strong reliability and validity (alpha ≥ 0.90 for undergraduate populations; Spitzer, 2000). The motivation items assess students' goals and value beliefs for a course, their beliefs about their skills to succeed, and their anxiety about tests in a course. The learning strategy items assess students' use of various cognitive and metacognitive strategies.

Neither measure found significant differences between treatment and control samples. Mean Lawson's scores for the control and treatment conditions were 14.61 ($SD = 4.69$) and 14.88 ($SD = 4.38$) out of 24, respectively ($p = 0.614$, ns). Mean scores for MSLQ subscales used were as follows: task value (control: 23.77 [$SD = 6.39$]; treatment: 23.95 [$SD = 6.07$]; $p = 0.814$, ns), self-efficacy (control: 30.46 [$SD = 7.84$]; treatment: 23.95 [$SD = 6.07$]; $p = 0.643$, ns), metacognition and self-regulation (control: 38.74 [$SD = 8.40$]; treatment: 39.97 [$SD = 7.67$]; $p = 0.198$, ns), study environment (control: 29.41 [$SD = 5.71$]; treatment: 29.07 [$SD = 6.16$]; $p = 0.631$, ns), effort regulation (control: 15.46 [$SD = 3.34$]; treatment: 15.32 [$SD = 3.51$]; $p = 0.716$, ns), and peer learning (control: 7.50 [$SD = 2.88$]; treatment: 7.57 [$SD = 3.02$]; $p = 0.841$, ns).

Experimental Design

The double-blind, random-assignment design of this pre-post, two group quasi-experimental study (Campbell & Stanley, 1966) addressed two very important analytical concerns that often confound research on instruction. With the use of double-blinding, neither participants, instructors, nor researchers knew which students were from the control or treatment groups. This averted both potential experimenter effects and the Hawthorne effect (Rosenthal, 1966). We also randomly assigned laboratory sections to either a treatment or control condition. Laboratory sections were facilitated by graduate teaching assistants (TAs) who were assigned to a single condition for the semester. Neither TAs nor their students were aware of the existence of the study.

The instructional content for the scientific process in biology was delivered via Internet as a series of streaming videos that were required viewing for students in the course. Students viewed them independently before each weekly laboratory session and a second time at the beginning of class with the TA, which was followed by a brief discussion of the content. Each successive video became available for viewing at the beginning of the week for which it was assigned. Once available, students were free to view it as many times as they wished for the duration of the course. Each TA was provided only the condition-appropriate set of videos for their instructional preparation. Every reasonable effort was made to ensure that the videos were as similar as possible except for the actual verbiage of the content being presented. The faculty member being taped even wore the same sweater for every taping session for both treatment and control in order to increase the consistency of presentation for the videos.

Students' viewing compliance was monitored via server logs and participation points were awarded for viewing the videos outside of class each week. Compliance rates ranged from 69% to 93% of students each week and on average 87% of students in the control condition and 82% of students in the treatment condition watched the videos on their own time. Compliance was consistently higher in the control condition each week with only the Observation I video being viewed more in the treatment condition (93% compared to 92%). As students watched the videos in laboratory each week as well, these viewings represent reinforcement of the material. Thus, there do not appear to be any meaningful distinctions between viewing rates in treatment and control and if such differences exist, they favor the control condition.

Cognitive Task Analysis

To conduct the CTA, three experts were recruited from the university biology faculty on the basis of the following criteria: each had been engaged in biological research for at least 10 years, had published articles in top-tier academic journals, and was acknowledged by peers as being highly skilled in the scientific process (cf. Ericsson & Charness, 1994). Each was interviewed for approximately 2 hours regarding the way in which he or she approached the scientific process, beginning with conducting observations and the formulation of a

research question. As the interviews progressed, attention was focused on the identification of decision points and cueing events that led to the selection of a specific strategy relevant to the problem-solving process. Frequently, the interviewer would attempt to reframe the information provided to generate a test case (e.g., ''So, if the situation were like this, you would do *X* and not *Y*. Is that correct?''). If the expert felt that the characterization was not fully accurate, she would articulate the way in which she would address the situation, and the interviewer and the expert would work to construct a refined decision rule that incorporated the appropriate differentiated elements.

Once the interviews were completed, the transcripts were analyzed to develop an abstract representation of the process reported by each individual. These representations took the form of outlines that incorporated actions and branching decision points (see Appendix 1, for examples, in final protocol). Each expert reviewed his or her own protocol to offer any necessary revisions or additions. The three reviewed protocols were then synthesized by the interviewer to represent a single, aggregate approach. This final protocol was then distributed to all three experts for an additional round of editing according to the following instructions:

> Please review this protocol and provide feedback. Because this is an aggregation of three people's procedures, do not worry if it does not represent precisely the way that you, yourself, engage in the scientific process. Instead, please decide if the procedure would generate appropriate results for a student trying to implement it. If you find a step that is missing, please add it. If you find a step that you believe is inappropriate or unnecessary, please note that as well.

Suggested edits were incorporated, and the resulting protocol was redistributed for final approval by the experts.

Creation of the Videos

The video-based instruction consisted of a tenured associate professor with multiple awards for outstanding teaching[2] delivering either traditional, lecture-based instruction in these skills (i.e., best practices based on self-report and theoretical knowledge) or lectures scripted from the full set of decision rules generated through the CTA process. All control condition videos were created prior to providing the instructor with the CTA-based scripts to prevent inadvertent contamination of the ''current best practice'' videos.

Topics for the videos were determined in advance based on the laboratory curriculum goals and focused on the major steps of scientific inquiry (i.e., observation, research question identification, hypothesis generation, experimental design, and data analysis). Eight treatment–control pairs of videos were made. Two condition-neutral videos (the same video was used for treatment and control) were also made as placeholders to maintain the students' viewing habit each week. Each video was 5–10 minutes long. Content analysis of treatment and control videos indicated that the primary differences in the explanations provided related to the level of specificity and detail in the statements made and whether content was framed procedurally or conceptually. The treatment condition (CTA-based) provided more specific and detailed statements and was framed as a set of step-by-step actions and decisions to be made. In contrast, the explanations provided in the control videos were found to be more abstract and presented as principles illustrated with examples (Feldon & Stowe, 2009). For example, in explaining how to generate an experimental hypothesis from observations, the traditional instruction included statements such as:

> As we design experiments to test the hypothesis, all of the data, the results, must be judged against the hypothesis.
>
> A hypothesis is something which is developed carefully . . . it provides the core or base around which you can design an experiment.
>
> It's a guess . . . that is based upon the ideas that you have accumulated, whether from previous data, from reading . . . it's not based upon simply an idea that you have. A working hypothesis as we normally use it in the laboratory represents our first approximation . . .
>
> In the world of science, it is absolutely essential that every hypothesis: (a) be testable . . .; (b) that the individual testing the hypothesis be willing to change the hypothesis at any time if the data are inconsistent with the working hypothesis.

While the CTA-based instruction said "write down . . . relationship you expect to see between the factor(s) you included in your research question." See Appendix 2 for detailed comparisons.

Students' Scientific Investigations

Using the videos and other laboratory exercises, students engaged in portions of the scientific process throughout the semester (e.g., learning how to develop testable hypotheses, isolate variables, analyze results, and draw conclusions). The laboratory portion of the course then culminated in a multiple-week, inquiry-based investigation of *Drosophila melanogaster* (fruit fly) genetics wherein students were required to make observations, generate hypotheses, collect, analyze and interpret data and form conclusions based on those data. The work product submitted for course credit was a formal paper, written in scientific format, reporting their findings. Due to logistical constraints, all students investigated the same unknown genetic cross. Their task was to determine the genotypes of the parental generation and if the alleles exhibited Mendelian inheritance patterns. Investigations were conducted in small groups within a laboratory section. Data were then pooled within that laboratory section to increase sample size. Thus, students were provided with the research question, hypothesis, and methods, but they had complete discretion in the scientific judgments articulated in their papers (i.e., discussions of intellectual context, study rationale, data analysis and interpretation, conclusions, and limitations). Lab reports were written and submitted individually by students via an online course management website. All papers were checked for plagiarism using SafeAssign™ and papers containing plagiarized material were not included in the sample.

Measures of Student Performance

Student performance was measured using the Universal Lab Rubric (Timmerman, Strickland, Johnson, & Payne, in press), a rubric specifically developed for evaluating written laboratory reports in biology. The criteria are derived from common perceptions of what constitutes effective scientific reasoning and writing in the science education literature (Timmerman, 2008), as well as criteria espoused by scientific journals (e.g., Cicchetti, 1991; Marsh & Bell, 1981). The rubric has been psychometrically validated across multiple biology courses (including previous semesters of the BIOL 101 course used in this study) and assignments including an earlier version of the *Drosophila* lab used in this study. Overall reliability of the rubric was calculated using generalizability analysis and found to be high ($g = 0.85$). Laboratory reports were rated by a pool of three raters, two of whom had prior experience rating biology papers with this rubric and all of whom had appropriate biological content backgrounds. Pairwise inter-rater reliability for this sample ranged from $g = 0.70$ to 0.86.

Because of students' limited control over the methodology for the experiment, analysis of student performance focuses on the Introduction, Results, and Discussion Sections. Only complete papers submitted by the assignment deadline were analyzed for the purposes of this study. Papers excluded from analysis for these reasons were randomly distributed across all course sections. Reflecting both exclusions and course attrition, the final sample used for analysis was $n = 252$ ($n = 119$ for the treatment condition; $n = 133$ for the control condition).

Measures of Student Attrition

Student attrition was measured by the university registrar's office. It is normal practice for students to frequently add and drop lab sections during the first week of the semester as they arrange their schedules. These students were not included in the attrition calculations. Attrition was measured by identifying those students who had committed to the course initially but then withdrew at some later point in the semester as indicted by a notation on their transcripts.

Results and Discussion

Using a double-blind design, this study examines the effects of CTA-based instruction in scientific inquiry skills on undergraduate students in a lab-based biology course. Specifically, it hypothesized that recipients of CTA-based instruction would perform better on written laboratory reports of an experiment and that course attrition would be lower for students in the CTA condition.

Laboratory Report Performance

In accordance with the first hypothesis, student performance favors the treatment (CTA) condition in the Discussion Section of the written laboratory reports. The lack of significant differences in other sections (i.e., Hypotheses, Methods, Data Selection, Data Presentation, and Statistical Analysis) is expected due to the common methodology dictated by the instructor and used across all laboratory sections. Table 1 shows the ANOVA of performance outcomes by individual rubric item and rubric section index score. Effect sizes calculated for significant differences using Cohen's d were in the small to medium range (0.232–0.391) (Cohen, 1988).[3] Effect sizes for each element in the Discussion Section of the rubric are as follows: conclusions based on selected data, $d = 0.265$; alternative explanations, $d = 0.311$; limitations, $d = 0.266$; significance, $d = 0.232$; and Discussion index score, $d = 0.391$.

Much of the CTA focused on the need to be explicit in predicting expected outcomes, and this constant reference to making connections between data and conclusions likely contributed to this effect. For example, the CTA instructions include details such as:

> If your hypothesis does not state your expected outcomes for every condition of your experiment, then you will need to list explicitly in your lab notebook what you expected the dependent variable to do for each combination of factorial levels that you will test.

The CTA-based instructions repeatedly reinforced the concept that predictions and conclusions should be made explicit and written down in the lab notebook. Engaging in such behaviors would make it more obvious to students when anomalies or discrepancies arise between data and a theory or model. Thus, their conclusions are more likely to be firmly grounded in their data. Recognition of anomalies and re-assessment of prior knowledge or misconceptions are critical components of conceptual change and knowledge acquisition (Chinn & Brewer, 1993; Posner, Strike, Hewson, & Gertzog, 1982). Instruction in the treatment condition frequently required students to articulate and compare new information with prior ideas, which likely stimulated significant reflection and metacognition. Reflection, especially through writing, has been shown to improve scientific reasoning and knowledge generation (Keys, 2000).

Given the notable level of instruction on how to frame a scientific question and the role of primary literature in framing a question and interpreting results, some might view the lack of significant differences between treatment and control in the Introduction or Use of Primary Literature sections as surprising. However, the rubric was originally designed to assess change over the span of an entire undergraduate career (Timmerman, 2008). Post hoc review of the rubric suggests that the scales for these items are insufficiently sensitive to pick up differences in these aspects of student writing at this early stage in their scholarly development. These items are currently undergoing further revision to improve their sensitivity to initial stages of student development. As there was no instruction geared toward writing style or mechanics, the lack of significant differences in overall writing quality likely reflects an equivalent preexisting level of writing ability among participants. It should also be noted that the Writing Quality criterion was effectively a holistic assessment reflected in a single criterion. Holistic rubrics typically generate far greater variation in scores between raters than do analytic rubrics (Klein et al., 1998) as many factors are compressed into a single score. Post hoc analysis of our inter-rater reliability data found this criterion to be the most variable.

The only significant unexpected finding was the higher mean in the control condition for the ''testable and consider alternatives'' element of the Hypotheses section. It is possible that the emphasis in the CTA on generation of research ideas from the literature and observations caused students in the treatment condition to attempt more innovative and meaningful hypotheses than were possible in a situation where the genetic cross was pre-selected. In contrast, students in the control condition may not have been searching as deeply for their hypotheses and constructed more straightforward and testable hypotheses.

Course Attrition

Course enrollment records indicate that participants in the control condition were almost six times as likely to withdraw from the course (8.1% of enrolled students) as students in the treatment condition (1.4% of enrolled students). Chi-square tests indicated significant differences between treatment and control sections

Table 1
Student performance in written laboratory reports by experimental condition and rubric element

Universal Laboratory Rubric Criteria and Definitions (Timmerman et al., in press)	Treatment, Mean (*SD*) ($n = 119$)	Control, Mean (*SD*) ($n = 133$)	*F*	*p*-Value (2-Tailed)
Introduction				
Context: Demonstrates a clear understanding of the big picture; Why is this question important/interesting in the field of biology?	1.37 (0.59)	1.34 (0.59)	0.150	0.699
Accuracy and relevance: Information is accurate, relevant and provides appropriate background for reader including defining critical terms	1.19 (0.60)	1.29 (0.59)	1.675	0.197
Introduction: index score	*2.56 (1.09)*	*2.63 (1.07)*	*0.252*	*0.616*
Hypothesis quality				
Testable and consider alternatives: Hypotheses are clearly stated, testable and consider plausible alternative explanations	0.74 (0.33)	0.83 (0.33)	4.500	0.035*
Scientific merit: Hypotheses have scientific merit	0.74 (0.35)	0.77 (0.35)	0.229	0.633
Hypotheses: index score	*1.49 (0.66)*	*1.60 (0.63)*	*1.826*	*0.178*
Methods				
Controls and replication; Appropriate controls (including appropriate replication) are present and explained.	0.47 (0.42)	0.49 (0.48)	0.235	0.628
Experimental design: Experimental design is likely to produce salient and fruitful results (tests the hypotheses posed)	0.72 (0.37)	0.70 (0.42)	0.168	0.682
Methods: index score	*1.19 (0.64)*	*1.19 (0.73)*	*0.006*	*0.937*
Results				
Data selection: Data are comprehensive, accurate and relevant	0.89 (0.35)	0.92 (0.38)	0.482	0.488
Data presentation: Data are summarized in a logical format. Table or graph types are appropriate. Data are properly labeled including units. Graph axes are appropriately labeled and scaled and captions are informative and complete	1.38 (0.74)	1.47 (1.17)	0.521	0.471
Statistical Analysis: Statistical analysis is appropriate for hypotheses tested and appears correctly performed and interpreted with relevant values reported and explained	1.09 (0.56)	1.05 (0.54)	0.487	0.486
Results: index score	*3.35(1.36)*	*3.42(1.61)*	*0.247*	*0.620*
Discussion				
Conclusions based on data selected: Conclusion is clearly and logically drawn from data provided. A logical chain of reasoning from hypothesis to data to conclusions is clearly and persuasively explained. Conflicting data, if present, are adequately addressed	0.90 (0.50)	0.77 (0.48)	4.378	0.037*
Alternative explanations: Alternative explanations are considered and clearly eliminated by data in a persuasive discussion	0.43 (0.52)	0.28 (0.44)	6.171	0.014*
Limitations: Limitations of the data and/or experimental design and corresponding implications discussed	0.70 (0.63)	0.54 (0.57)	4.703	0.031*
Significance: Paper gives a clear indication of the implications and direction of the research in the future	0.31 (0.46)	0.21 (0.40)	3.463	0.064^
Discussion: index score	*2.34 (1.49)*	*1.78 (1.37)*	*9.501*	*0.002***
Use of Primary Literature: Reasonably complete discussion of how research project relates to other relevant work	0.70 (0.42)	0.74 (0.38)	0.666	0.415
Writing quality: Grammar, word usage and organization facilitate the reader's understanding of the paper	1.22 (0.49)	1.19 (0.52)	0.231	0.631
Total rubric score	*12.85 (4.08)*	*12.56 (3.9)*	*0.322*	*0.571*

*Significant at $p < 0.05$.

**Significant at $p < 0.01$.

^Significant at $p < 0.05$ using one-tailed test.

Table 2
Course attrition by condition, gender, and biology major status

	Initial Enrollment	Number of Students Who Withdrew	Fisher's Exact Test (1-Tailed), *p*-Value
Men (T)	52	1	0.072
Men (C)	57	6	
Women (T)	90	1	0.041*
Women (C)	115	8	
Biology majors (T)	60	1	0.334
Biology majors (C)	64	3	
Allied health majors (T)	82	1	0.010**
Allied health majors (C)	108	11	
Overall (T)	142	2	0.005**
Overall (C)	172	14	

Note. Initial enrollment numbers were recorded during the second week of the semester. Number of students who withdrew was recorded at the end of the semester based on transcript grades. T, Treatment (CTA-based instruction); C, Control (traditional instruction).
*Significant at $p < 0.05$.
**Significant at $p \leq 0.01$.

overall ($p = 0.005$; Cramér's $V = 0.152$), with subpopulations consistently reflecting higher attrition in the non-CTA condition (Table 2).

Effect of the CTA-based instruction was significant for women ($p = 0.041$; Cramér's $V = 0.142$) and non-majors ($p = 0.010$; Cramér's $V = 0.183$). The relatively small number of students withdrawing from both conditions likely limited significance and measurable effect size, but the trends for each subset of participants indicated a higher rate of attrition in the control condition. The impact on declared biology majors (three withdrawals in the control, one withdrawal in the treatment) is notable, though non-significant. Future data collection may shed greater light on possible differential effects among majors. With the current data, the association between exposure to CTA-based instruction and student retention reflects a small effect size using Cramér's V ($0.10 < V < 0.30$; Cohen, 1988).

Conclusions

The value of higher education lies in the subject matter expertise of its professors. It seems, however, that STEM faculty may unintentionally fail to communicate this knowledge effectively. Even when instruction is engaging and student-centered (e.g., problem-based, inquiry-oriented), failure to provide the information necessary to solve problems successfully, can impede students' academic progress (Jonassen, Tessmer, & Hannum, 1999; Kirschner et al., 2006).

Findings from this study suggest that CTA-based procedural instruction offers notable benefits over traditional methods of generating instructions on how to engage in scientific inquiry, such as personal reflections of the instructor or the textbook descriptions of the ''scientific method.'' The combination of performance and attrition data is consistent with our hypotheses and the predictions of cognitive load theory. It is important to remember that these results were obtained at a single institution during a single semester, so the generalizability of our findings is somewhat limited. Within that context, CTA appears to provide two specific benefits. The first is that the instructions provided to students are more complete (fewer steps, criteria or decision points are likely to be omitted). Second is that the explicit nature of the instructions generated by CTA provides a level of precision and detail that is otherwise unavailable to students, so they likely have lower levels of extraneous cognitive load and fewer knowledge gaps. The decrease in cognitive load potentially leads to fewer instances of burnout, because sustained task demands are less likely to exceed working memory capacity for students receiving CTA-based instruction (Clark, 1999).

It is interesting to note that the significant differences between conditions for the performance on the final scientifically formatted paper are associated with the most analysis-intensive facets of the scientific writing process—interpreting findings, articulating the underlying chain of logic, identifying limitations, and providing alternative explanations for results. Because abstract scientific reasoning processes are typically the most difficult for experts to self-report accurately (Dunbar, 2000; Feldon, in press), CTA may be a

particularly fruitful educational innovation for instructors concerned with developing students' problem-solving and scientific reasoning skills.

On a broader level, our findings raise important questions about instruction in ill-structured domains. It is commonly argued that providing highly explicit, procedural explanations can interfere with the development of flexible and adaptable knowledge that will enable students to solve new problems within the domain. As articulated by Chin and Chia (2005), ''in the actual implementation of project work, the essence of inquiry may get diluted, displaced, or distorted, if students merely follow prescribed procedures'' (p. 45). However, in our study, the procedural instruction provided procedural heuristics that can be applied to a wide range of authentic research tasks. The instruction was not specifically geared to the experiment on which the students were assessed. Further, the assessment rubric was developed completely independently of the instruction used in either condition and was not adapted to meet any specific elements of the laboratory activity. The effectiveness of the CTA-based instruction in the treatment condition indicates that there are circumstances under which the assumption that ''favorable conditions for learning . . . [entail] conditions for which no known procedures are available'' (Roth, 1994, p. 216) does not hold. Future studies should assess the impacts of CTA for tasks that offer students greater autonomy to define their own research questions and experiments. Results would help to further characterize the impact that CTA-based instruction can have on transfer of knowledge and development of inquiry skills in ill-structured science domains.

The work reported in this paper is supported in part by a grant of the National Science Foundation (NSF-0653160) to David Feldon (P.I.), Kirk Stowe, and Richard Showman. The views in this paper are those of the authors and do not necessarily represent the views of the supporting funding agency. The authors would also like to thank Denise Strickland for her assistance in data collection and analysis. Previous drafts of this paper and related work were presented at the annual meetings of the National Association for Research on Science Teaching and the American Educational Research Association.

Notes

[1]Thjs definition of inquiry is one of three identified by Minner, Levy, and Century (in press, p. 3) as ''what scientists do (e.g., conducting investigations using scientific methods).'' It differs from the other two which address either the manner in which people learn or a pedagogical approach for teaching scientific content.

[2]These awards include: [State] College Teacher of the Year in Science and six university-wide teaching awards, including two created by named endowments. The name of the state and the specific university-wide awards are withheld to preserve anonymity.

[3]Cohen (1988) defines small effect sizes as $d \geq 0.2$, medium effect sizes as $d \approx 0.5$, and large effect sizes as $d > 0.8$.

References

Aarts, H., & Dijksterhuis, A. (2000). Habits as knowledge structures: Automaticity in goal-directed behavior. Journal of Personality and Social Psychology, 78(1), 53–63.

Anderson, J.R., Reder, L.M., & Simon, H.A. (1997). Situative versus cognitive perspectives: Form versus substance. Educational Researcher, 26, 18–21.

Astin, A.W., & Astin, H.S. (1993). Undergraduate science education: The impact of different college environments on the educational pipeline in the sciences. Los Angeles, CA: Higher Education Research Institute, UCLA.

Bargh, J.A., Gollwitzer, P.M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: The nonconscious activation and pursuit of behavioral goals. Journal of Personality and Social Psychology, 81(6), 1014–1027.

Barnett, S.M., & Ceci, S.J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128(4), 612–637.

Bianchini, J.A., Whitney, D.J., Breton, T.D., & Hilton-Brown, B.A. (2001). Toward inclusive science education: University scientists' views of students, instructional practices, and the nature of science. Science Education, 86, 42–78.

Blessing, S.B., & Anderson, J.R. (1996). How people learn to skip steps. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(3), 576–598.

Bransford, J.D., & Schwartz, D.L. (1999). Rethinking transfer: A simple proposal with multiple implications. Review of Research in Education, 24, 61–100.

Britt, T.W. (2005). The effects of identity-relevance and task difficulty on task motivation, stress, and performance. Motivation and Emotion, 29(3), 189–202.

Campbell, D.T., & Stanley, J.C. (1966). Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. Cognition and Instruction, 8, 293–332.

Chao, C., & Salvendy, G. (1994). Percentage of procedural knowledge acquired as a function of the number of experts from whom knowledge is acquired for diagnosis, debugging, and interpretation tasks. International Journal of Human-Computer Interaction, 6(3), 221–233.

Chin, C., & Chia, L. (2005). Problem-based learning: Using ill-structured problems in biology project work. Science Education, 90(1), 44–67.

Chinn, C.A., & Brewer, W.F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. Review of Educational Research, 63(1), 1–49.

Cicchetti, D.V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences, 14, 119–135.

Clark, R.E. (1999). Yin and yang: Cognitive motivational processes operating in multimedia learning environments. In: J.J.G. van Merriënboer (Ed.), Cognition and multimedia design. Herleen, Netherlands: Open University Press.

Clark, R.E. (2009). How much and what type of guidance is optimal for learning from instruction? In: S. Tobias & T.M. Duffy (Eds.), Constructivist theory applied to instruction: Success or failure? New York: Routledge, Taylor and Francis.

Clark, R.E., & Estes, F. (1996). Cognitive task analysis. International Journal of Educational Research, 25(5), 403–417.

Clark, R.E., Feldon, D.F., Van Merriënboer, J.J.G., Yates, K.A., & Early, S. (2008). Cognitive task analysis. In: J.M. Spector M.D. Merrill J.J.G. van Merriënboer & M.P. Driscoll (Eds.), Handbook of research on educational communications and technology (3rd ed., pp. 577–593). New York: Routledge.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooke, N.J. (1992). Modeling human expertise in expert systems. In: R.R. Hoffman (Ed.), The psychology of expertise: Cognitive research and empirical AI (pp. 29–60). Mahwah, NJ: Lawrence Erlbaum Associates.

Cooke, N.J., & Breedin, S.D. (1994). Constructing naive theories of motion on the-fly. Memory and Cognition, 22, 474–493.

Crandall, B., & Getchell-Reiter, K. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the "intuition" of NICU nurses. Advances in Nursing Sciences, 16, 42–51.

Delamont, S., & Atkinson, P. (2001). Doctoring uncertainty: Mastering craft knowledge. Social Studies of Science, 31(1), 87–107.

Dunbar, K. (2000). How scientists think in the real world: Implications for science education. Journal of Applied Developmental Psychology, 21(1), 49–58.

Eilam, B. (2002). Phases of learning: Ninth graders' skill acquisition. Research in Science & Technological Education, 20(1), 5–24.

Ericsson, K.A., & Charness, N. (1994). Expert performance: Its structure and acquisition. American Psychologist, 49(8), 725–747.

Feldon, D.F. (2007). Implications of research on expertise for curriculum and pedagogy. Educational Psychology Review, 19(2), 91–110.

Feldon, D.F. (in press). Do psychology researchers tell it like it is? A microgenetic analysis of research strategies and self-report accuracy. Instructional Science.

Feldon, D.F., & Stowe, K. (2009). A case study of instruction from experts: Why does cognitive task analysis make a difference? Technology, Instruction, Cognition, and Learning, 7(2), 103–120.

Gobet, F. (1998). Expert memory: A comparison of four theories. Cognition, 66, 115–152.

Hmelo-Silver, C.E., Duncan, R.G., & Chinn, C.A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). Educational Psychologist, 42(2), 99–107.

Hodson, D. (1996). Laboratory work as scientific method: Three decades of confusion and distortion. Journal of Curriculum Studies, 28(2), 115–135.

Jonassen, D.H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. Educational Technology Research & Development, 45(1), 65–95.

Jonassen, D.H., Tessmer, M., & Hannum, W.H. (1999). Task analysis methods for instructional design. Mahwah, NJ: Lawrence Erlbaum Associates.

Keys, C.W. (2000). Investigating the thinking processes of eighth grade writers during the composition of a scientific laboratory report. Journal of Research in Science Teaching, 37(7), 676–690.

Kirschner, P., Sweller, J., & Clark, R.E. (2006). Why minimally guided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. Educational Psychologist, 41(2), 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction. Psychological Science, 15(10), 661–667.

Klein, S.P., Stecher, B.M., Shavelson, R.J., McCaffrey, D., Ormseth, T., Bell, R.M., Comfort, K., & Othman, A.R. (1998). Analytic versus holistic scoring of science performance tasks. Applied Measurement in Education, 11(2), 121–138.

Koedinger, K.R., & Anderson, J.R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. Cognitive Science, 14, 511–550.

Koslowski, B., & Maqueda, M. (1993). What is confirmation bias and when do people actually have it? Merrill-Palmer Quarterly, 39, 104–130.

Lawson, A.E. (1978). The development and validation of a classroom test of formal reasoning. Journal of Research in Science Teaching, 15, 11–24.

Lawson, A.E. (2000). Classroom test of scientific reasoning: Multiple choice version (revised edition). Tempe, AZ: Arizona State University.

Lawson, A.E., Clark, B., Cramer-Meldrum, E., Falconer, K.A., Sequist, J.M., & Kwon, Y.-J. (2000). Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? Journal of Research in Science Teaching, 37, 81–101.

Leckie, G.J. (1996). Desperately seeking citations: Uncovering faculty assumptions about the undergraduate research process. The Journal of Academic Librarianship, 22, 201–208.

Lederman, N. (1998). The state of science education: Subject matter without context. The Electronic Journal of Science Education, 3(2).

Lee, R.L. (2003). Cognitive task analysis: A meta-analysis of comparative studies. Unpublished doctoral dissertation, University of Southern California, Los Angeles, California.

Lee, J.Y., & Reigeluth, C.M. (2003). Formative research on the heuristic task analysis process. Educational Technology Research and Development, 51(4), 5–24.

Marsh, H.W., & Bell, S. (1981). Interjudgmental reliability of reviews for the Journal of Educational Psychology. Journal of Educational Psychology, 73(6), 872–880.

Merrill, M.D. (2002). A pebble-in-the-pond model for instructional design. Performance Improvement, 41(7), 39–44.

Minner, D.D., Levy, A.J., & Century, J. (in press). Inquiry-based science instruction—What is it and does it matter? Results from a research synthesis years 1984 to 2002. Journal of Research in Science Teaching. DOI 10.1002/tea.20347

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231–259.

Onwuegbuzie, A.J., Slate, J.R., Paterson, F.R.A., Watson, M.H., & Schwartz, R.A. (2000). Factors associated with achievement in educational research courses. Research in the Schools, 7(1), 53–65.

Paas, F., Tuovinen, J.E., van Merriënboer, J.J.G., & Darabi, A.A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. Educational Technology Research and Development, 53(3), 26–34.

Palmquist, B.C., & Finley, F.N. (1997). Preservice teachers' views of the nature of science during a postbaccalaureate science teaching program. Journal of Research in Science Teaching, 34, 595–615.

Pintrich, P.R., Smith, D.A.F., García, T., & McKeachie, W.J. (1991). A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). Ann Arbor: University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning.

Pintrich, P.R., Smith, D.A.F., García, T., & McKeachie, W.J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). Educational and Psychological Measurement, 53, 801–813.

Posner, G.J., Strike, K.A., Hewson, P.W., & Gertzog, W.A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education, 66(2), 211–227.

Renkl, A., & Atkinson, R.K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. Educational Psychologist, 38(1), 15–22.

Rikers, R., Schmidt, H.G., & Boshuizen, H. (2000). Knowledge encapsulation and the intermediate effect. Contemporary Educational Psychology, 25(2), 150–166.

Rosenthal, R. (1966). Experimenter effects in behavioral research. New York: Appleton, Century, Crofts.

Roth, W.-R. (1994). Experimenting in a constructivist high school physics laboratory. Journal of Research in Science Teaching, 31, 197–223.

Sadeh, I., & Zion, M. (2009). The development of dynamic inquiry performances within an open inquiry setting: A comparison to guided inquiry setting. Journal of Research in Science Teaching, 46(10), 1137–1160.

Savery, J.R., & Duffy, T.M. (2001). Problem based learning: An instructional model and its constructivist framework. Center for Research on Learning and Technology technical report No 16-01.

Schaafstal, A., Schraagen, J.M., & van Berlo, M. (2000). Cognitive task analysis and innovation of training: The case of the structured troubleshooting. Human Factors, 42(1), 75–86.

Schraagen, J.M., Chipman, S.F., & Shute, V.J. (2000). State-of-the-art review of cognitive task analysis techniques. In: J.M. Schraagen, S.F. Chipman, & V.L. Shalin (Eds.), Cognitive task analysis (pp. 467–468). Mahwah, NJ: Lawrence Erlbaum Associates.

Schunn, C.D., & Anderson, J.R. (2001). Acquiring expertise in science: Explorations of what, when, and how. In: K. Crowley, C.D. Schunn, & T. Okada (Eds.), Designing for science: Implications from everyday, classroom, and professional settings (pp. 83–113). Mahwah, NJ: Lawrence Erlbaum Associates.

Schwartz, D.L., & Martin, T. (2004). Inventing to prepare for learning: The hidden efficiency of original student production in statistics instruction. Cognition & Instruction, 22, 129–184.

Seymour, E. (2001). Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology. Science Education, 86, 79–105.

Seymour, E., & Hewitt, N. (1997). Talking about leaving: Why undergraduates leave the sciences. Boulder, CO: Westview.

Spitzer, T.M. (2000). Predictors of college success: A comparison of traditional and nontraditional age students. NASPA Journal, 38(1), 82–98.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. Journal of Experimental Psychology: General, 119(2), 176–192.

Timmerman, B.E. (2008). Peer review in an undergraduate biology curriculum: Effects on students' scientific reasoning, writing, and attitudes. Unpublished doctoral dissertation, Curtin University of Technology, Perth, Western Australia.

Timmerman, B.E., Strickland, D., Johnson, R.L., & Payne, J. (in press). Development of a 'universal' rubric for assessing students' scientific reasoning skills using scientific writing. Assessment and Evaluation in Higher Education.

Tuovinen, J.E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. Journal of Educational Psychology, 91(2), 334–341.

Velmahos, G., Toutouzas, K., Sillin, L., Chan, L., Clark, R.E., Theodorou, D., & Maupin, F. (2004). Cognitive task analysis for teaching technical skills in an animate surgical skills laboratory. The American Journal of Surgery, 187, 114–119.

Wigboldus, D.H.J., Dijksterhuis, A., & van Knippenberg, A. (2003). When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. Journal of Personality and Social Psychology, 84, 470–484.

Appendix 1: Final Cognitive Task Analysis Protocol

Observation

(1) Observe phenomena of personal and scholarly interest until a pattern (visual, temporal, spatial and/or interactional) or the disruption of a known pattern is noticed in environment or reported in primary research literature.
(2) IF pattern implicates the form or function of phenomena, organisms, or systems as discussed in relevant literature base THEN describe patterns/disruptions and circumstances under which they are observed in a lab notebook.
(3) IF pattern does not implicate form or function of phenomena, organisms, or systems as discussed in relevant literature base THEN return to Step 1.
(4) Seek replication of pattern/disruption in different instances, locations, or conditions.
(5) IF pattern/disruption cannot be replicated during additional observations THEN return to Step 1.
(6) IF pattern is consistent across replications THEN formulate questions asking about the characteristics of variables of interest and the nature (magnitude, direction, causality) of the relationships among them and write them in lab notebook.
(7) IF any question asks about historical fact (e.g., *Did* change *x* occur through mechanism *y*?) THEN change wording and tense of question to ask about existence of relationship in planned experiments (e.g., *Will* change *x* occur through mechanism *y* under experimental conditions *a*, *b*, and *c*?) and write it in lab notebook.

(8) Refer to research literature discussing variables, phenomena, organisms, or systems that are part of observed pattern.
(9) IF phenomenon of interest cannot be reliably measured THEN return to Step 1.
(10) IF literature does not report findings of correlations between observed outcomes/variables THEN formulate research questions to test significance of correlation and identification of additional relevant factors, write it in lab notebook, and begin Refine Research Question procedure.
(11) IF literature reports findings of correlations between observed outcomes/variables of interest THEN identify theoretical explanations and remaining research questions.
  (a) IF theoretical explanations are missing or inadequate THEN reformulate question from Step 6 such that an answer to the question would provide/expand theoretical explanation, write it in lab notebook, and begin Refine Research Question procedure.
  (b) IF theoretical explanations are present/sufficient THEN compare observation with predictions of theory.
    (i) IF observations do not match predictions of theory THEN reformulate question from Step 6 to ask what this basis of difference is, write it in lab notebook, and begin Refine Research Question procedure.
    (ii) IF observations do match predictions of theory THEN determine if facet of theory is underspecified or requires further empirical evidence to support it.
      (1) IF theory is adequately specified and/or does not require further empirical evidence to support it THEN return to Step 1.
      (2) IF theory is inadequately specified or requires further support THEN formulate question to which an answer will clarify relationship or provide confirmatory or disconfirmatory evidence of theory's validity, write it in lab notebook, and begin Refine Research Question procedure.

Refine Research Question

(1) IF previous findings reported in the literature suggest a partial answer to your question THEN identify the factors in those studies that are relevant and correlate with the phenomenon you observed.
(2) Select at least one of the factors that is likely to be helpful in answering your question and revise your research question to ask about potential interactions with other factors in the pattern you observed. Then write the revised question in your lab notebook.
(3) IF knowledge base/theory suggests more sources of variance than are feasible to incorporate into an experiment or meaningfully analyzed THEN select no more than three factors that potentially interact in order of their apparent relevance to the observed pattern and the research question.
(4) IF the process of answering the research question as written does not require the use of specific variables or operational definitions THEN identify a scientific model or theory from the literature to narrow scope of question and repeat Steps 1–3.
(5) IF the research question as written cannot be answered to a reasonable degree of certainty through the analysis of a finite number of experiments THEN narrow scope of question and repeat Steps 1–3.
(6) IF the research question as written can be answered to a reasonable degree of certainty through the analysis of a finite number of experiments THEN go to Formulate Hypothesis procedure.

Formulate Hypothesis

(1) Specify expected relationship between factor of interest (independent variable) and the measured dependent variable(s) that fully answers Refined Research Question. Write it in lab notebook.
(2) IF literature base of relevant theory or model reports existing measurement techniques that are appropriate for investigating current research question THEN select most feasible measures given available resources.
(3) IF literature base does not report existing measurement techniques appropriate for investigating the current research question THEN select measures from alternative literature or develop and validate measure.

(4) Specify the means by which variables will be measured and levels of independent variables will be verified and write in lab notebook.
(5) Specify mechanism and/or condition that will generate expected outcome.
(6) Write information from Step 1 and Step 4 into a single sentence in lab notebook as formulated hypothesis.
(7) IF independent variable and/or dependent variable cannot be measured with as much precision or accuracy as used in published literature on the phenomena or factors of interest THEN return to Step 1.
(8) IF it would not be possible for another researcher to replicate or verify the measures used in Step 2 THEN return to Step 1.
(9) IF all mechanisms or conditions in hypothesis statement cannot be manipulated, controlled, or accurately and systematically measured THEN return to Step 1.
(10) IF all mechanisms or conditions in hypothesis statement can be manipulated, controlled, or accurately and systematically measured THEN go to Design Experiment procedure.

Experimental Design

(1) Identify and list in lab notebook which factors from the literature, personal knowledge of relevant biological system(s), and conditions surrounding observation that led to the research question are likely to influence obtained measurements of dependent variable stated in hypothesis.
(2) IF listed factor is not independent variable in hypothesis THEN:
   (a) IF elimination of factor will not disrupt biological system or threaten organism health/survival THEN list in lab notebook as factor to be eliminated.
   (b) IF elimination of factor will disrupt biological system or threaten organism health/survival THEN list in lab notebook as factor to be held constant at level recommended in literature or observed in environment as satisfactory for maintenance of biological system or health/survival of organism.
   (c) IF factor cannot be controlled or eliminated to prevent influence on dependent variable THEN identify and list in lab notebook the locations or other environmental conditions that will ensure equivalent impact on all experimental conditions.
   (d) IF no locations or environmental conditions exist where equivalent impact of factor on measurement of dependent variable can be identified THEN identify appropriate measure of factor from literature and include measurement of factor in protocol for reference during data analysis.
(3) IF a listed factor is an independent variable in the hypothesis THEN:
   (a) IF the factor can be controlled THEN list in the lab notebook the levels of factor that the theory or hypothesis predicts will be different enough to produce different means in measurements of dependent variable(s) for each condition without producing ceiling or floor effects in measurement data from dependent variable.
   (b) IF the factor cannot be controlled THEN list in lab notebook the placements or other conditions for measuring dependent variable(s) that the theory or hypothesis predicts will be different enough to produce different means in measurements of dependent variable(s) for each condition without producing ceiling or floor effects in measurement data from dependent variable.
   (c) IF the factor cannot be controlled and there is only one condition for the factor THEN identify and list in lab notebook appropriate ways to measure the level of the factor and note for data analysis that study is non-experimental.
(4) IF levels listed for each independent variable are not likely to replicate observed pattern in at least one, but not all, experimental conditions THEN identify and list in lab notebook necessary additional level(s) to meet condition.
(5) For each item on the list of factors to be eliminated or controlled, generate and list in lab notebook the techniques for doing so from literature, personal knowledge of relevant biological system(s), and conditions surrounding observation that led to research question.
(6) IF one or more techniques listed for factor elimination or control are feasible for conditions and resources of study THEN select technique considered most effective in literature.
   (a) IF multiple techniques are considered equally effective THEN select easiest and cheapest listed technique and write it into protocol.

(7) IF no techniques listed for factor elimination or control are feasible for conditions and resources of study THEN identify and recruit/obtain necessary expertise or resources.
   (a) IF necessary expertise or resources are unattainable THEN return to Step 2.
(8) IF hypothesis does not predict explicitly the outcomes for each condition THEN write in lab notebook additions to hypothesis that specify predictions for each condition in either absolute (prediction of measured outcome) or relative terms (prediction of higher or lower measurement than other conditions).
(9) Conduct experiment according to protocol.

Data Analysis

(1) Record values of dependent variables and measured independent variables for each condition in experiment.
(2) IF event occurs differently from protocol THEN record events and note potential impact on measurements in lab notebook.
(3) Compute descriptive statistics for all measurements. Record results in lab notebook.
(4) Graph data:
   (a) IF data reflect a single dependent variable changing in relation to a known factor (e.g., time, location, experimental condition, etc.) THEN:
   (i) IF points represent a linear trend THEN connect points to form line.
   (ii) IF points represent a nonlinear trend THEN draw smooth curve along points maintaining equal distances between the line and points above or below it.
   (b) IF data reflect multiple variables changing in relation to a known factor (e.g., time, location, experimental condition, etc.) THEN create a clustered bar graph with one vertical bar per dependent variable clustered at each time point. Every bar representing a particular variable must use the same color at every point on the graph. The color must be easily distinguishable from the colors used for the other variables in the cluster (e.g., black, dark gray, white, etc.).
(5) Based on trends apparent in graph and descriptive statistics, select appropriate statistical method to test hypothesis.
(6) IF outliers are detected THEN note values of outliers and conditions in which they occurred in lab notebook, withhold from dataset, and compute statistical outcomes.
(7) IF results are statistically significant THEN repeat analysis including outliers and record results in lab notebook.
   (a) IF results still significant THEN determine extent to which results support hypotheses.
   (b) IF results not significant THEN determine likely causes of outliers.
   (i) IF outliers appear to be systematic THEN redesign experiment to include likely source of outliers as relevant factor when designing protocol.
   (ii) IF outliers do not appear to be systematic and resources permit THEN replicate experiment.
   (1) IF replication experiment does not produce the same statistically significant results THEN carefully review protocols and notes in lab notebook to determine differences between original experiment and replication that could explain different results.
(8) IF results are not significant, but there is a trend in the data, THEN conduct a statistical power analysis and repeat Step 9 of Experimental Design procedure with increased sample size if feasible.
(9) IF results are not statistically significant and there is no trend in the data THEN return to Step 1 of Observation procedure.

Appendix 2: Comparison of Control and Treatment Video Content and Alignment With Performance Outcome Criteria

Video Title	Control (Traditional Content)	Treatment (CTA Generated Content)
Transcribed examples of content		
Observation I	'. . . The skill of observation is essential, because what comes out is often not what we expect . . .' 'Essentially observation allows me to test the hypothesis by coming back over several (time periods) . . .' 'Observation is not done only by looking at something directly . . ., but by reflecting back upon . . . what's been written.' 'So, the verifiable observations of others are crucial in the process of observation.'	'. . . All of these observations occur looking for a pattern . . . or alternatively information that contradicts a pattern.' 'This means that you are going to continue your observations . . ., until you develop a series of observations . . . relevant to biological sciences.' 'If in replicating your experiment you do not obtain the same observations . . . If your observations are not replicated . . . this is a strong indication you need to go back to step 1 (initial observations).'
Observation II	'The concept of standards, coupled with observation is essential. Any piece of equipment you use to observe a phenomenon is only as good as its reproducibility and its accuracy.'	'. . . When you read other people's published research . . . At this point you will need to evaluate what has been done in order to know how to proceed. The first thing you need to do is to determine if any theories exist that explain the pattern in your observations.' '. . . The next step is to go to the literature and read what scientists have already learned about these things.' 'When the situation arises where a theory does not explain . . . You need to decide if the theoretical explanation that is offered: (1) provides all of the detail that is needed to understand the phenomena of interest; (2) has sufficient empirical evidence . . . to support it.' 'However, it is possible that nothing has been published . . . If this is the case, you need to review your list of questions and select a . . . focuses on measuring only the reliability, or significance, of the relationship you observed.' 'It is important to understand how the variables in your pattern are measured. If no effective and reliable means exists . . . return to step 1 [initial observation] . . . If variables can be reliably measured then you need to look for previous reports . . .' 'But if theory needs to be further developed . . . then should reformulate a question, if answered, will either clarify the theory or alternatively contribute significantly to supporting or refuting it.'

Video Title	Control (Traditional Content)	Treatment (CTA Generated Content)
Refine research question	'. . . A research question comes out of our observations . . . It may come from a previous set of experiments, it may come from something we read . . .' '. . . You start breaking that question down until you can handle it intellectually, objectively, in the laboratory, in the field . . .' 'We're going to simplify the question, we're going to formulate a question which can then be tested and hopefully a specific answer is achieved or arrived at.'	'. . . You're going to look in the primary literature . . . If . . . find partial answers to your questions then you want to incorporate that information along with any factors that you've obtained while you're observing and include those factors as a list within your notebook.' "Does the question that I formulated obviously present to me how and what I'm going to be measuring?" If it does proceed. 'If . . . not, then you're going to need to return to the literature . . .' '. . . select one question if at all possible and focus on this. Now, taking that question, reword it so that it takes into account all of the specific things that you've noted during your observations.'
Hypothesis	'. . . As we design experiments to test the hypothesis, all of the data, the results, must be judged against the hypothesis.' 'An hypothesis is something which is developed carefully . . . it provides the core or base around which you can design an experiment.' 'It's a guess . . . that is based upon the ideas that you have accumulated, whether from previous data, from reading . . . it's not based upon simply an idea that you have. A working hypothesis as we normally use it in the lab represents our first approximation . . .' 'In the world of science, it is absolutely essential that every hypothesis: (a) be testable . . .; (b) that the individual testing the hypothesis be willing to change the hypothesis at any time if the data are inconsistent with the working hypothesis.' 'If in fact the experiment is valid, it's reproducible, and is inconsistent with the hypothesis . . . there has to be an adjustment to the hypothesis.' 'We refine the hypothesis as we collect new data, as long as data continues to be consistent with the hypothesis and verified, you are fine. The minute that you find an inconsistency, then whether the hypothesis needs to be changed or you need to go back . . . and ask, "why might there be this inconsistency?"' 'Any information that you gather which is inconsistent with the hypothesis requires one of two things: you either must go back and reconsider the experiment and ask, "did the experiment ask the question I wanted?' Or, "did the experiment succeed, did I do something wrong . . .?"' 'The variable is never the data. The data either verifies the hypothesis or nullifies it, and, as a consequence, if the hypothesis is nullified, you must go back and adjust the hypothesis.'	'First step is to write down . . . relationship you expect to see between the factor(s) you include in your research question.' 'Remember your research question was your independent variable and the result you plan to measure, your dependent variable . . . Once you have decided which measure to use . . .' 'Next you need to think about the information you gathered when you read the scientific literature on your topic. Did it talk about specific ways to measure changes in your independent variable? If it did, then decide which method of measuring is most practical for you' 'Next review the details and predictions that you've written down and be sure that they all come together as one or two clear and grammatically correct sentences. This type of editing . . . is very important to help make it clear what it is you think will happen and why. This edited statement is your formal hypothesis.' 'Similarly, if it would not be possible for another researcher to replicate or verify the measures you have specified in your hypothesis, then you need to return to step 1 and rewrite your prediction statement.' 'Using the approach I just described for the dependent variable, repeat the process of choosing measures for each of you factors or independent variables . . . If it turns out that any of your dependent or independent variables cannot be controlled, measured, or manipulated . . . the it is time to return to step 1 and rewrite your prediction statement.'

Video Title	Control (Traditional Content)	Treatment (CTA Generated Content)
Experimental design I	'These are called control variables and those are the components of the experiment which we want to keep constant during the length of the experiment.' 'An independent variable is that variable which you yourself have under control . . .' 'The dependent variable in your experiment is the one which is going to vary as a consequence of the change in the pH.' 'We all have to agree that there is a universe, we can see it, we can measure it, we can design experiments and through these experiments determine something about the nature of the universe.' '. . . Data are as simple as that . . . You can't manipulate it, you can't change it, you can't ignore it.' '. . . No matter how many controlled variables, you consider and control for, you still have an infinite number of additional ones which you may or may not considered . . . as a consequence, may very well have affected the outcome of the experiment. And thus, the data acquired is subject to an alternative interpretation.'	'For those factors that you list, which are not independent variables in your hypothesis, you need to evaluate them to determine whether or not they should be eliminated, held constant, avoided, or measured for later analysis.' 'If you can eliminate a relevant factor that is not in your hypothesis, without disrupting the biological system or threatening the organism's health, then list it in your lab notebook as a factor to be eliminated. However, if by eliminating the factor you will disrupt the system or harm the organism, then list it in your notebook as a factor to be held at constant levels that are recommended in the literature or sufficient in nature to keep the organism healthy. If the factor you list cannot be controlled or eliminated to keep it from influencing your measurements of the dependent variable, then identify and write down in your lab notebook the locations or other environmental conditions that will make sure that the impacts are the same across all conditions or versions of the experiment.' 'If this is a factor you can control then you need to list in your lab notebook which levels of the factor will be different enough to produce meaningful and significant differences in the dependent variable when you measure it . . . If the levels you've listed for each of your independent variables is not likely to reproduce the initial pattern that you originally observed in at least one condition, then identify and list in your lab notebook the additional levels that will allow you to do that.'
Experimental design II	'One of the last things one has to consider in designing experiments is the concept of controls. Now, controls are in fact essential for any experiment. These are perhaps more important when you're dealing with a chemical reaction or a physics experiment, or any experiment in which instrumentation is required.' 'There, the number of organisms involved becomes critical, because variation is directly related with the number of subjects that you're examining. The key to this is replication. And replication is determined in great part by the sample size that you have available to you.' '. . . Data analysis is something that you have to take into consideration when you first begin the design of your experiment . . . But, thinking ahead and asking yourself "what is it that I'm actually asking?" and "What is it that I actually expect to get?"'	'For each one, write down in your lab notebook which techniques you will use to deal with it. These strategies can come from scientific literature, your personal knowledge of the relevant biological systems, and/or the conditions surrounding your original observation that led to your research question. If you find that you can identify more than one feasible strategy to eliminate or control a particular factor, then select a technique in the literature that is considered the most effective.' 'If your hypothesis does not state your expected outcomes for every condition of your experiment then you will need to list explicitly in your lab notebook what you expected the dependent variable to do for each combination of factorial levels that you will test.'

Video Title	Control (Traditional Content)	Treatment (CTA Generated Content)
Data analysis I	'. . . One has to recognize that as you collect your data, not all of it is going to fit perfectly. The question then is, ''are you allowed to do something with that data?'' As a matter of practice, the answer is no. You must include all of your data . . . Furthermore, you must not move, delete, add, or in any other way modify the data once it has been collected. As I've said before, the data are.' 'You can choose a number of different types of statistic methodology, but what's most important, is that you understand what that statistical methodology tells you . . . If it was not, it's entirely possibly that at the end of your analysis, you're going to discover there simply was not an answer available. There was not enough statistical power in your experiment, and, as a consequence, you can't get an answer.'	'Now you need to graph your data. To do this you should recall from your design how many dependent variables you have. If you have only one and you're measuring has changed in relation to the independent variable, plot the points on the graph where the ''*x*'' axis represents the value of the independent variable, and the ''*y*'' axis represents the value of the dependent variable . . .' 'Also, if anything happened during your experiment that was different from how it was designed, in other words, if the protocol changed in some way, then you need to record the details of those events and how they may have impacted your measurements in your lab notebook. This information will be helpful later if you need to explain why things did not turn out as you predicted in your hypothesis. Next you should compute the appropriate descriptive statistics for all of the data and record the results in your lab notebook . . . at the minimum, you should calculate the mean and the standard deviation.'
Data analysis II	'And out of the pattern that appears of the plotting of the data you begin to see a relationship between unknown factors and known factors. This plotting of the data requires that you think about how to deal with individual points . . . connecting points . . . draw a line between . . . straight line . . . curved . . .'	'Based on the trends in your graphs and the descriptive statistics from your variables, you next need to select an appropriate statistical method to test your hypothesis.' 'If your results are not statistically significant but you do see a trend in your data, then you should conduct what's called a statistical power analysis to figure out a better sample size for getting significant results and rerun your experiment accordingly.' 'If your results are not statistically significant and you do not see any meaningful trend in your data then it is unfortunately time to go back to the very beginning and start conducting initial observations to look for a pattern of interest to investigate.' '. . . You need to check your measurements to see if there are outliers in your data. Outliers are specific data points that deviate substantially from an otherwise stable trend that you can see in your graphs. If you do have outliers then write down in your lab notebook the values for these data points and the experimental conditions under which they occurred. Next, pull them out of the larger data set and run your statistics.' '. . . If they [outliers] are [due to unplanned source of variation], . . . you need to . . . redesign the experiment to control or measure the new factor.' '. . . If similar outliers show up again in your replication data then you need to keep looking for another source of variance in your protocol.'