Higher Education Quality Council of Ontario

An agency of the Government of Ontario

Evaluating Critical Thinking and Problem Solving in Large Classes:
Model Eliciting Activities for Critical Thinking Development

James Kaupp, Brian Frank and Ann Chen
Queen's University

Published by

The Higher Education Quality Council of Ontario

1 Yonge Street, Suite 2402
Toronto, ON Canada, M5E 1E5

Phone: (416) 212-3893
Fax: (416) 212-3899
Web: www.heqco.ca
E-mail: info@heqco.ca

Cite this publication in the following format:

Kaupp, J., Frank, B., & Chen, A. (2014). *Evaluating Critical Thinking and Problem Solving in Large Classes: Model Eliciting Activities for Critical Thinking Development.* Toronto: Higher Education Quality Council of Ontario.

Executive Summary

This report describes a study to investigate the impact of a MEA-integrated curriculum on critical thinking (CT) development in a first-year engineering course at Queen's University. The course focuses on developing problem solving, modeling and critical thinking skills in part by using complex contextualized problems known as model eliciting activities (MEAs). In addition, the MEAs provide a means for the rigorous, authentic and sustainable course-embedded assessment of the aforementioned skills. The study was conducted over the course of the fall semester in the 2012-2013 school year, with 542 students participating.

Explicit critical thinking instruction using the Paul & Elder model was embedded into the course experience alongside a series of model eliciting activities to develop students' critical thinking skills. Several standardized tests and other instruments were used to assess critical thinking skill at the beginning and end of the course experience, including:

1) the Cornell Critical Thinking Test: Level Z
2) the International Critical Thinking Essay Test (ICTET)
3) the Collegiate Learning Assessment (CLA)
4) course surveys
5) think aloud protocols: an interview session in which participants are asked to "think aloud" their answers to a problem or scenario

Upon conclusion of the study and analysis of the data, we observed no significant gain in critical thinking skill (CTS) over the duration of the course in two of the standardized instruments (Cornell Critical Thinking Test Level Z and International Critical Thinking Essay Test), and one instrument was only available for administration as a pre-test (Collegiate Learning Assessment). Participating control groups faced recruitment and attrition challenges and did not have sufficient participation for comparison. However, we observed improvement in student performance on critical thinking outcomes embedded in course activities (MEAs) and in think aloud exercises. Student responses to survey questions asking them about their perceptions of their critical thinking development during the course identified course elements and MEAs as useful for developing critical thinking. Of all the standardized instruments, the Collegiate Learning Assessment exhibited the highest correlations with the scores given by graders on the model eliciting activities, as the application of critical thinking skills between this instrument and the model eliciting activities are very similar.

We have identified several elements that we believe should be carefully considered in future work:

1) The alignment between the critical thinking framework used for instruction, embedded activities and standardized instruments should be carefully considered. Divergence between these elements may affect measured CT outcomes and the utmost care should be taken to maintain a high degree of alignment.
2) The task alignment between standardized tests and embedded activities should be preserved. Specifically, how critical thinking is applied in the standardized tests and how it is applied in embedded activities should be virtually indistinguishable.
3) Student motivation and engagement was a considerable challenge. Testing fatigue resulting from the standardized instruments resulted in superficial approaches and performance issues. Assessments for pre-post-testing should be indistinguishable from the course experience to minimize these effects.
4) The standardized instruments may not possess the sensitivity to measure gains in CTS over the course of a single semester. Standardized tools may be more apt for longitudinal assessment of long-term development of CTS over the course of a program. The assessment of short-term development of CTS should be measured using student artifacts and authentic practices.

In order to address these challenges, we recommend careful selection of instructional frameworks, embedded activities and standardized instruments to maintain alignment and application of frameworks. To this end, future work in assessing and developing critical thinking in engineering, including work with the HEQCO Learning Outcomes Consortium, will continue using some standardized instruments, but will also use program-level rubrics (e.g., VALUE rubrics from the Association of American Colleges & Universities, AAC&U) to score student artifacts generated for academic work, to evaluate critical thinking longitudinally from first to fourth year.

Table of Contents

List of Figures

List of Tables

Introduction

The ability to solve problems and think critically are considered by many to be desired outcomes of the education system, both within K-12 and higher education. They are ever-present skills measured by many accreditation frameworks in the professional and higher education sectors, and consistently rank among the top skills and abilities desired in graduates, according to employer surveys (Hart Research Associates, 2008; 2013). Despite this prevalence, critical thinking and problem solving are often identified by employers as skills that require more emphasis in higher education (Hart Research Associates, 2008; Arum & Roksa, 2011). Recent evidence questions the degree to which current undergraduate education supports the development of critical thinking and complex problem solving skills (Arum & Roksa, 2011; Astin, 1993a; 1993b; Blaich & Wise, 2008; Klein et al., 2009; Pascarella, Blaich, Martin & Hanson, 2011). The development of critical thinking skills (CTS) is itself a complex issue, complicated by a lack of agreement on the definition of critical thinking and on an associated framework for its development (Ku, 2009). Popular frameworks of critical thinking include the Cornell-Illinois model (Ennis, Millman & Tomko, 1985), the Paul-Elder model (Paul & Elder, 2005; Paul & Elder, 1996), the CLA model (Shavelson, 2008), the APA Delphi model (Facione, 1990), and Halpern's Model for Critical Thinking (Halpern, 1999; Halpern & Riggio, 2002). Each of these frameworks or models proposes a different definition for critical thinking and a different set of skills, traits and abilities that comprise it. Instruction and assessment of CTS is also an area of particular difficulty, with the efficacy of pedagogical strategies for critical thinking development and the authenticity of critical thinking assessment under much scrutiny (Bensley & Murtagh, 2011; Solon, 2003).

Despite these underlying issues, there is general agreement that CTS are crucial for dealing with complex real-world problems. One approach to developing the ability to solve complex real-world problems in mathematically intense disciplines involves model eliciting activities (MEAs), realistic problems used in the classroom that require learners to document not only their solution to the problems but also their processes for solving them (Shuman, 2012; Shuman & Besterfield-Sacre, 2008). MEAs involve the creation of a mathematical description, procedure or system as part of the solution, a model which students use to develop and refine their process and solution (Chamberlin, 2004; Shuman & Besterfield-Sacre, 2008). MEAs have been developed and used in a variety of subject areas, including mathematics, economics and environmental engineering. Studies have shown MEAs to be valuable in helping students to develop conceptual understanding, knowledge transfer and generalizable problem-solving skills (Self, Shuman & Besterfield-Sacre, 2012; Yildirim, Shuman, Besterfield-Sacre & Yildirim, 2010).

Research Objectives

This report describes an investigation of CTS development in a first-year engineering course (APSC 100: Engineering Practise Module 1) at Queen's University. The primary objective of our study is to investigate the impact of the MEA-integrated curriculum on the development of students' critical thinking skills. With respect to this objective, we set out to benchmark the CTS of first-year engineers at the beginning of the fall semester of the 2012/2013 academic year, expose them to the MEA-integrated curriculum, and then assess their CTS at the conclusion of the course experience at the end of the fall semester. In addition, we intend to provide further analysis regarding the critical thinking instruments used in the study through the following research questions:

1) Is there a correlation between critical thinking instrument scores and MEA scores?
2) Is there a correlation between critical thinking instrument sub-scores and MEA sub-scores?
3) Is there a correlation between critical thinking ability and motivational factors?
4) Is there a correlation between critical thinking ability and specific course experiences?
5) Is there a correlation between critical thinking ability and specific extrinsic factors?
6) Are the critical thinking instruments used reliable and valid?

The secondary objective of the study is to provide an overview of how to assess critical thinking within an engineering context. With respect to this objective, we hope to provide other parties interested in critical thinking development and assessment with a starting point for future work. We provide a short review of critical thinking instruments, together with conclusions and recommendations regarding these tools and additional observations resulting from the study. To this end, we pose the following questions:

1) Is there evidence that MEAs have a significant positive impact on students' critical thinking skills?
2) Which critical thinking framework and which critical thinking instrument reflect the application of critical thinking skills in solving complex engineering problems?
3) To what extent does alignment of tasks between critical thinking instrument and complex engineering problems need to be preserved?
4) What are effective approaches to evaluating critical thinking skills in a course environment?

Three instruments were used in this study to evaluate the CTS of first-year engineering students. These instruments were used as both a pre- and post-test in order to benchmark the CTS of the incoming first-year students and determine the effectiveness of MEA instruction on developing students' critical thinking ability.

Literature Review

Model Eliciting Activities

MEAs have been used in engineering education at the university level since 2004 (Diefes-Dux et al., 2004; Moore & Diefes-Dux, 2004; Shuman & Besterfield-Sacre, 2008). Originally developed as an assessment tool in mathematics (Lesh, 1999; Lesh & Doerr, 2003) and still a topic of study at the middle school level, MEAs are also currently the focus of a four-year research project at seven American universities (MEDIA project, n.d.), which looks at their use in a variety of contexts in both large and small classes. Thus far, MEAs are showing promising results in developing students' topical conceptual understanding, information fluency, problem solving and communication skills. MEAs require students to draw upon prior knowledge and often help to identify and address misconceptions in the course of learning and promote connections between information.

MEAs are designed according to a set of six principles, adapted for use in engineering curriculum from their original mathematical context, outlined below (Lesh & Doerr, 2000; Moore & Diefes-Dux, 2004):

1) Model construction: The activity requires the construction of an explicit description, explanation or procedure for a mathematically significant situation.
2) Reality: Requires the activity to be posed in a realistic engineering context and to be designed so that the students can interpret the activity meaningfully from their different levels of mathematical ability and general knowledge.
3) Self-assessment: The activity contains criteria that students can identify and use to test and revise their current ways of thinking.
4) Model documentation: Students are required to create some form of documentation that will reveal explicitly how they are thinking about the problem situation.
5) Construct shareability and reusability: Requires students to produce solutions that are shareable with others and modifiable for other engineering situations.
6) Effective prototype: Ensures that the model produced will be as simple as possible yet still mathematically significant for engineering purposes.

MEA instruction places a considerable emphasis on the process used to solve the problem and the reasoning and thinking students used to develop their solutions rather than on the product of that methodology. The solution of an MEA requires participants to apply and combine multiple engineering, physics or mathematical

concepts drawn from their educational experience and previous background to formulate a general mathematical model that can be used to solve the problem. Students typically employ an iterative process approach to the MEA, first generating a model, testing the model and revising the model to develop a suitable solution (Lesh & Doerr, 2003). Students solve the MEAs as a part of a group, emulating the team-based experience typical of professional practice. The students' solutions to the MEA typically take the form of a comprehensive report outlining the process used to generate their solution to the problem.

There have been several studies investigating the impact of MEA instruction on student learning outcomes and general skill development. These studies have shown that MEAs:

1) Encourage a different perspective regarding the use of engineering concepts, with students applying concepts to achieve a broad, high-level solution rather than a low-level formulaic, rote approach (Shuman & Besterfield-Sacre, 2008).
2) Encourage students to work collaboratively and cooperatively as a group, honing teamwork and interpersonal skills and delivering a higher quality solution than individual submissions (Gokhale, 1995).
3) Encourage integration and synthesis of information and concepts spanning engineering and other disciplines (Yildirim et al., 2010).
4) Encourage reasoning and higher-order thinking skills through the ill-structured and complex nature of MEA instruction (Chamberlin, 2002).

The aforementioned benefits of MEAs lead to a more meaningful learning experience for students by engaging them in an exercise that reflects professional engineering practise. This meaningful learning experience helps foster both higher-level skills and desired generic learning outcomes of complex problem solving, communication, information literacy and critical thinking.

Think Aloud Protocols

In addition to critical thinking assessment tools, this study used think aloud exercises. Such protocols originate from cognitive science (Ericsson & Simon, 1993; Fonteyn et al., 1993; Van Someren, Barnard & Sandberg, 1994) as a way to observe and study concurrent reasoning for the purpose of analyzing participants' reasoning and information processing (Boren & Ramey, 2000; Ericsson & Simon, 1993; Ungson & Braunstein, 1982). Think aloud protocols have been used as a means to assess cognitive activities such as problem solving and critical thinking ability in engineering and other related fields (Daly, 2001; Ku & Ho, 2010a; Norris, 1990; Steif, Lobue, Kara & Fay, 2013). In these sessions, participants are presented with a task and an objective and are asked to "think aloud" their thought process as they work towards a solution. A facilitator is present as a passive observer and to prompt and remind participants to verbalize their thought processes. These sessions are recorded and transcripts are produced. The transcripts are then coded using protocol analysis according to a selected cognitive model or, in the case of this study, a critical thinking framework.

Critical Thinking Frameworks

Critical thinking frameworks each describe a different viewpoint on the complex construct of critical thinking. Each model is based on a working definition of critical thinking and provides a framework for the component skills, attributes, standards and dispositions according to the working definition. Many of these frameworks do not contain an explicit pedagogical strategy or developmental sequence for students; they simply provide a succinct definition of the construct and its components. However, a definition and framework form the basis of, and are essential to, the infusion of critical thinking into course curriculum. The descriptions of the models in the following section are included as an introduction to critical thinking frameworks and to provide insight into the assessments of CTS used in this study. For additional critical thinking frameworks, please consult

Appendix 9.

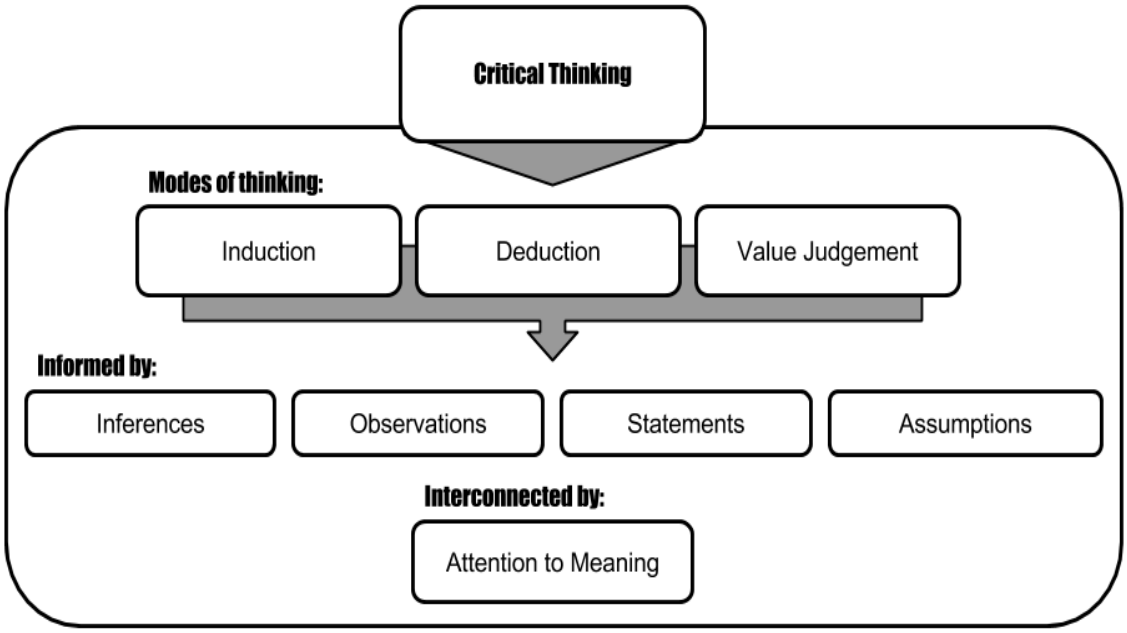
Cornell-Illinois Model

The Cornell/Illinois model of critical thinking was developed and refined by Robert Ennis based on the following working definition of critical thinking:

> Critical thinking is reasonable and reflective thinking focused on deciding what to believe or do (Ennis et al., 1985)

The model, illustrated in Figure 1 is divided and sub-classified based on three modes of critical thought (induction, deduction and value judging) and four methods on which they are based: the results of inferences, observations, statements and assumptions. Lastly, the model is connected by a common thread of attention to meaning which is interwoven throughout the four methods and three elements (Ennis et al., 1985).

Figure 1: The Cornell-Illinois Model



Paul-Elder Model

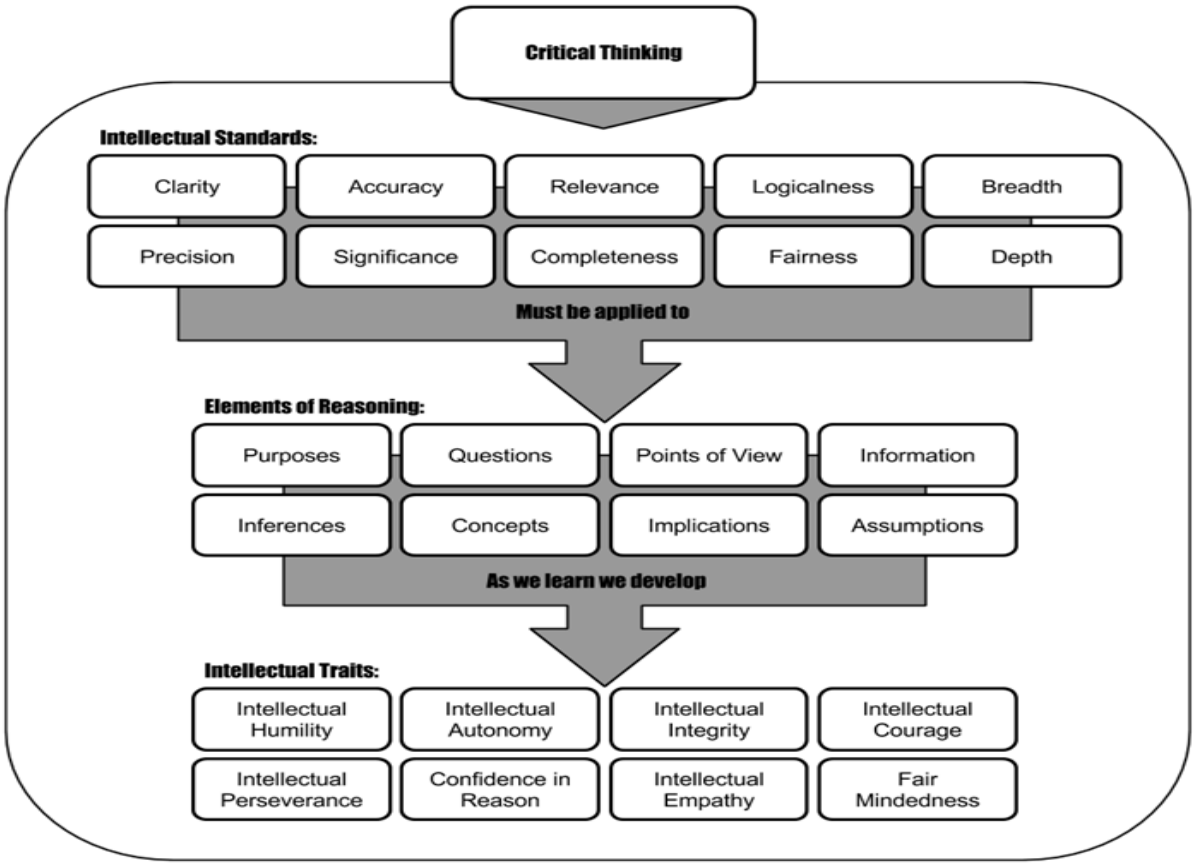
The Paul-Elder model, developed originally by Paul (Paul, 1993; Paul et al., 1993) and further refined by both Paul and Elder (Paul & Elder, 2001), is associated with the Foundation for Critical Thinking (www.criticalthinking.org), an educational non-profit organization which promotes essential change in education and society through the cultivation of fair-minded critical thinking (Foundation for Critical Thinking, n.d.). The Paul-Elder model is based on the following working definition of critical thinking as:

> that mode of thinking — about any subject, content, or problem — in which the thinker improves the quality of his or her thinking by skillfully analyzing, assessing, and reconstructing it. Critical thinking is

> self-directed, self-disciplined, self-monitored, and self-corrective thinking. It presupposes assent to rigorous standards of excellence and mindful command of their use. It entails effective communication and problem-solving abilities, as well as a commitment to overcome our native egocentrism and sociocentrism. (Paul & Elder, 2005)

The Paul-Elder model divides critical thinking into three key components: elements of reasoning, intellectual standards and intellectual traits. The elements of reasoning are universal elements that inform and describe all reasoning or thought. The intellectual standards are standards applied to elements of reasoning or thought to interpret or assess quality. Lastly, the intellectual traits are desired traits or characteristics of a skilled practitioner of critical thinking. These three components are interrelated and each contributes to the development of a critical thinker. In the Paul-Elder model, critical thinkers apply the intellectual standards to the elements of reasoning in order to develop intellectual traits (Figure 2). There are two essential dimensions of thinking that students need to master in order to learn how to upgrade their thinking. They need to be able to identify the component parts of their thinking, and they need to be able to assess their use of these parts of thinking (Paul et al., 1996). These two essential dimensions, in concert with the intellectual standards, elements of thought and intellectual traits, can be organized into a rubric for the evaluation of critical thinking.

Figure 2: The Paul-Elder Model



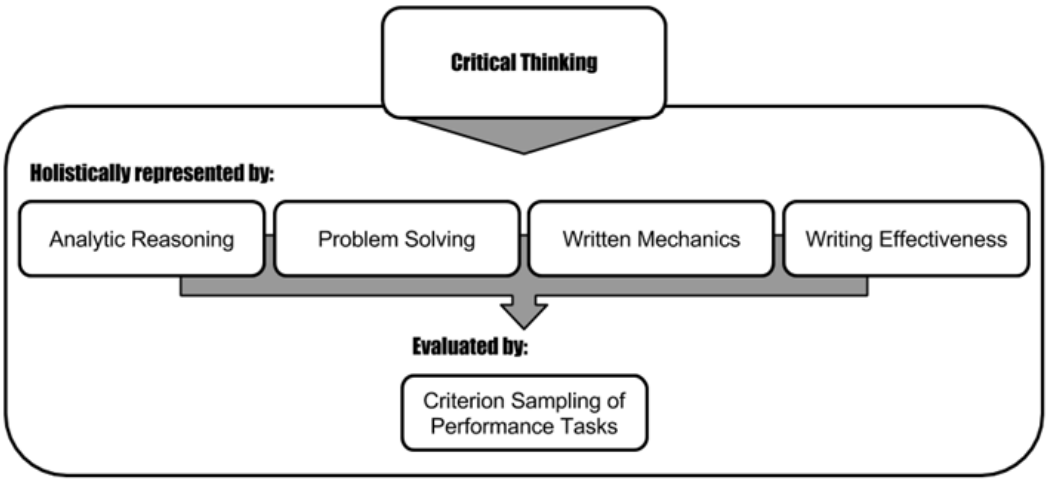
CLA Model

The CLA model was developed for the holistic evaluation of critical thinking through problem solving. This model is used solely for assessment and was developed for the Collegiate Learning Assessment, a test that is discussed in the next section. The CLA model is not an explicit framework, unlike the Paul-Elder or Cornell-Illinois models, which reduces critical thinking into constituent parts. Rather, the CLA views critical thinking in the broadest sense, as summarized by (Bok, 2006):

> The ability to think critically—ask pertinent questions, recognize and define problems, identify arguments on all sides of an issue, search for and use relevant data and arrive in the end at carefully reasoned judgments—is the indispensable means of making effective use of information and knowledge.

The CLA model holds that critical thinking assessment is best approached holistically, arguing that critical thinking cannot be broken down into component parts and measured. Instead, the CLA views the larger construct of critical thinking as being closely connected to and represented by several criteria or skills that students utilize in their responses on the test, as shown in Figure 3.

Figure 3: The CLA Model



The CLA model relies on a criterion sampling approach that is relatively straightforward and seeks to determine the abilities of a student by sampling tasks from the domain in which the student is to be measured, observing their response and inferring performance and learning on the larger construct. Shavelson (2008) explains criterion sampling by using the example of driving a car:

> For example, if you want to know whether a person not only knows the laws that govern driving a car but also if she can actually drive a car, don't just give her a multiple-choice test. Rather, also administer a driving test with a sample of tasks from the general driving domain such as starting the car, pulling into traffic, turning right and left in traffic, backing up, and parking. Based on this sample of performance, it is possible to draw valid inferences about her driving performance more generally. (Shavelson, 2008)

The CLA follows the criterion sampling approach by presenting students with holistic, real-world problems. Through these problems, it samples tasks and collects students' responses, which are then graded according

to a set of generic skills and formed into rubrics. In order to generate a successful response to the task, students would have to apply problem solving successfully, reason analytically, and write convincingly and effectively. Since these are all underlying components of critical thinking as defined by the CLA model, critical thinking ability can thus be inferred from student responses to test questions.

Critical Thinking Assessments

There are numerous critical thinking assessments available, each constructed from a different framework of critical thinking. This leads to a wide variety in the style and application of each test, each with their own strengths and weaknesses. This allows an instructor a great deal of latitude to select a test that best suits their own definition of critical thinking or to adopt a framework for instruction that has a corresponding test to maintain alignment between instruction and assessment. The reviews of these assessments in the following section are included to provide information regarding each assessment's respective strengths and weaknesses, and to comment on the assessments selected for use in this study. For additional information about other critical thinking assessments, please consult Appendix 10.

Cornell Critical Thinking Test Level Z

The Cornell Critical Thinking Test Level Z (CLZ) is a 52-item, multiple choice test aimed at gifted high school students and university students. The CLZ was developed by Robert Ennis, Jason Millman and Thomas Tomko, with the most recent edition of the test being released in 2005 (Ennis et al., 1985). The CLZ measures seven aspects of critical thinking consistent with the Cornell-Illinois model (Ennis et al., 1985):

1) Induction
2) Deduction
3) Observation & Credibility
4) Assumptions
5) Meaning & Fallacies

Two scoring options exist for the CLZ: a "rights only" scoring, which counts correct responses, and a "rights minus half wrongs" scoring, which penalizes students for incorrect answers. Both options are valid for scoring the test, with the authors recommending the latter scoring as guessing is not consistent with meaningful thinking habits (Ennis et al., 1985). The test can be completed either on paper or online. Additional sub-scores can be calculated for the seven individual aspects of the Cornell/Illinois model, as listed above. The CLZ has been validated by several studies, with observed validity measures ranging from α=0.5 to 0.87 (Ennis et al., 1985; Frisby, 1992).

There are some potential issues with using a multiple choice assessment of CTS, arising from the fact that the test does not assess dispositional aspects of critical thinking. Multiple choice CT assessments in general have been criticized as tests assessing verbal and quantitative knowledge and not critical thinking, since the format prevents test-takers from applying CTS to develop their own solution to the problem (Abrami et al., 2008; Halpern, 2003; Ku, 2009). Additionally, multiple choice tests can only narrowly assess a single concept of thought in a question, whereas the real-world application of critical thinking typically employs a wide variety of concepts and skills (Bensley & Murtagh, 2011; Ku, 2009). More specifically, criticisms of the CLZ point to the low validity scores associated with the test and to potential gender bias issues with test items (Stein et al., 2003).

Collegiate Learning Assessment

The Collegiate Learning Assessment (CLA) is comprised of a set of web-administered task assignments targeted to first-year and fourth-year university students. The CLA was developed and is administered by the

Council for Aid to Education (CAE) and is formulated around the CLA model of critical thinking and problem solving. The CLA is scored by an automated system using a series of grading rubrics (Council for Aid to Education, n.d.; Shavelson, 2008). Both overall scores and sub-scores are compiled from aspects of critical thinking, including:

1) Analytic reasoning
2) Problem solving
3) Writing mechanics
4) Writing effectiveness

The CLA consists of two distinct tasks, of which students generally complete one: a "performance task" and an "analytic writing task" containing two subtasks, "make an argument" and "critique an argument." The CLA has a high reported validity (α=0.80) but only at the institutional level, as the CLA displays poor validity at the student level (α=0.45) (Klein et al., 2009; Klein, Benjamin, Shavelson & Bolus, 2007). There has also been some concern raised about the holistic assessment methods of the test not accurately measuring the component cognitive skills of critical thinking, and some critique on the grading method of the CLA. This, alongside the cost of the CLA and the narrow administration window for the test, are potential barriers to its use (Possin, 2013). Despite these potential challenges, the CLA is a comprehensive assessment, with the tasks requiring the identification, integration and use of multiple skills and critical thinking concepts in both tasks. Additionally, the CAE has recently addressed the student-level reliability issue through the development of the new CLA+, which increases the student-level validity to α=0.85-0.87 (Zahner, 2013).

International Critical Thinking Essay Test

The International Critical Thinking Essay Test (ICTET) was developed by Richard Paul and Linda Elder of the Foundation for Critical Thinking. The ICTET is an essay-style test designed to provide an assessment of the fundamentals of critical thinking. The ICTET has two areas of focus. The first is to provide a reasonable way to measure CTS, while the second is to provide a test instrument that stimulates the faculty to teach their discipline in a manner that fosters critical thinking in the students (Paul & Elder, 2010). The ICTET is divided into two separate forms: an analysis of a writing prompt and an assessment of the writing prompt. In the analysis segment (Form A) of the test, the student must accurately identify the elements of reasoning within a prompt. In the assessment segment of the test (Form B), the student must critically analyze and evaluate the reasoning used in the original prompt. Student responses are graded according to a rubric based on the elements of reasoning that comprise Paul's model of critical thinking (Paul & Elder, 2005):

1) Purpose
2) Questions
3) Information
4) Conclusions
5) Concepts
6) Assumptions
7) Implication
8) Point of view

Both a total score and related sub-scores can be calculated. The ICTET was authored to have high consequential validity, such that the consequence of using the test would be significant and highly visible to instructors (Paul & Elder, 2007). This encourages discipline-specific adoption of critical thinking and the redevelopment of curriculum that "teach to the test." Statistically speaking, criterion and concurrent validity, as well as reliability measures for the ICTET, have yet to be determined due to both the lack of a universal criterion to measure CTS and the relative infancy of the instrument. Content and construct validity for the ICTET are addressed through the use of discipline-specific prompts for the test and the well-established critical thinking model. That is to say, the test measures what it is supposed to measure through the use of

the Paul-Elder model as a framework and the test accurately assesses CTS through the use of prompts containing subject matter relevant to the discipline.

There are a few potential challenges that may be encountered with this style of test. First, the prompts task students with the recall-based identification and evaluation of the elements of thought. While these skills are of vital importance within critical thinking, the specific prompts cannot evaluate how students apply CTS in a real-world setting (Bensley & Murtagh, 2011; Butler & Butler, 2012; Butler et al., 2012; Halpern, 2006). Second, the specificity of the questions may limit the breadth of response in test-takers, leading to a reduced inclination to engage in critical thinking (Taube, 1997). Lastly, inter-rater reliability (IRR) in this style of test is a potential issue that should be considered when administering the test on a large scale (Shavelson, Baxter & Gao, 1993).

Instructional context

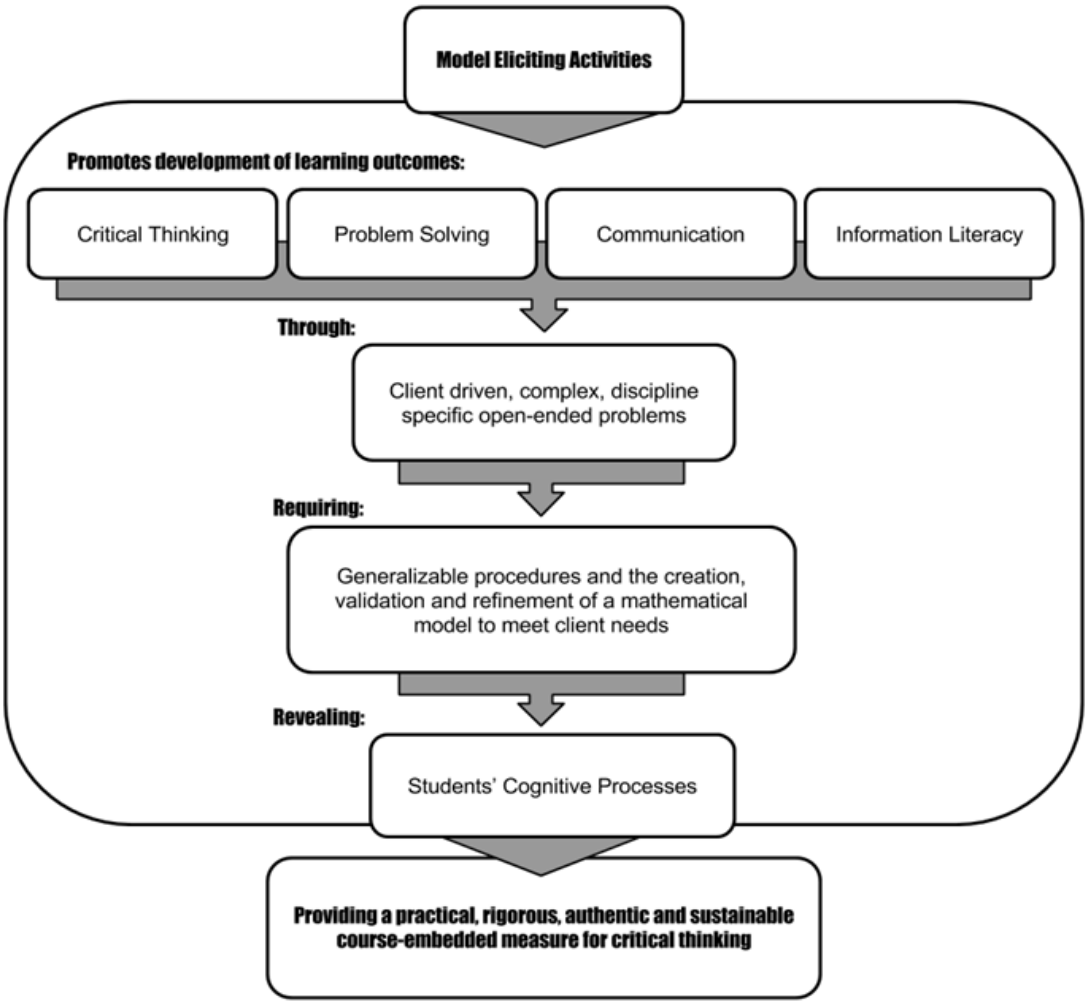
APSC 100 Module 1

The study was set in a project-based course in the first year of the undergraduate engineering program at Queen's University. The program has a common first year; all 650 first-year engineering students take the same courses before being allowed free choice between ten engineering programs in their second year. In the first semester of the program, students take courses in statics, chemistry, earth systems, engineering graphics and calculus, and a full-year course, APSC 100, focused on engineering design and practice, in which this study is situated Appendix 1.

APSC 100 is a team- and project-based course designed to promote a sense of curiosity about engineering and provide opportunity for students to develop judgment and problem solving skills by tackling tasks that emulate engineering activities. The course is divided into three modules: module 1 on problem analysis and modeling; module 2 on experimentation and measurement; and module 3 on engineering design. Each of these is one semester long and equivalent in weight to a standard one-semester engineering course (Frank, Strong, Sellens & Clapham, 2012; Frank, Strong & Sellens, 2011). This study was embedded into the delivery of the problem analysis and modeling module (APSC 100 Module 1). APSC 100 Module 1 is a semester-long integrative experience that uses concepts from engineering sciences, natural sciences and mathematics courses to solve complex open-ended problems. The course is structured around three complex problems known as model eliciting activities (MEAs) that were addressed sequentially in three-week blocks over the semester.

The situations described in the MEAs require students to create and use a mathematical model of a physical system using a numerical computation tool (MATLAB) and to deal with professional issues including ethical dilemmas, conflicting information and incorrect/missing information. While each MEA requires students to employ different areas of subject knowledge, students are taught to approach all three MEAs using critical thinking skills. For example, students are guided to draw concept maps, question the credibility of information sources, incorporate a range of factors into their decision-making and consider the implications of their conclusions. These skills are what Paul calls "elements" of critical thinking – invaluable thinking processes involved in any complex problem-solving activity (Paul & Elder, 2005). Most importantly, the MEAs provide a practical, course-embedded means for the authentic, rigorous and sustainable measure of critical thinking development and assessing critical thinking skills illustrated below in Figure 4.

Figure 4: Critical Thinking Assessment Using MEAs



MEAs have been used in the course for the past three years (Frank & Kaupp, 2012). In the 2010-2011 academic year at Queen's, engineering students were observed to improve in their ability to solve complex problems and meet course expectations over the year. Student scores on MEA rubric elements of problem definition and creating solution process generally improved over the year, and more detailed analysis on these results is being conducted for other purposes. Students also viewed these activities as beneficial to their development; in a course survey, most students reported that the MEAs improved their skills in solving open-ended problems.

Learning Outcomes & Course Structure

The module learning outcomes were to:

1) Apply a prescribed process for solving complex contextualized client-driven problems (ill-defined, multiple constraints, problems, unknown information)
2) Create and apply appropriate quantitative model and analysis to solve problems
3) Effectively communicate technical information following a prescribed format and using standard grammar and mechanics
4) Apply concepts including occupational health and safety principles, economics, law, and equity to engineering problems
5) Identify and resolve a simple ethical dilemma by applying professional codes of ethics and engineering standards
6) Apply critical and creative thinking principles to solve contextualized problems
7) Apply a numerical modeling tool (MATLAB) to create a model used for solving complex problems

The module was structured to help students develop confidence and skill in solving complex engineering problems – problems for which all information is not known, in which there is ambiguity, and where the goals are not necessarily clearly defined. In most weeks, the one-hour-per-week lecture followed a structure similar to the one described below:

1) The instructor presented a recent problem or news article related to the lecture objective; in some cases, the students responded in teams to a problem posed using a web-based audience response system (ARS)
2) The instructor presented or reviewed the problem being solved during the three-week session
3) The instructor led a short discussion on a topic related to the problem being studied
4) Students worked on some component of the problem in small groups; in many cases, this included an open-text question answered using the web-based ARS

The students also attended a two-hour studio each week. The studio opened with a short quiz on prior reading and/or online videos on MATLAB, followed by a short discussion of some MATLAB concept. The majority of the studio focused on a problem that contributed to the current MEA. Students received a small mark each week for completing the task, encouraging them to keep up with the course material.

MEA Characteristics and Outcomes

The three MEAs were:

1) MEA 1: Cable ferry failure (weeks 2-4): This problem focused on the failure of a cable ferry (see Appendix 2. MEA 1 Objectives).
2) MEA 2: Wind turbine design turbine (weeks 6-8): This problem focused on the analysis and design of a wind turbine (see Appendix 3. MEA 2 Objectives)
3) MEA 3: Building heat loss (weeks 9, 11, 12): This problem focused on the design of the insulation for a net zero home (see Appendix 4. MEA 3 Objectives)).

Each MEA required students to develop a model of a physical system that could solve a problem presented by a fictitious client, write MATLAB code to implement the model, and evaluate their report against three to five of nine critical thinking elements identified in the Paul and Elder critical thinking model (clarity, accuracy, relevance, logicalness, breadth, precision, significance, completeness, fairness and depth)(Paul & Elder, 2005). Table 1 shows the elements embedded into each MEA. Critical thinking elements were explicitly targeted in all three exercises by discussing principles in class, using in-class activities and embedding them into the MEA requirements. During one of the lectures focused on MEA 1, students created lists of the kinds

of questions that should be asked when investigating an accident, which led to a discussion about asking questions. In their final deliverable they were required to identify the kinds of questions they would ask upon arriving at an accident investigation site. In MEA 2, students were required to summarize relevant information, including an assessment of its source and credibility, uncertainties and biases.

Table 1: Characteristics of the Individual MEAs

Category	MEA 1	MEA 2	MEA 3
Technical	Stress and strain, drag	Fluid flow, lift	Heat transfer
Design	Problem definition, concept mapping	Decision making (e.g., weighted evaluation matrices)	Decision making (e.g., weighted evaluation matrices)
Professional	Safety, risk assessment, concept maps	Associations, codes and standards	Economics, codes of ethics, equity
Critical thinking	Asking questions, uncertainty in information, identifying erroneous or conflicting information	Assessing information credibility, argumentation, assumptions, inferences	Bias, inferences
Communications	Report format, English usage, argumentation	Report format, English usage, argumentation, concision	Report format, English usage, argumentation, concision

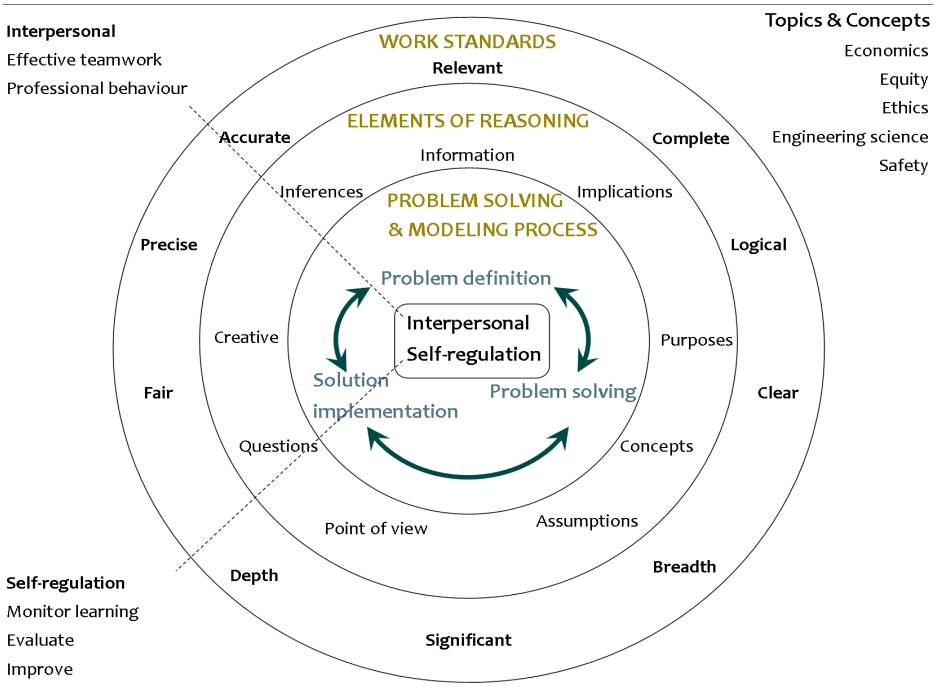
Each MEA had five common outcomes and two task-specific outcomes. The common outcomes, illustrated below in Table 2, are based on the process by which students generate their solutions, while the specific outcomes for each MEA address aspects of professional practise. Both specific and common outcomes are organized into a rubric, with established performance criteria for consistent assessment across a variety of scenarios.

Table 2: Common MEA Outcomes

MEA Outcome	Description
Information Summary	Accurately summarizes relevant information pertaining to the problem (background, contextual, content and methodological information), and includes an assessment of the credibility, uncertainty and biases of the information and its source.
Model Generation	Creates, compares and contrasts quantitative models in MATLAB using approximations and assumptions generated from a justified problem solving process supported by information.
Model Results	Evaluates validity of both the model and its results for error and uncertainty, drawing well-supported conclusions to support and strengthen the solution.
Critical Evaluation	Critically assesses conclusions on the basis of intellectual standards of clarity, precision, accuracy, relevance, logicalness, breadth, depth, significance, completeness and fairness.
Argumentation	Rationally supports claims and conclusions with data and comprehensive description of the context in which they apply.
Communication	Information is clearly and concisely presented, demonstrating consistent use of important engineering and technical reporting conventions, including organization, content, presentation and stylistic choices.

The conceptual framework for the module is shown in Figure 5. Course activities were designed to introduce teamwork skills, to encourage students in their learning, and to use processes to solve a problem and continue improving them (self-regulation). These are shown in the centre of the circles below. Students were encouraged to apply the elements of reasoning to the problems they solved (shown in the middle ring below), some of which were discussed in class. Students evaluated their own submission against the work standards shown in the outermost ring.

Figure 5: Conceptual Framework Used for APSC100



Method and Procedure

Overview of Study Design and Variables

During the fall semester of the 2012/2013 academic year, three instruments were used to evaluate the critical thinking skills (CTS) of first-year engineering students. The ICTET and CLZ critical thinking assessments were used as both a pre- and post-test in order to benchmark the CTS of the incoming first-year students and to determine the effectiveness of MEA instruction at developing student CTS in APSC 100. In our study, the MEA-integrated curriculum is the independent variable and students' CTS are the dependent variable.

All the first-year students were invited to participate in the broad study, which was granted approval by Queen's General Research Ethics Board (GREB). Stratified sampling was used to assign various pre and post instruments according to a within-subjects design. These assessments of CTS are part of the course requirements, so the participation rate was close to 100%.

Given similarities in incoming student characteristics and/or learning environment, we had planned to collaborate with engineering programs at three other universities to evaluate critical thinking using one of the instruments at the start and end of the 2012/2013 fall semester. Unfortunately, due to recruitment issues, our collaborators were unable to obtain a sufficient number of volunteers to provide a statistically comparable sample to serve as a control group. First-year student volunteers from the physics department at Queen's, who followed a similar curriculum, were solicited for participation in the study to serve as a control group. Unfortunately, recruitment issues again resurfaced and only a small number of participants volunteered. These volunteers served as a control group for think aloud interviews.

Study Instruments

For our study, we carefully considered a variety of factors before selecting the instruments to assess CTS. After deliberating on issues concerning purchasing, administration and scoring of each test, we selected the Cornell Critical Thinking Test Level Z, the International Critical Thinking Essay Test, and think aloud protocols as the instruments to assess critical thinking skills. The Collegiate Learning Assessment was also used in this study, in conjunction with another study funded by the Higher Education Quality Council of Ontario (HEQCO) at Queen's University. Each test was administered as a pre- and post-test, with the exception of the CLA, which was used solely as a pre-test. Groups for each test were randomly created from the incoming students and further divided into cohorts with different pre- and post-test pairings to investigate potential test-retest effects, as illustrated below in Table 3. Additionally, some students were asked to participate in think aloud problems.

Table 3: Measurement Approach, Cohort Grouping and Study Instruments

Approach	Cohorts	N	Pre	Post
Collegiate Learning Assessment (CLA)	A	151	Pre-test Survey	
Cornell Critical Thinking Test Level Z (CLZ)	B	96	Cornell Level Z Pre-test survey	Cornell Level Z Post-test survey
	C	84	ICTET Pre-test survey	Cornell Level Z Post-test survey
International Critical Thinking Essay Test (ICTET)	D	109	ICTET Pre-test survey	ICTET Post-test survey
	E	101	Cornell Level Z Pre-test survey	ICTET Post-test survey
Think aloud protocol	Control	3	Mini-MEA A	Mini-MEA B Exit Interview
	Experimental	2		
MEA scores	All	542	Evaluated in Oct, Nov and Dec by graders	

The initial pre-test benchmarking of CTS took place at the beginning of the 2012/2013 fall semester, prior to any critical thinking or MEA instruction. The final post-test measurement of CTS took place at the end of the 2012/2013 fall semester, after the conclusion of critical thinking or MEA instruction. Students were also invited to participate in two think aloud activities, also run as pre- and post-tests to measure students' CTS.

In the following figures depicting the different study groups, **X** represents the intervention, (i.e., the MEA-integrated curriculum), and the measurement or observation of CTS (i.e., a pre- or post-test, interview session) is explicitly labeled.

CLA Group

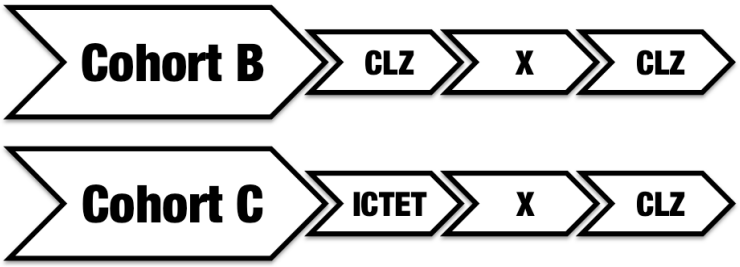
The CLA group is comprised of 151 students who wrote the CLA as a pre-test and no associated post-test, as illustrated in Figure 6. As previously stated, this was due to the CLA being part of another HEQCO-funded study, and also because the time commitment required to take the CLA was considerable and the testing window for the CLA did not occur within the study time frame. However, we hypothesize that the CLA scores can possibly be used as a predictor of MEA performance and have presented sub-scale correlations.

Figure 6: Structure of the CLA Group

Cornell Level Z Group

The CLZ group is further divided into two sections, Cohort B and Cohort C (97 and 84 students, respectively). Cohort B will take the CLZ as a pre-test and then as a post-test. Cohort C will take the ICTET as a pre-test and then the CLZ as a post-test (Figure 7). The post-test results of Cohort B should not differ significantly from the post-test results for Cohort C, given that students from both sections are comparable. However, if results show otherwise, there may be test-retest effects in the CLZ.

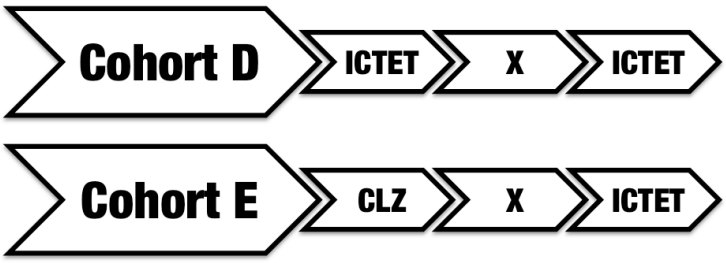
Figure 7: Structure of the Cornell Level Z Group



ICTET Group

The ICTET group is further divided into two sections, Cohort D and Cohort E (109 and 101 students, respectively). Cohort D will take the ICTET as a pre-test and then as a post-test. Cohort E will take the CLZ as a pre-test and then the ICTET as a post-test (Figure 8). The post-test results of Cohort D should not differ significantly from the post-test results of Cohort E, given that students from both sections are comparable. However, if results show otherwise, there may be test-retest effects with the ICTET.

Figure 8: Structure of the ICTET Group

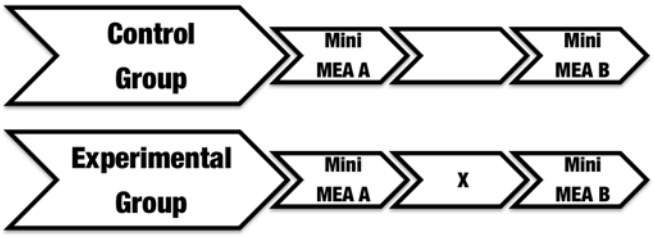


Think Aloud Groups

A group of three students, drawn from the first-year engineer student pool, will be assigned to the experimental condition, i.e., participation in the MEA-integrated curriculum. A group of three students, drawn from the first-year physics student pool, will be assigned to the control condition, i.e., no exposure to a MEA-integrated curriculum (Figure 9).

Both groups of students will be asked to solve an open-ended task (Mini-MEA A, Appendix 5)) as a pre-test and then another open-ended task as a post-test (Mini-MEA B, Appendix 6). Ideally, members from both conditions should remain the same from the pre-test to the post-test. The post-test observation of the experimental group should differ significantly from the post-test observation control group in terms of the quality of students' CTS, as students in the experimental condition will have received training to use their CTS to solve complex tasks.

Figure 9: Think Aloud Group Division



Survey Creation

In order to measure intrinsic and extrinsic factors affecting critical thinking and to investigate student perceptions of CTS development, we used two separate surveys administered during the pre- and post-tests. Survey responses were collected and coded using Scantron test answer cards and an EZData scanner with Remark OMR scanning software.

The pre-test survey assessed student motivation and English proficiency, and consisted of ten questions targeting motivation and three questions regarding English proficiency (Appendix 7). The ten motivation questions were drawn from the Academic Motivation Scale (Vallerand et al., 1992), consisting of three questions targeting intrinsic motivation, three questions targeting integrated and identified motivation, and four questions targeting external and introjected motivation, grouped according to Deci & Ryan's self-determination theory (Deci & Ryan, 1985; 2010; Ryan & Deci, 2000). The three questions for English proficiency asked students if they had to write an English proficiency exam for admission, which test they

wrote and their test results, as English proficiency has been shown to be a factor affecting critical thinking skills (Dunham, 1997; Rashid & Hashim, 2008).

The post-test survey assessed external factors affecting critical thinking, such as workload, and student perception of CTS development with respect to both general university experience and specific APSC 100 Module 1 experiences (Appendix 8), Three questions targeted student workload in individual work, group-based teamwork and group-based individual work. Five questions targeted university experience, APSC 100 Module 1 experiences relating to MEA individual and group work, lectures and interaction with course personnel. Four questions targeted student involvement in transformation and synthesis of ideas from different courses, tutoring and discussion of ideas from different courses with faculty, family or peers. The last item on the post survey targeted student perceptions of first-year courses and extracurricular activities on CTS development and asked them to rank those experiences from most to least important.

Methodology

Test Administration

With the CTS assessment embedded within the course experience, the administration of the tests occurred during scheduled lab sessions in week 1 and week 12 of the fall semester. The tests other than the CLA were not timed and students had full use of the time slot to finish the tests and surveys. All participants finished in under an hour for the CLZ and in under 90 minutes for the ICTET. Proctors supervised testing and were instructed only to help with basic questions and not to respond to questions that would influence the students' responses.

Test Scoring

The CLA was scored using the automated scoring method developed by the CAE, with the resultant scores and sub-scale data provided. The maximum achievable score for the CLA is 1400, with each sub-scale maximum achievable score being 6.

The CLZ tests were scored using Scantron EZdata systems, with a maximum achievable score of 52, using the authors' suggestions and the "rights only" method selected for the overall score (Ennis & Weir, 1985). This scoring method was selected as it is the mode with which students would be most familiar. Sub-scale data were scored using the same "rights only" method, with each sub-scale possessing a different maximum achievable score dependent upon the number of questions comprising each respective scale (Deduction: 10, Observations & Credibility: 4, Meaning & Fallacies: 11, Induction: 17, Assumptions: 10).

The ICTET scoring required comprehensive training, provided by the Foundation for Critical Thinking under the "train-the-trainer" model, in order to establish inter-rater reliability (IRR). IRR was assessed using Krippendorff's alpha (Hayes & Krippendorff, 2007), measured at the end of the sessions with the Foundation for Critical Thinking and after the grader training sessions prior to grading the ICTET. The maximum achievable score for the ICTET is 80, with the maximum achievable score for each sub-scale being 10.

MEA Scoring

Due to the large number of students in APSC 100 Module 1, there were nine graders responsible for scoring MEAs, with graders consistently marking the same students across all of the MEAs. In order to establish and preserve grader reliability, a calibration session was conducted prior to grading each MEA. During this session, the course instructor and graders evaluated a subset of student submissions until consensus was established and each grader was comfortable scoring according to the criteria set forth in the rubric. Each

MEA was graded using a MEA-specific rubric that consisted of the six common outcomes and two MEA specific outcomes (Appendix 2, Appendix 3, Appendix 4).

After the calibration session, the graders then graded all student MEA submissions independently. The mean and standard error of MEA scores and sub-scores for each grader were monitored, with graders validating outlier student submissions to the course instructor to ensure consistency, accuracy and to combat grade inflation.

Think Aloud Sessions

For our study, we constructed a task that emulated the MEA activities used in APSC 100, although on a much smaller scale. These mini-MEA activities are a loosely defined, open-ended scenario with a system based on a fundamental physics problem. The scenario asked the subjects to evaluate the system and provide solutions for a specific request, along with any additional safety recommendations they saw fit. Subjects were provided with supplemental information of varying authenticity, reliability and credibility to help them with their recommendations. Subjects had one hour to solve the problem while "thinking aloud" their solution. At the end of the hour, the subjects were asked to present their recommendations.

At the beginning of the sessions, students were introduced to the facilitator and the expectations for the think aloud exercises. The subjects were then taken through a warm-up exercise in which they had to provide improvements for a common appliance or simple machine (e.g., bicycle, washing machine) and present them to the facilitator. Following the conclusion of the warm-up activity, the actual think aloud began.

Think Aloud Pre-Test

The pre-test think aloud tasked subjects to provide safety recommendations to a city council regarding its proposed toboggan hill for a winter festival. An email to the team from the city council was provided, outlining its request and the details of the problem. Subjects were provided with supplemental material to help address the problem:

1) An independent opinion on toboggan safety
2) A newspaper article on tobogganing safety
3) A student-created list of friction coefficients
4) A textbook excerpt of friction coefficients
5) Information about average mass of American children and adults
6) A scientific article on human tolerance and crash survivability
7) A physics equations sheet

The primary areas of contention for this scenario concerned the pedigree of the supplemental information and that the initial parameters provided resulted in very unsafe slope conditions. The initial email from the city council included an unsolicited reference from a councilor with a military background regarding human impact tolerance. This reference, alongside the independent opinion on toboggan safety, parts of the newspaper article and the student-created list of friction coefficients, should have been identified in some way by the subjects as potentially unreliable. Ultimately, the subjects should have realized that the hill, under its initial conditions, was very unsafe and should have proposed recommendations resulting from informed analysis.

Think Aloud Post-Test

The post-test think aloud tasked subjects to provide safety recommendations to an amusement park, FunZone Amusements, regarding its proposed prototype rollercoaster. An email to the team from the

company was provided, outlining its request and the details of the problem. Subjects were provided with supplemental material to help address the problem:

1) Summary of American Society of Testing and Materials (ASTM) Standards on Amusement Park Device Design
2) Scholarly articles on rollercoasters and G-forces
3) Reports and articles on roller coaster safety
4) A physics equation sheet

The primary areas of contention for this scenario concerned the conflicting information of the supplemental material, the interpretation of a professional standard and the rollercoaster parameters. The initial email from the company included some unsolicited guidelines regarding track dimensions and average velocity. These reference dimensions led to potentially unsafe conditions for passengers. Upon further analysis of supplemental materials, students should have integrated multiple references to form a cogent argument. The question regarding re-using existing carts and restraints also challenged the students' ability to interpret information on a technical chart and classify their solution by a professional standard.

Data and Statistical Analyses

All study data were anonymized then analyzed using IBM SPSS Statistics 21 and a variety of parametric and nonparametric techniques. More detailed information on the analysis techniques for specific study items are provided in their respective sections.

Critical Thinking Tests

In order to assess gains in CTS, pre and post scores and sub-scores measuring specific critical thinking elements (Table 4) were compared using paired different *t*-tests for all testing groups. Independent *t*-tests comparing post-test scores and difference scores (post-pre) across cohorts with similar post-tests were conducted as a measure of validating experimental design and as an element of test reliability. Additional measures of test reliability included Cronbach's alpha and pre-post score correlations using groups with the consistent pre-post-tests.

The sub-scores for the Cornell Level Z featured duplicate measures for the elements of induction and assumptions of the critical thinking model used by the test. Cronbach's alpha was calculated in order to evaluate the internal reliability of this grouping.

Table 4: Critical Thinking Test Sub-Score Items

Test	Sub-Score Items							
CLA	Analytic reasoning	Writing effectiveness	Writing mechanics	Problem solving				
CLZ	Induction	Deduction	Observation & credibility	Meaning & fallacies	Assumptions			
ICTET	Purpose	Questions	Information	Conclusions	Concepts	Assumptions	Implications	Point of view

Surveys

Survey results were used to assess the motivation of students taking the test (using the Academic Motivation Survey questions (Vallerand et al., 1992)) and to assess students' perceptions of specific course experiences and external factors developing CTS. The survey questions pertaining to motivation were divided into categories measuring intrinsic, external and integrated motivation, and the five-point Likert scale responses were recoded to a three-item scale for further comparison and analysis. Questions pertaining to external factors affecting critical thinking were processed in a similar fashion and reduced to a two-item scale where necessary. All scales and responses are illustrated below in Table 5. Students were also asked to rank specific course experiences within the first-year engineering curriculum and extracurricular activities according to their importance in developing CTS skills ("What do you think has contributed to developing the type of thinking used for the critical thinking post-test over the past three months?").

An independent *t*-test was used to investigate the differences between English as a second language (ESL) and English as a first language (EFL) status on CTS, measured by post-test performance. For re-coded scales, Cronbach's alpha was used to assess consistency with the parent five-item scale. The Kruskal-Wallis H-test was used to determine differences between CTS and respective sub-scores (measured by the post-test score) and motivation or factors affecting CTS.

Table 5: Descriptions of Survey Likert Scales (Appendices 7 and 8)

Likert Scale	1	2	3	4	5
5 point	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
5 point workload	0-5 Hours	6-10 Hours	11-20 Hours	20-30 Hours	30+ Hours
4 point participation	Very Often	Often	Sometimes	Never	
3 point	Disagree	Neutral	Agree		
2 point	Disagree	Agree			

Model Eliciting Activities (MEA)

A repeated measures ANOVA was conducted to determine differences in the three MEAs and respective sub-scores, illustrated in Table 6, over the course of the fall semester. Motivational and additional factors affecting MEA performance were also explored, using the recoded three-point motivation categories, with a Kruskal-Wallis H-test. The relationship between MEA scores and critical thinking test scores (post-test scores for the CLZ and ICTET, pre-test score for the CLA) and between MEA and critical thinking test sub-scores was assessed using the Pearson's product-moment correlation coefficient. Associations between MEA sub-scores and grades were also explored to assess the internal consistency and correlation of MEA sub-scores and grades.

Table 6: MEA General and Specific Sub-Scores

	Sub-Scores							
MEA	**Common**						**MEA-Specific**	
1	Information summary	Model generation	Model results	Critical evaluation	Argumentation	Communication	Proposed process	Safety analysis
2	Information summary	Model generation	Model results	Critical evaluation	Argumentation	Communication	Cover letter	Power generation alternatives
3	Information summary	Model generation	Model results	Critical evaluation	Argumentation	Communication	Cover letter	Ethical reasoning

Think Aloud Sessions

Videos of the pre and post think aloud interview sessions were transcribed and annotated. The research team divided the think aloud transcripts into five segments, with each segment consisting of a particular issue that the group addressed. The research team selected the safety recommendations segment for further analysis, as this unit was thought to display the greatest amount of elements of critical thinking. The safety recommendations segment was then coded for elements of critical thinking, corresponding to the Paul-Elder model, as illustrated in Table 7. The quality of each element was assessed, with codes being separated into those of acceptable and unacceptable quality. Quality was assessed by comparing each code against the indicator for the corresponding element; a negative response to the indicator was assigned to low quality, whereas a positive response was assigned to high quality.

Once coded, the safety recommendations unit was analyzed for common themes and notable differences between pre- and post-tests in the experimental and control group. The two groups were also compared and contrasted.

Table 7: Coding by the Paul-Elder Critical Thinking Model

Elements	Indicators
Purpose	Did the participants clarify the purpose of the task given? Did the participants ask about the purpose of supplemental material given?
Questions	Did the participants clarify what questions they were supposed to answer? Did the participants have a plan of action to answer the questions that they identified?
Points of View	Did the participants ask whether there were other relevant viewpoints that should be considered? Did the participants ask about the viewpoints expressed in supplemental material given?
Assumptions	Did the participants identify or question their own assumptions? Did the participants ask the extent to which their own assumptions were valid? Did the participants identify assumptions made by authors of supplemental material provided?
Information	Did the participants identify what information they were lacking? Did the participants ask how they could get the information that they needed? Did the participants question the source of supporting information?
Concepts	Did the participants ask whether the concept or theory considered applicable to the given situation? Did the participants ask whether there was another theory or principle that would better explain the given situation?
Conclusions	)id the participants ask whether their conclusions were supported by their analysis or supplemental material provided? Did the participants ask whether there were alternative conclusions that would also fit the data?

Results

Pre-Post-Testing Results

CLA Group: Cohort A

The summary of the critical thinking pre-test scores for Cohort A is shown below in Table 8. Students in the cohort (*M*=1204, *SD*=161) scored higher than the average score of American schools (*M*=1050, *SD*=97) participating in the CLA administration (156 institutions). Additionally, the CLA group's performance on the four sub-scores of analytic reasoning, problem solving, writing mechanics and writing effectiveness was higher than the American (CAE, 2012).

Table 8: CLA Group Mean Scores and Sub-Scores

	N	Test Score		Analytic Reasoning		Problem Solving		Writing Mechanics		Writing Effectiveness	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cohort A	151	1204	161	3.6	0.9	3.5	0.9	3.8	0.8	3.7	0.9
American Institutions		1050	97	2.9	0.8.	2.7	0.8	3.2	0.9	2.9	0.8

Cornell Level Z Group: Cohort B

The students of Cohort B wrote the CLZ for both the pre- and post-test. The summary of the critical thinking pre-test and post-test scores is illustrated below in Table 9. Student performance on the pre-test *(M*=30.90, *SD*=4.51) and post-test (*M*=30.47, *SD*=5.77) was similar, with 96 students completing both pre- and post-tests. Despite the similarity in overall score, sub-scale measurements illustrated a significant decrease in the deduction sub-scale, $t(95)=3.416$, $p<0.005$, and a significant increase in the semantics and meaning sub-scale, $t(95)=-2.562$, $p<0.05$.

Table 9: Cornell Level Z Group: Cohort B Mean Scores and Sub-Scores

	N	Test Score		Deduction		Semantics & Meaning		Credibility		Induction		Assumptions	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pre-test	96	30.90	4.51	7.63*	1.12	4.94*	1.77	2.62	1.04	10.42	1.71	5.29	1.94
Post-test	96	30.47	5.77	7.09*	1.39	5.48*	1.95	2.70	1.02	9.91	2.43	5.29	1.86

Cornell Level Z Group: Cohort C

The students of Cohort C wrote the ICTET for the pre-test and the CLZ for the post. The main purpose of this group was to assess any test-retest effects with the CLZ. The summary of the critical thinking pre-test (*M*=4.78, *SD*=1.41) and critical thinking post-test (*M*=30.13, *SD*= 4.99) scores are illustrated below in Table 10, with 84 students completing both pre- and post-tests. The sub-scores for each test were compiled but are not shown due to differences in the sub-scores measured by each test.

Table 10: Cornell Level Z Group: Cohort C Mean Scores

	N	Test Score	
		Mean	SD
Pre-test (ICTET)	84	4.78	1.41
Post-test (CLZ)	84	30.13	4.99

Using the Cornell Level Z user norms for rights-only scoring, the mean post-test score reported in this study (*M*=30.47, *SD*=5.77) was comparable to two of the mean scores of studies assessing the CTS ability of freshman undergraduate students, and greater than the mean scores of two other studies (*M*=29.8, *SD*=4.4; *M*=27.8, *SD*=4.7; *M*=25.9, *SD*=4.2; *M*=31.2, *SD*=3.9)(Ennis et al., 1985).

ICTET Group: Cohort D

The students of Cohort D wrote the ICTET for both the pre-test and the post-test. The summary of the critical thinking pre-test (*M*=4.78, *SD*=1.06) and critical thinking post-test scores (*M*=4.60, *SD*=1.32) is illustrated below in Table 11, with 109 students completing both pre- and post-tests. There was no significant difference between the pre-test and post-test assessment of any of the ICTET sub-scores.

Table 11: ICTET Group: Cohort D Mean Scores and Sub-Scores

	N	Test Score		Purpose		Questions		Information		Conclusions		Concepts		Assumptions		Implications		Points of View	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	M	SD	Mean	SD	Mean	SD	Mean	SD
Pre-test	109	4.78	1.06	6.77	1.73	5.48	2.16	5.52	2.13	5.70	1.93	4.00	1.92	2.44	1.19	3.60	2.04	4.78	1.94
Post-test	109	4.60	1.32	6.28	1.89	5.59	1.95	5.47	1.70	5.32	2.09	3.90	1.88	2.68	1.69	3.24	2.14	4.32	2.28

ICTET Group: Cohort E

The students of Cohort E wrote the CLZ for the pre-test and the ICTET for the post. The main purpose of this group was to assess any test-retest effects with the ICTET. The summary of the critical thinking pre-test (*M*=30.13 *SD*=4.25) and critical thinking post-test scores (*M*=4.67 *SD*=1.21) is illustrated below in Table 12.. The mean pre-test score was 30.17±4.25, compared to the post-test score of 4.67±1.21, with 101 students completing both pre and post-tests. The sub-scores for each test were compiled but are not shown due to differences in the sub-scores measure by each test.

Table 12: ICTET Group: Cohort E Mean Scores

	N	Test Score	
		Mean	SD
Pre-test(CLZ)	101	30.17	4.25
Post-test (ICTET)	101	4.67	1.21

Performance measures and comparison of the ICTET were not conducted due to a lack of published studies using the ICTET.

Hypothesis Validation

Upon comparison by paired *t*-test, there was no significant difference between the CTS of cohorts with the same post-test, as shown in below in Table 13. This suggests the robustness of CTS measurement by the CLZ or the ICTET; potential test-retest effects may not be as important a factor as suspected.

Table 13: Hypothesis Validation: Comparison of Cohorts Post-Test Scores

Cohort			Critical Thinking Test Score	
			Mean	SD
	Cohort B (CLZ-CLZ)	Post	30.47	5.77
	Cohort C (ICTET-CLZ)	Post	30.13	4.99
	Cohort D (ICTET-ICTET)	Post	4.60	1.32
	Cohort E (CLZ-ICTET)	Post	4.67	1.21

MEA Results

The MEA scores for all students, analyzed by repeated measures ANOVA with a Bonferroni adjusted post-hoc test, elicited statistically significant increases over the course of the semester at $p<0.005$, as shown in Table 14. Common MEA sub-scores of information summary, model generation, critical evaluation and argumentation exhibited statistically significant increases over the course of the semester. Model results and communication exhibited a statistically significant increase across the first two MEA activities, analyzed by repeated measures ANOVA with a Bonferroni adjusted post-hoc test. There was no observed difference between the second and third MEA activity.

Table 14: MEA and MEA Sub-Score Comparison over the Duration of APSC100

MEA	N	MEA Score		Information Summary		Model Generation		Model Results		Critical Evaluation		Argumentation		Communication	
		M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.
1	536	61.65	12.22	5.30	1.37	4.72	1.47	4.88	1.41	4.30	1.68	4.66	1.30	5.49	1.14
2	541	70.74	9.27	5.69	1.17	5.82	1.12	5.67	1.07	5.50	1.12	5.64	.97	5.89	1.02
3	542	73.85	7.75	5.98	1.00	6.16	1.09	5.71	.92	5.74	.88	5.81	.76	5.89	.92

Correlations between MEAs and Critical Thinking Tests

The relationship between the separate MEA scores and sub-scale scores and the critical thinking tests and their sub-scores were investigated. There were weak, significant correlations between certain critical thinking test scores and sub-scores and MEA scores and sub-scores. The CLA exhibited the most items correlated with MEA scores and sub-scale items as shown in Table 15. Interestingly, MEA 1 and 2 exhibited the greatest number of correlations, with MEA 3 exhibiting the smallest number of correlated elements. The ICTET exhibited fewer correlations with the MEA scores and sub-scores, less than the number of correlations between MEAs and the CLA, as illustrated in Table 16. The CLZ exhibited a small number of significant, negative correlations with the MEA scores and sub-scores, and exhibited the smallest number of correlated elements with the MEA scales compared to the other critical thinking tests, as illustrated in Table 17.

Table 15: MEA-CLA Item Correlations

		CLA Score	Analytic Reasoning	Problem Solving	Writing Mechanics	Writing Effectiveness
MEA 1 (*N*=148)	Information Summary	.048	.044	.024	-.045	.085
	Model Generation	.130	.108	.093	.068	.149
	Model Results	.178*	.197*	.168*	.053	.201*
	Critical Evaluation	.277**	.250**	.257**	.170*	.280**
	Argumentation	.154	.139	.112	.135	.204*
	Communication	.202*	.185*	.166*	.131	.177*
	Overall Score	.248**	.233**	.206*	.143	.281**
MEA 2 (*N*=150)	Information Summary	.165*	.137	.079	.077	.155
	Model Generation	.119	.089	.050	-.095	.067
	Model Results	.231**	.214**	.183*	.099	.206*
	Critical Evaluation	.217**	.278**	.163*	.104	.201*
	Argumentation	.275**	.289**	.219**	.039	.211**
	Communication	.276**	.205*	.194*	.244**	.223**
	Overall Score	.258**	.240**	.167*	.073	.212**
MEA 3 (*N*=151)	Information Summary	.039	.037	.011	-.022	.044
	Model Generation	-.174*	-.128	-.160*	-.109	-.130
	Model Results	-.090	-.098	-.064	-.143	-.092
	Critical Evaluation	.115	.060	.089	.099	.079
	Argumentation	.123	.161*	.079	.045	.100
	Communication	-.044	-.053	-.056	-.080	-.032
	Overall Score	.002	.007	-.016	-.026	.010

** Correlation is significant at the 0.01 level (2-tailed)

* Correlation is significant at the 0.05 level (2-tailed)

Table 16: MEA-ICTET Item Correlations

		ICTET Score	Purpose	Questions	Information	Conclusions	Concepts	Assumptions	Implications	Points of View
MEA 1 (*N*=82)	Information Summary	.033	-.024	.054	-.059	.004	.023	.144	.025	.017
	Model Generation	.011	-.024	-.056	.005	-.023	.034	-.052	.041	.109
	Model Results	-.174	-.149	-.226*	-.098	-.110	-.037	-.121	-.117	-.087
	Critical Evaluation	-.082	-.097	-.158	-.105	.036	-.009	-.028	-.152	.051
	Argumentation	-.061	-.202*	-.145	-.013	.036	.065	-.086	-.049	.042
	Communication	.215*	.029	.117	.181	.207*	.227*	.207*	.023	.186
	Overall Score	-.070	-.148	-.105	-.035	-.026	.042	-.013	-.105	.012
MEA 2 (*N*=84)	Information Summary	.177	.127	.191	.086	.159	.153	.194	.051	.030
	Model Generation	.044	.122	.236*	.122	.014	.005	-.034	-.070	-.114
	Model Results	.147	.066	.175	.057	.041	.102	.061	.204*	.080
	Critical Evaluation	.037	.088	.175	.078	.063	-.011	.004	-.028	-.134
	Argumentation	.206*	.121	.181	.119	.151	.136	.121	.125	.156
	Communication	.315**	.244*	.292**	.113	.277**	.163	.378**	.126	.139
	Overall Score	.212*	.172	.293**	.128	.151	.090	.175	.105	.057
MEA 3 (*N*=84)	Information Summary	-.013	-.086	-.046	-.015	-.092	.111	-.044	.060	.029
	Model Generation	.191	.201*	.198*	.192	.138	.087	.063	.147	.024
	Model Results	-.091	-.083	-.014	-.145	.012	.050	-.032	-.179	-.096
	Critical Evaluation	-.093	-.106	-.045	-.034	-.081	.045	-.081	-.040	-.143
	Argumentation	.058	.026	.067	.096	.111	.113	.025	-.099	-.001
	Communication	.044	.055	.177	.153	.041	-.013	-.031	-.115	-.002
	Overall Score	.031	.007	.086	.077	.047	.097	-.019	-.092	-.015

** Correlation is significant at the 0.01 level (2-tailed)

* Correlation is significant at the 0.05 level (2-tailed)

Table 17: MEA-CLZ Item Correlations

		CLZ Score	Deduction	Induction	Semantics & Meaning	Observations & Credibility	Assumptions
MEA 1 (*N*=97)	Information Summary	-.138	-.053	-.122	-.152	-.067	-.034
	Model Generation	.039	-.034	.104	.028	.047	-.043
	Model Results	.146	.109	.155	.041	.133	.052
	Critical Evaluation	-.025	-.022	.057	-.169	.080	-.004
	Argumentation	-.028	-.217*	.086	-.018	.012	-.027
	Communication	.291**	.180	.172	.150	.221*	.265**
	Overall Score	.058	-.025	.091	-.062	.071	.104
MEA 2 (*N*=97)	Information Summary	.165	.070	.123	.173	.009	.114
	Model Generation	.303**	.098	.163	.311**	.161	.242*
	Model Results	-.069	-.098	-.072	.022	.016	-.078
	Critical Evaluation	.048	-.059	.097	.045	-.067	.056
	Argumentation	.053	-.101	.042	.072	.061	.077
	Communication	.062	-.099	.113	-.004	-.030	.138
	Overall Score	.092	-.081	.094	.101	.003	.116
MEA 3 (*N*=97)	Information Summary	.091	-.044	.149	.222*	-.158	-.026
	Model Generation	.162	.065	.129	.183	.110	.035
	Model Results	.140	.069	.092	.119	.091	.088
	Critical Evaluation	.002	-.155	.003	.072	-.107	.101
	Argumentation	-.017	-.078	.006	.023	-.116	.038
	Communication	.145	.079	.113	.077	.066	.127
	Overall Score	.185	.020	.195	.170	.018	.117

** Correlation is significant at the 0.01 level (2-tailed)

* Correlation is significant at the 0.05 level (2-tailed)

MEA Reliability

The reliability of each MEA was assessed through item-total correlations between sub-scores and overall MEA score and Cronbach's alpha. The results of the item-total correlations (sub-score to total score) and the internal consistency, measured by Cronbach's alpha, are illustrated below in Table 18. Each MEA exhibited high measures of internal consistency ($0.7< \alpha <0.9$) and strong, significant correlations with the total score ($r >0.5$, $p<0.005$). As previously stated, we are using the MEAs as a course-related measure for the development and assessment of critical thinking and problem solving skills. These statistics demonstrate how well each sub-score contributes to the overall MEA score (item-total correlations) and how closely related and reliable each sub-score is to the overall MEA score (Cronbach's alpha).

Table 18: MEA Reliability Measures

	Cronbach's Alpha	N	Information Summary	Model Generation	Model Results	Critical Evaluation	Argumentation	Communication
MEA 1	0.813	536	0.623**	0.647**	0.756**	0.754**	0.797**	0.621**
MEA 2	0.802	540	0.689**	0.620**	0.762**	0.677**	0.754**	0.622**
MEA 3	0.771	542	0.602**	0.690**	0.601**	0.600**	0.732**	0.701**

** Correlation is significant at the 0.005 level (2-tailed)

Think Aloud Sessions

The results presented for this section of our study give an overview of the coding and analysis of the safety recommendations segment from the transcribed pre and post think aloud sessions for the control (n=3) and experimental group (n=2). This segment was coded for elements of critical thinking corresponding to the Paul-Elder model. The quality of each element was assessed, with codes being separated into those of acceptable and unacceptable quality. Once coded, the safety recommendations unit was analyzed for common themes and for notable differences between pre- and post-tests in the experimental and control group. The two groups were also compared and contrasted as presented below.

Think Aloud Pre-Post Comparison: Control Group

The three physics subjects serving as the control group displayed considerable improvement between the pre- and post-tests. In the pre-tests, participants approached the problem in reverse, reviewing the supplemental information and forming recommendations at the beginning of the session, prior to any analysis. These conclusions, presented as safety recommendations, were formed primarily on assumptions grounded in students' own personal experiences and not on a comprehensive analysis and solution of the key elements of the problem. The subjects eventually attempted an analysis of the scenario, but became confused and lacked a clear plan or method to solve the physics aspect of the problem. The subjects continued to remain stuck on the analysis past the allotted time and failed to provide any conclusions or safety recommendations resulting from that analysis.

In the post-test, subjects approached the problem similarly to the way displayed by the engineering group. In the post-tests, subjects based their conclusions primarily on assumptions, with little use of concepts and supplemental information. Subjects used personal and anecdotal experience to formulate conclusions rather than support conclusions with supplemental information. There was little validation of questioning of their

assumptive conclusions, and recommendations were given without any reference to the limitations of their knowledge, expertise or any societal implications. However, these elements of critical thought were present in their conclusions, which can be viewed as an improvement from their performance on the pre-test despite the relative quality of these elements.

Think Aloud Pre-Post Comparison: Experimental Group

The experimental group, consisting of two engineering students who participated in an MEA-integrated curriculum, showed improvements in elements of critical thinking, specifically in areas of concepts, information and implications. In the pre-test, the subjects balanced their conclusions between assumptions, either drawing from personal or anecdotal experience or from their previous knowledge base of physics concepts, supplemented occasionally by data and literature. The supporting literature was adopted and used alongside assumptive reasoning to form their conclusions, without any further investigation. Subjects did not question the validity and accuracy of their own assumptions and gave the briefest consideration to the potential implications of their solutions.

In the post-test, the subjects based their conclusions on their initial analysis supplemented by concepts and supplemental information. There were few conclusions based on assumptions formed by first-person or anecdotal evidence or experience. Concepts of physics and forces were used in a progressive manner, with subjects developing mathematical relationships between variables rather than simply using the equations. The validity of supplemental information was questioned, with the subjects considering the authority of the provided information. Subjects also considered the implications of their conclusions, highlighting limitations and possible areas of concern within their conclusions to be addressed in the future and even identifying a possible conflict with building codes.

The subjects demonstrated improvement from the pre-test, particularly in the elements of concepts, assumptions, information and implications. Subjects questioned the validity and source of supplemental information, in addition to questioning the accuracy and validity of their own assumptions and conclusions, in contrast to the pre-test where this was poorly demonstrated. Concepts were used beyond simple application, as the subjects used concepts along with mathematical formulae to produce a mathematical model relating the design parameters that required recommendations. Lastly, the subjects vastly improved in considering implications, identifying additional codes and requirements outside of the scope of the recommendations, and readily highlighted potential gaps in their knowledge upon which to improve.

Survey Analysis

Motivation Sub-Scales

The ten pre-survey motivation questions were transformed from component questions into three sub-scales that each probed a different type of motivation: intrinsic motivation (IM), introjected and external motivation (IJ/ER), and internal and identified motivation (IN/ID). The reliability (internal consistency) measures for each sub-scale are IM: α=0.769, IJ/ER: α=0.706, IN/ID: α=0.712. The results of the sub-scale grouping are shown below in Table 19, Table 20 and Table 21. The student responses showed that the majority of students agreed or strongly agreed with all of the motivational questions, which placed them in a high motivation grouping. A correlation analysis was conducted to investigate any relationship between motivation and performance on MEAs and critical thinking post-test performance and each instruments related sub-scores. There were no observed correlations between motivation, MEAs and critical thinking test performance.

Table 19: Intrinsic Motivation Classification of Study Participants

Intrinsic Motivation				
		Frequency	Percent	Valid Percent
	Low Motivation	47	8.7	8.8
	Neutral	44	8.1	8.3
	High Motivation	441	81.4	82.9
	Total	532	98.2	100.0
	Missing	10	1.8	
Total		542	100.0	

Table 20: Introjected and External Motivation Classification of Study Participants

Introjected & External Motivation				
		Frequency	Percent	Valid Percent
	Low Motivation	53	9.8	10.0
	Neutral	31	5.7	5.8
	High Motivation	448	82.7	84.2
	Total	532	98.2	100.0
	Missing	10	1.8	
Total		542	100.0	

Table 21: Internal and Identified Motivation Classification of Study Participants

Internal & Identified Motivation				
		Frequency	Percent	Valid Percent
	Low Motivation	9	1.7	1.7
	Neutral	7	1.3	1.3
	High Motivation	516	95.2	97.0
	Total	532	98.2	100.0
	Missing	10	1.8	
Total		542	100.0	

ESL Status

The majority of students participating in the study identified English as their first language (EFL) (n=498), with a small percentage of students identifying English as a second language (ESL) (n=34) and ten students declining to respond. Stratified sampling reduced these numbers when grouping by similar critical thinking test and English proficiency. The CLA was the only test that exhibited a significant difference in critical thinking test performance by ESL status, $t(142)=-5.364$, $p<0.005$), with ESL students ($M=979.5$, $SD=40.36$) achieving a lower score than EFL students ($M=1218.92$, $SD=12.94$), as illustrated below in Table 22.

Table 22: Differences in Critical Thinking Score by ESL Status

	ESL Status	N	Mean	S.D.
CLA Score	English as a second language	12	979.50	139.81
	English as a first language	132	1218.92	148.71
Cornell Z Post Score	English as a second language	10	29.60	5.56
	English as a first language	171	30.27	5.50
ICTET Post Score	English as a second language	12	45.36	10.11
	English as a first language	199	46.35	12.76

External Factors Affecting CTS

Results from the post survey questions regarding student effort (three questions, Figure 10) student perceptions of university and course experience on CTS development (six questions, Figure 11) and student perceptions of knowledge integration, discussion and peer instruction (four questions, Figure 12) are presented below. Student rankings of first-year course-based and extracurricular activities are also presented (Figure 13). While there was no significant relationship between any of these questions and CTS, some insight can be gained into which course experiences students found essential for developing the type of thinking used during their critical thinking post-test. Students viewed working on MEAs in both an individual and group setting and the APSC 100 lectures to contribute to the type of thinking used on their critical thinking post-test. Also, when asked to rank first-year course experiences according to the development of the type of thinking used in their critical thinking post-test, students ranked APSC 100 Module 1 as most important for CTS development (46.4%). Other course experiences, APSC 111(Physics) (22.9%) and APSC 131(Chemistry) (22.4%), were indicated as the next two most important course experiences.

Figure 10: Student Effort Survey Question Results

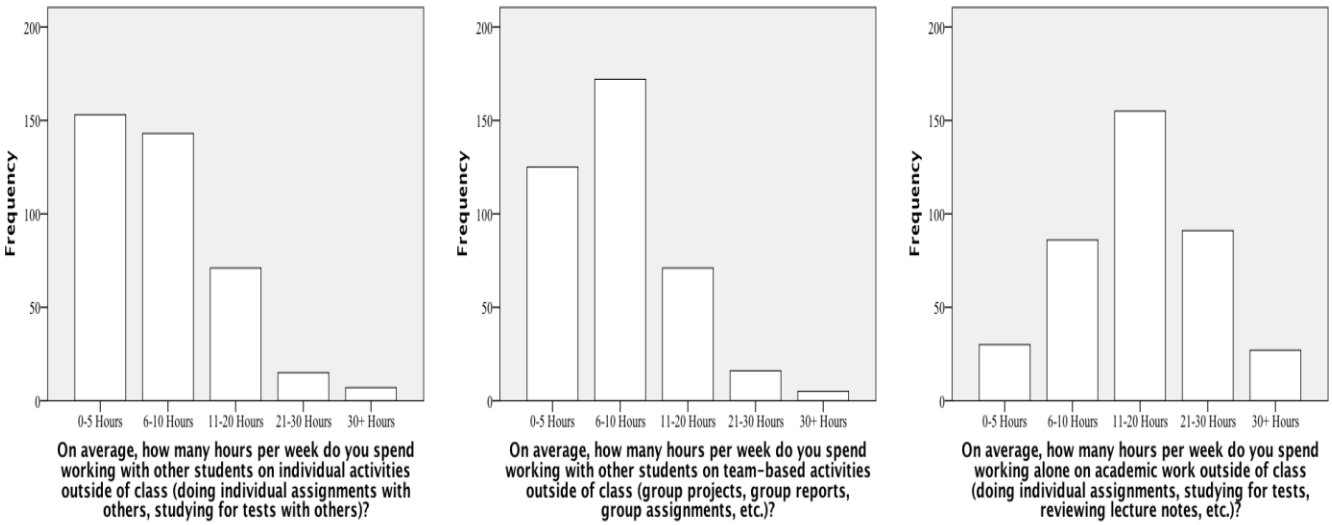


Figure 11: Student Perceptions of Course Activities on CTS Development

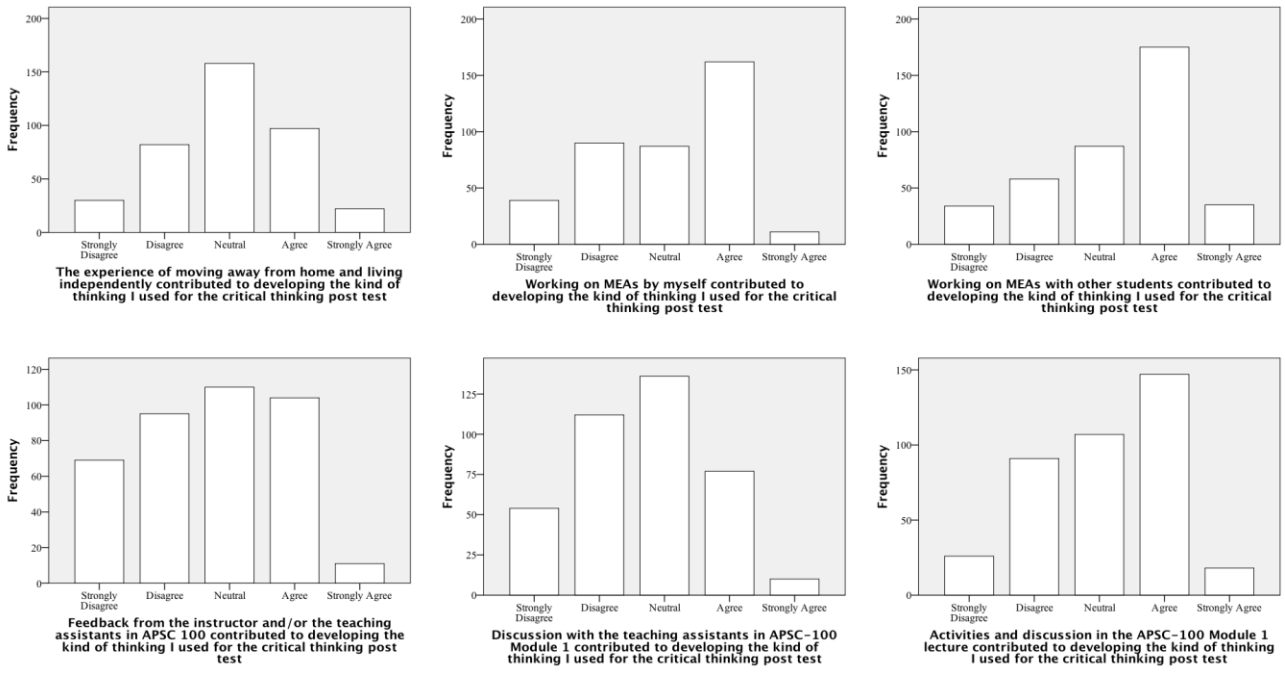


Figure 12: Student Perceptions Regarding Knowledge Integration, Content Discussion and Tutoring

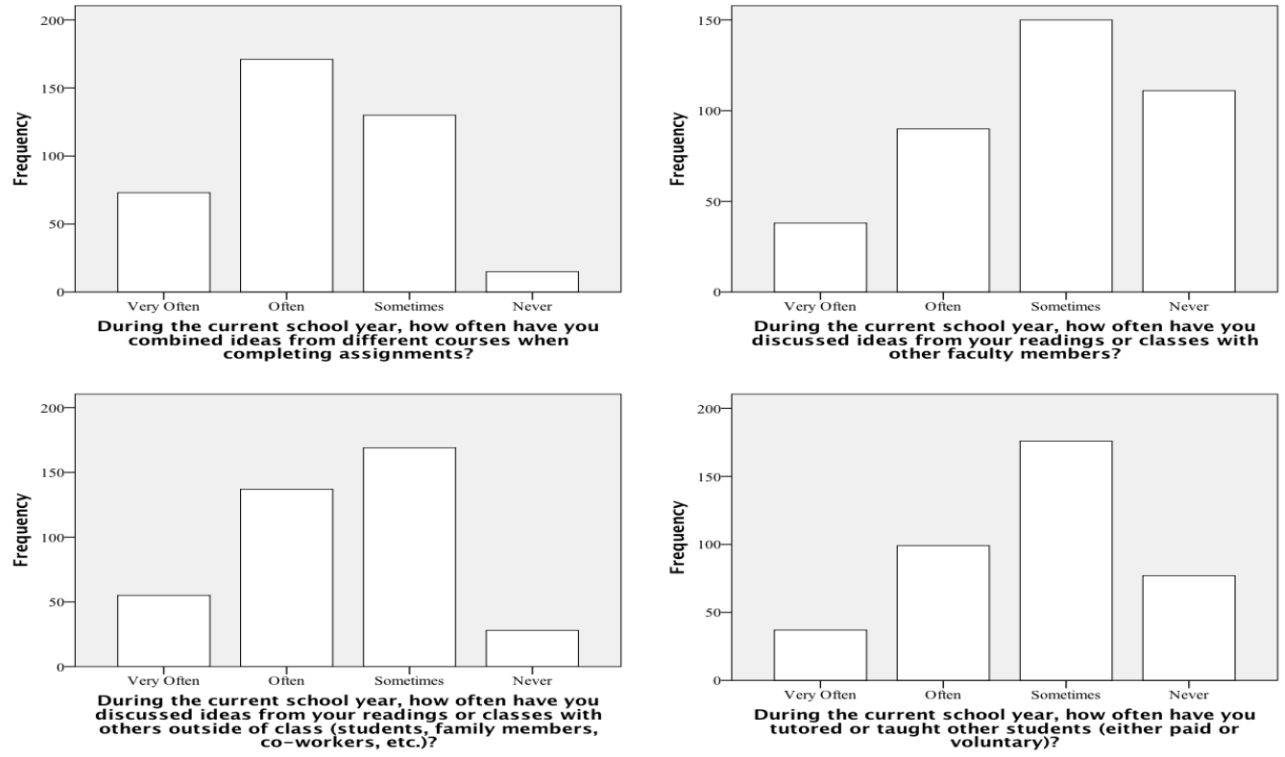
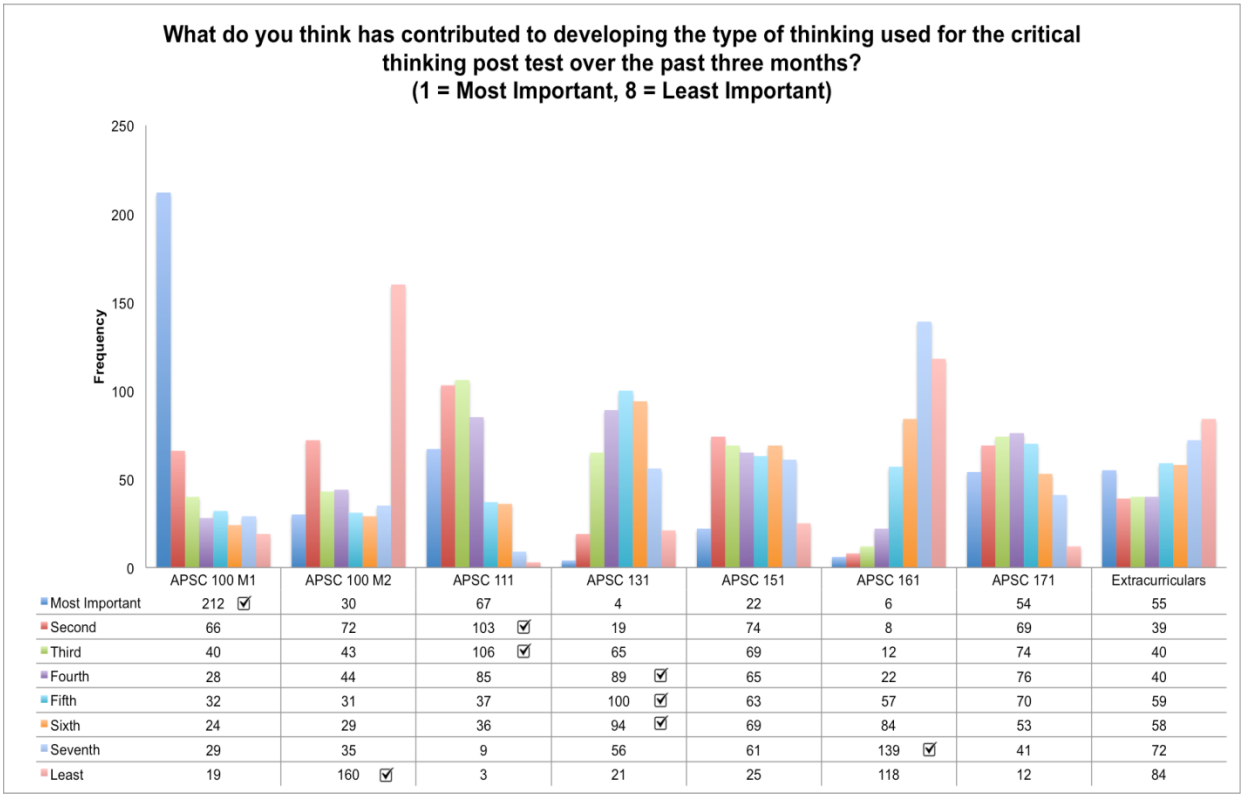


Figure 13: Student Ranking of First-Year Experiences Contributing to CTS



	APSC 100 M1	APSC 100 M2	APSC 111	APSC 131	APSC 151	APSC 161	APSC 171	Extracurriculars
Most Important	212 ☑	30	67	4	22	6	54	55
Second	66	72	103 ☑	19	74	8	69	39
Third	40	43	106 ☑	65	69	12	74	40
Fourth	28	44	85	89 ☑	65	22	76	40
Fifth	32	31	37	100 ☑	63	57	70	59
Sixth	24	29	36	94 ☑	69	84	53	58
Seventh	29	35	9	56	61	139 ☑	41	72
Least	19	160 ☑	3	21	25	118	12	84

Critical Thinking Test Reliability

Reliability for each test was calculated from results of the study by using multiple measures. Test-retest reliability was assessed using Pearson's product-moment correlation coefficient and pre- and post-test scores. Internal consistency was measured by Cronbach's alpha using the test sub-scores.

CLA

Since the CLA was used only as a pre-test, test-retest reliability statistics could not be determined. Cronbach's alpha for the CLA was $\alpha=0.920$, calculated using the CLA sub-scores of problem solving, analytic reasoning, writing mechanics and writing effectiveness. This high measure of internal consistency allows us to infer that the sub-scales and overall scores of the CLA are very closely related.

Cornell Level Z

The test-retest reliability, measured by Pearson's moment correlation coefficient, was calculated at $r(95)=0.471$, $p<0.005$. A correlation of this magnitude indicates that there are no test-retest effects present with Cohort B, supported by comparison of post-test scores with Cohort E. Cronbach's alpha, using sub-scores of deduction, semantics and meaning, observation and credibility, induction and assumptions, was calculated at $\alpha=0.645$. This moderate measure of internal consistency suggests that the sub-scales and overall scores of the CLZ are closely related, but raises some questions regarding the reliability of the tool.

International Critical Thinking Essay Test

An additional reliability statistic was investigated for the ICTET, due to the multiple graders assessing the test. IRR was calculated using Krippendorff's alpha, measured as previously stated (Hayes & Krippendorff, 2007), at the end of the sessions with the Foundation for Critical Thinking and after the grader training sessions. The measures of IRR were α=0.932 and α=0.974, respectively, which indicate a high level of agreement between graders and ensure consistent grading. Upon collecting the scored papers and assigning them to graders, a single grader scored all papers for Cohort D.

The test-retest reliability was calculated as r=0.150 (n=101). This low correlation indicates a potential test-retest effect with the ICTET. However, cross-cohort comparison indicates that Cronbach's alpha using the ICTET sub-scores of purpose, questions, information, conclusions, concepts, assumptions, implications and point of view was calculated at α=0.828, which suggests a strong relation between the sub-scales and overall score of the ICTET.

Discussion

The tests for reliability of the CTS tools presented above are consistent with previous studies in the literature, yet there is no significant change in means between the pre- and post-test. These results do not show critical thinking gains over the term, which contrasts with evidence from the MEAs themselves, survey results and think aloud analyses, which do show learning gains. The authors feel the evidence points to issues with the instruction, alignment, assessment and the administration of standardized tests.

Anecdotal comments from the proctors suggest that students were quite keen during the pre-tests, but were quite fatigued at the end of their stressful first semester in the intense first-year engineering program. Student surveys run as part of normal program delivery two weeks before the post-test found that half of the class was struggling to keep up with the demands of the program. During the post-test, the proctors observed significant motivational issues, including many comments like "Why are we taking this same test again?", and a large number of students who appeared to put little effort into the post-test.

The MEAs showed high item-total correlations between sub-scales and scores, and good internal consistency of sub-scales. In contrast to the results from the CTS tools, significant performance gains were observed in the MEA score means, particularly from MEA 1 to MEA 2, with relatively little change between MEA 2 and MEA 3. Significant improvements were observed in argumentation, information evaluation, and in students' ability to evaluate their own work critically.

There is a question of alignment between the critical thinking tests and the approaches to evaluating critical thinking in the engineering context through the MEAs. The correlation between MEA scores and CLZ and ICTET was not significant, despite the fact that Paul and Elder's model of critical thinking was the framework for critical thinking instruction in the course and the ICTET was built around the same framework. However, it used tasks requiring students to identify points of view, inferences, etc., from a prompt, rather than requiring a constructed logical argument, as is required by the CLA and MEAs. The CLA, which uses constructed response tasks, had the highest correlation with the MEA scores. In this study, similarity of task provides better alignment than similarity of critical thinking framework.

Context, framework and task are important for assessing generic skills like critical thinking, problem solving, etc. The CLZ was not set in an engineering context, used a different framework from the Paul and Elder model, and as a multiple choice test was a very different task than the constructed responses expected for the contextual and complex problems provided in the course activities. Interestingly, several of the CLZ scales were negatively correlated with MEA performance. Critical thinking is a multidimensional construct involving

skills, disposition and metacognition; the CLZ and ICTET offer highly structured prompts or multiple choice, which do not take into account disposition and metacognition.

It is noteworthy that other researchers have identified no significant learning gains over the course of a semester, even when a course is focused exclusively on critical thinking (which was not the case for this study). Nieto and Saiz (2008) show increases in performance in course activities (grades), but no difference in pre-post-test scores measured by the Cornell. CTS and activities were evaluated according to Halpern's model, illustrating that the difference in test model and instructional model (alignment) has an effect.

There was no significant correlation between scales on the motivational surveys and any of the instruments. Students generally responded positively ("agree" or "strongly agree") to all of the items on the motivational surveys and were not asked to select between the motivational factors. This is a weakness of the Academic Motivation Survey items used in this study.

In the think aloud problems, there were very noticeable differences between the engineering and physics control groups. The engineering subjects exhibited greater use of aspects of critical thinking in their safety recommendation conclusions. In the pre-tests, both groups displayed significant reliance on assumptive personal experiences to form conclusions, yet the engineering subjects utilized additional elements, incorporating concepts and supplemental information into their analysis and ultimately into their conclusions, whereas the physics control group relied solely on assumptive reasoning. Despite these improvements in exhibiting and applying elements of critical thinking, both groups did not question the credibility or validity of the supplemental information or display a consideration to the potential implications of their recommendations.

Both groups exhibited improvement in using elements of critical thinking during the post-tests. The physics group continued to rely primarily on assumptive personal experiences to form conclusions, but did begin to incorporate conceptual elements into recommendations, utilizing combinations of methodologies and extensions of theory to inform their conclusions. The physics group also showed the beginnings of using supplemental material to support their conclusions, similar to how the engineering group utilized supplemental material in their pre-test.

The engineering group showed a reduction in personal experience-based assumptive reasoning. The conclusions presented by the engineering group were supported by concepts, information, implications and evaluation of assumptions. The engineering subjects utilized a number of concepts to create a mathematical relationship between the design parameters in order to solve the problem, which could ultimately lead to a descriptive model. This use of concepts was similar to the methodology covered by instruction and model eliciting activities the subjects experienced in APSC 100. Compared to the pre-test, the engineering subjects considered the credibility of supplemental information and used multiple sources to assist their solution and form the basis for safety recommendations. The engineering group, during the post-test, incorporated new elements of critical thinking not demonstrated during the pre-test, considering the implications of their solutions with respect to potential zoning regulations and questioning their own assumptions and limits of knowledge.

These differences between the groups may be attributed to the MEAs and to the critical thinking instruction present in APSC 100. It may also be attributed to the varying educational backgrounds that the different groups may possess. As a means to determine potential reasons, a small exit interview was conducted at the end of the think aloud post-test. Subjects were asked if there was anything in their university or life experience that helped them with the thinking in this type of problem. The control group responded that the pre-test, alongside a similar roller coaster-based question from their physics class, helped them during the post-test. Students in the experimental group responded that they felt the experiences in APSC 100 helped them, and alluded to creating a model and relationships between variables and a self-defined solution with this type of

thinking. Students in the end-of-course think aloud relied less on personal assumptions, considered alternative points of view and evaluated the validity of information more than students in the beginning of course think aloud. The themes from the think aloud studies align well with the instructional goals and observed performance gains in the sub-scores of the MEAs.

The student survey demonstrated that students overwhelmingly identified the activities in this course, primarily MEAs, as the most influential experience in developing critical thinking. Given the gains observed in MEA performance and observations in the think alouds, the authors conclude that there is a misalignment between the two CTS tools used for pre- and post-course evaluation, the Cornell Level Z and International Critical Thinking Essay Test, and the course

Recommendations and Future Research

There are a number of recommendations resulting from this study, stemming from the research objectives outlined at the beginning of the report:

Is there a correlation between critical thinking instrument scores and MEA scores?

There were no observed correlations between student scores on the standardized critical thinking instruments and MEA scores. The authors believe that this lack of correlation may be explained by alignment between the standardized tools and the MEAs and issues regarding instrument sensitivity. Alignment will be addressed in a later recommendation. Regarding sensitivity, standardized assessments may not possess sufficient sensitivity or resolution to measure the development of CTS over the course of a semester. The authors of the CLA do not recommend using their test as a pre and post pairing over the course of the semester as they have found that it takes longer to detect development. The authors of the ICTET express a similar concern, that any gains in CTS measured by a standardized instrument over a semester may only be reporting temporary gains, as the development of CTS typically occurs through continual application, practice and reflection (Halpern, 2002; Paul & Elder, 2005). The CLZ has been used in several studies in a pre and post arrangement, and has illustrated student gains in CTS measured by the assessment. However, the criticisms summarized in this report raise concerns about the validity and accuracy of multiple-choice assessments of CTS.

Is there a correlation between critical thinking instrument sub-scores and MEA sub-scores?

There were some observed correlations between the sub-scores of standardized critical thinking instruments and MEA sub-scores. The CLA had the greatest numbers of correlated items between the sub-scores, with the ICTET having the second most correlated items, and the CLZ having the least number of correlated items between the respective sub-scores. Similar to the overall scores, the authors believe that the differences in items correlated may be explained by alignment between the specific standardized tools and the MEAs that will be addressed in a later recommendation.

Is there a correlation between critical thinking ability and motivational factors?

There was no observed correlation between critical thinking ability, as measured by student scores on the standardized critical thinking tests, and motivational factors. Student responses to the survey instrument indicated that students reported high motivation for all sub-scales of the instrument. This may be attributed to only using a select group of questions to measure a complex construct such as motivation, instead of instrument as a whole. Also, students were not asked to rank or compare motivational factors, which may have contributed to students reporting high motivation for the different types of motivation reported by the sub-scales.

Is there a correlation between critical thinking ability and specific course experiences?
Is there a correlation between critical thinking ability and specific extrinsic factors?

These two questions are addressed together. From student self-reports, the think alouds and survey information, there was a correlation between critical thinking ability and specific course experiences. A large number of students identified that APSC 100 Module 1 and the MEAs themselves as responsible for developing critical thinking skills. Students viewed working on MEAs and the APSC 100 Module 1 lectures as beneficial for critical thinking ability. With regard to extrinsic factors, students indicated that critical thinking ability was developed in their first-year physics (APSC 111) and chemistry (APSC 131) courses. Extracurricular activities and university experience were indicated as relatively unimportant in the development of critical thinking ability.

Are the critical thinking instruments used reliable and valid?

The standardized instruments possess different strengths and weaknesses with respect to reliability and validity. Reliability was assessed by Cronbach's alpha, test-retest reliability and, where possible, inter-rater reliability. The CLA exhibited high internal consistency, but test-retest reliability could not be established due to its use solely as a pre-test. The CLZ displays high test-retest reliability but moderate internal consistency, which raises some concerns regarding the accuracy of the test. The ICTET exhibited poor test-retest reliability, high inter-rater reliability and high internal consistency. The low test-retest reliability is not an inherent problem with the ICTET test itself but indicates that using the same prompt for a pre- and post-test may not be advisable.

Traditional measures of validity of any instruments are difficult to establish. With respect to this study, a valid instrument was one whose tasks that reflects the application of CTS outlined by the course objectives and MEAs. This was determined by two approaches: first, by how well the instrument tasks reflected in the course objectives, which view the application of CTS in solving complex engineering problems; second, by correlations between the MEA sub-scales and instrument sub-scales. Performance-based instruments, such as the CLA, that seek to holistically evaluate CTS appear to be a suitable approach to assess engineering CTS. Out of the three instruments, the CLA exhibited highest number of correlated items between sub-scales, and despite its generic context, the CLA tasks parallels the course objectives. Multiple-choice instruments such as the CLZ should be avoided, as the recognition and recall nature of these assessments don't accurately reflect the course objectives. The ICTET, despite maintaining some domain-specific knowledge through an engineering-related prompt, doesn't reflect the application of CTS outlined in the course objectives. These differences are further illustrated through the low to moderate number of correlations between the MEAs and each instruments respective sub-scales.

The secondary objectives of the study sought to provide an approach of how to assess critical thinking in an engineering context, and provide a starting point for other parties interested in critical thinking development and assessment. The conclusions resulting from these objectives include:

Is there evidence that MEAs have a significant positive impact on students' critical thinking skills?

Despite no reported gain in students' CTS as measured by the standardized tools, the authors conclude the MEAs have a significant positive impact on CTS development. This conclusion is supported by the increase in MEA scores and critical thinking-related outcomes over the course experience. The post-test survey results and course rankings illustrate that students indicated that the MEAs and APSC 100 Module 1 lectures and studios were key contributors to the development of CTS. Lastly, analysis of the think alouds illustrated that the engineering group utilized CTS more than their physics counterparts, and identified the MEAs and APSC 100 Module 1 as the factors that contributed to the development of these skills.

> Which critical thinking framework, and which critical thinking instrument, reflects the application of critical thinking skills in solving complex engineering problems?

The selection of an instructional framework should be based on how well it reflects with the course objectives regarding the application of CTS. Additionally, the dimensions of the framework should be consistent with how critical thinking skills are applied within discipline-specific practice. Any model or framework for critical thinking would be suitable for discipline-specific use, provided explicit instruction in how the dimensions of the framework, and how CTS are applied in disciplinary practice.

Selecting a standardized instrument required careful attention and investigation, and cannot be adapted as easily as a framework. The prompts in each instrument are very specific and are crafted in such a way to assess the dimensions of the framework on which they are based. An assessment structured in this fashion may only serve to measure how well the student understands or recognizes the dimensions of the framework and does not measure how CTS are applied to solve complex engineering problems.

> To what extent does alignment of tasks between critical thinking instrument and complex engineering problems need to be preserved?

Utmost care should be taken to establish and maintain task alignment between a critical thinking instrument and how critical thinking is applied to solve complex engineering problems, or problems in any specific discipline. Misalignment in tasks leads to inaccurate results and questions regarding the reliability, validity and authenticity of the instrument. Establishing task alignment using standardized tools is a difficult challenge. This is evident with respect to the standardized instruments used in this study, specifically the CLZ and ICTET, as the tasks within those instruments do not accurately reflect complex engineering problems. The CLA possesses a greater measure of task alignment regarding engineering CTS, although the performance tasks presented in the instrument are generic in nature. Customized instruments, such as MEAs, can be crafted to possess a high degree of task alignment resulting in a valid, reliable way to assess CTS.

> What are effective approaches to evaluating critical thinking skills in a course environment?

Upon reflection on the results and challenges from the study, effective approaches to evaluating critical thinking skills in a course environment are ones that are embedded and well aligned with course outcomes. These are approaches that are well integrated, seamless and virtually indistinguishable from any other activity in the course experience and provide students with a meaningful link between CTS and their discipline. Standardized instruments are typically external to course activities, which can lead to reduced engagement and motivation for students to complete the instruments in a meaningful fashion resulting in unreliable measures. These instruments also are difficult to align with the discipline specific use of CT, as outlined in the previous paragraph. Using tools as pre- and post-tests also affects motivation and is specifically attributed to test fatigue. Lengthening the time between pre and post measures will allow sufficient time for students to reflect on and integrate CTS, reduce test fatigue and attrition, allowing additional time for engaging with tasks and receiving feedback, and avoid scheduling near final exams. These issues can be managed through the use of an authentic task for instruction, development and assessment of CTS. The authors believe that MEAs are an example of such an approach. MEAs are developed as real-world applications of discipline-specific problems and provide an easily embedded, well-aligned, rigorous, authentic way to simultaneously teach and assess CTS.

Taken together, the results of the study and the conclusions addressing the research objectives establish a set of common themes related to the use of standardized instruments for assessing critical thinking as part of a large course in a higher education institution:

1) Motivation – Pre-post-testing using standardized instruments will generally be viewed as being divorced from course activities, which may reduce motivation for completing the instruments thoughtfully, particularly at the end of the course. In cases where these instruments are not embedded in course activities, there are often large self-selection bias problems; in cases where it is embedded, there may be motivational problems.

2) Alignment – The tasks required by an instrument should represent an authentic assessment of critical thinking consistent with the course objectives. With the majority of instruments measuring CTS as a generic skill, developing and maintaining task alignment is highly important to collecting valid and reliable data.

3) Sensitivity – An instrument should possess sufficient resolution to measure changes in CTS over the duration of a typical semester-long course. An instrument possessing insufficient resolution will be of little use to determine the effectiveness of an intervention, for the purposed course or curricular improvement. If standardized instruments lack the resolution to assess CTS gains over a course, they may only be suitable for measuring long-term development for program evaluation and improvement purposes.

One approach that can resolve many of these issues is to evaluate the development of CTS by scoring a selection of student artifacts created at the beginning and end of some academic session (semester, year, or entire program) using a consistent scoring rubric. An example that has been gaining traction are the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics created by the Association of American Colleges and Universities ("VALUE: Valid Assessment of Learning in Undergraduate Education," n.d.). These allow for longitudinal assessment of student artifacts created for academic credit, reducing issues with motivation and alignment. This approach has been used to evaluate the success of initiatives, including collaborative course design, lifelong learning skill development and general education learning outcomes (Finley, 2012; Pusecker, Torres, Crawford, Levia & Lehman, 2012; Rhodes, 2012; Siefert, 2012).

This approach will be used in the future at Queen's University, as part of the HEQCO Learning Outcomes Assessment Consortium, in conjunction with a constructed response task, like the CLA+, to measure critical thinking as demonstrated in student artifacts, including MEAs, across the engineering and other participating faculties. These can be scored by trained graders at a singular point in time, reducing drift in scores due to grader disposition over a semester or year. The VALUE rubrics offer a widely used and validated approach to evaluating generic cognitive processes, including CTS, and it is possible to develop discipline-specific variants of these rubrics (e.g., to assess design process skill in engineering). A representative sample of student submissions could be selected for evaluating development to reduce the large amount of scoring that would otherwise be required in large classes. It may be possible to identify a standardized instrument that could use a rubric identical to that used to score CT using student artifacts from academic courses, ensuring alignment between the constructs.

In conclusion, the use of MEAs is a singular approach to the rigorous, authentic and sustainable development and measurement of higher-order skills. The approach to measuring critical thinking skills in engineering will likely be quite different from the approach used in other disciplines to measure critical thinking as part of regular course or program improvement activities. No matter the approach, careful consideration should be given to the conclusions outlined in this report. Continuous improvement processes are becoming frequent topics of discussion for quality improvement and accreditation, and identifying sustainable and reliable approaches for embedding assessment of generic learning outcomes and higher-order skills, such as critical thinking, should be a high priority.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, *78*(4), 1102-1134.

Adams, M. H., Whitlow, J. F., Stover, L. M., & Johnson, K. W. (1996). Critical thinking as an educational outcome: An evaluation of current tools of measurement. *Nurse Educator*, *21*(3), 23.

Arum, R., & Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.

Astin, A. W. (1993a). *What matters in college?: Four critical years revisited*. San Francisco, CA: Jossey-Bass.

Astin, A. W. (1993b). What Matters in College. *Liberal Education*, *79*(4), 4-15.

Bensley, D. A., & Murtagh, M. P. (2011). Guidelines for a Scientific Approach to Critical Thinking Assessment. *Teaching of Psychology*, *39*(1), 5-16. doi:10.1177/0098628311430642.

Blaich, C., & Wise, K. (2008). *Overview of findings from the first year of the Wabash National Study of Liberal Arts Education*. Unpublished manuscript.

Bok, D. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton, NJ: Princeton University Press.

Bondy, K. N., Koenigseder, L. A., Ishee, J. H., & Williams, B. G. (2001). Psychometric Properties of the California Critical Thinking Tests. *Journal of Nursing Measurement*, *9*(3), 309-328.

Boren, T., & Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261-278. doi:10.1109/47.867942.

Butler, H. A. (2012). Halpern Critical Thinking Assessment Predicts Real-World Outcomes of Critical Thinking. *Applied Cognitive Psychology*, *26*(5), 721-729. doi:10.1002/acp.2851.

Butler, H. A., Dwyer, C. P., Hogan, M. J., et al. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity*, *7*(2), 112-121. doi:10.1016/j.tsc.2012.04.001.

Council for Aid to Education. *2011-2012 CLA Institutional Report: Queen's University*. (2012). *2011-2012 CLA Institutional Report: Queen's University*. New York: Council for Aid to Education.

Council for Aid to Education. (n.d.). Architecture of the CLA Tasks. Retrieved from http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf

CAT Instrument Technical Information. (2010). *Tennessee Tech University*. Retrieved from http://www.tntech.edu/files/cat/reports/CAT_Technical_Information_V7.pdf

Chamberlin, M. T. (2004). Design principles for teacher investigations of student work. *Mathematics Teacher Education and Development*, *6*(1), 61-72.

Chamberlin, S. (2002). *Analysis of interest during and after model eliciting activities: A comparison of gifted and general population students*. Unpublished doctoral dissertation. Retrieved from http://docs.lib.purdue.edu/dissertations/AAI3099758/

Chan, N.-M., Ho, I. T., & Ku, K. Y. L. (2011). Epistemic beliefs and critical thinking of Chinese students. *Learning and Individual Differences*, *21*(1), 67-77. doi:10.1016/j.lindif.2010.11.001.

Daly, W. M. (2001). The development of an alternative method in the assessment of critical thinking as an outcome of nursing education. *Journal of Advanced Nursing*, *36*(1), 120-130. doi:10.1046/j.1365-2648.2001.01949.x.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior* (Reprint.). New York: Springer.

Deci, E. L., & Ryan, R. M. (2010). Self-Determination. *onlinelibrary.wiley.com*. Hoboken, NJ: John Wiley & Sons, Inc. doi:10.1002/9780470479216.corpsy0834.

Diefes-Dux, H. A., Moore, T., Zawojewski, J., et al. (2004). A framework for posing open-ended engineering problems: model-eliciting activities (pp. 455-460). Presented at the 34th Annual Frontiers in Education, 2004. FIE 2004. IEEE. doi:10.1109/FIE.2004.1408556.

Dunham, R. A. (1997). Assessing EFL Student Progress in Critical Thinking With the Ennis-Weir Critical Thinking Essay Test! *Launch Your Career*, *19*(1), 43.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. *psycnet.apa.org*. W H Freeman/Times Books/Henry Holt & Co.

Ennis, R. H., & Weir, E. E. (1985). The Ennis-Weir Critical Thinking Essay Test: An Instrument for Teaching and Testing. Retrieved from http://faculty.education.illinois.edu/rhennis/tewctet/Ennis-Weir_Merged.pdf

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). Cornell Critical Thinking Tests Level X & Level Z: Manual. Boise, ID: Midwest Publications.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. Revised edition. Cambridge, MA: MIT Press.

Facione, P. A. (1990). The delphi report. *Committee on pre-college philosophy.* American Philosophical Association.

Facione, P. A., Facione, N. C., & Blohm, S. W. (2007). *The California Critical Thinking Skills Test: CCTST*. San Jose, CA: California Academic Press.

Fawkes, D., O'meara, B., Weber, D., & Flage, D. (2005). Examining the exam: a critical look at the California critical thinking skills test. *Science & Education*, *14*(2), 117-135.

Finley, A. (2012). Reliable Are the VALUE Rubrics? *Peer Review*, *13/14*(4/1).

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, *3*(4), 430-441. doi:10.1177/104973239300300403.

Foundation for Critical Thinking. (n.d.). Retrieved from http://www.criticalthinking.org

Frank, B. M., Strong, D. S., Sellens, R., & Clapham, L. (2012). Progress with the Professional Spine: A Four-Year Engineering Design and Practice Sequence. Presented at the Proceedings of the 8th International CDIO Conference, Brisbane, Australia.

Frank, B., & Kaupp, J. (2012). Evaluating Integrative Model Eliciting Activities in First Year Engineering. Presented at the Proceedings of the 2012 Canadian Engineering Education Association (CEEA12) Conference, Winnipeg, Manitoba.

Frank, B., Strong, D., & Sellens, R. (2011). The Professional Spine: Creation of a Four-year Engineering Design and Practice Sequence. *Proceedings of the Canadian Engineering Education Association*.

Frisby, C. L. (1992). Construct Validity and Psychometric Properties of the Cornell Critical Thinking Test (Level Z): a Contrasted Groups Analysis. *Psychological Reports*, *71*(1), 291-303. doi:10.2466/pr0.1992.71.1.291.

Gasper, B. J., & Gardner, S. M. (2013). Engaging Students in Authentic Microbiology Research in an Introductory Biology Laboratory Course is Correlated with Gains in Student Understanding of the Nature of Authentic Research and Critical Thinking. *Journal of Microbiology & Biology Education*, *14*(1), 25-34. doi:10.1128/jmbe.v14i1.460.

Gokhale, A. A. (1995). Collaborative Learning Enhances Critical Thinking. *Journal of Technology Education*, *7*(1). Retrieved from http://scholar.lib.vt.edu/ejournals/JTE/v7n1/gokhale.jte-v7n1.html

Gottesman, A. J., & Hoskins, S. G. (2013). CREATE Cornerstone: Introduction to Scientific Thinking, a New Course for STEM-Interested Freshmen, Demystifies Scientific Thinking through Analysis of Scientific Literature. *CBE-Life Sciences Education*, *12*(1), 59-72.

Halpern, D. F. (2002). Teaching for Critical Thinking: Helping College Students Develop the Skills and Dispositions of a Critical Thinker. *New Directions for Teaching and Learning*, *1999*(80), 69-74. doi:10.1002/tl.8005.

Halpern, D. F. (2000). Thinking critically about critical thinking: Lessons from cognitive psychology (p. 22). Presented at the Training Critical Thinking Skills for Battle Command: ARI Workshop Proceedings.

Halpern, D. F. (2003a). The “how” and “why” of critical thinking assessment. In D. Fasko (ed.), *Critical thinking and reasoning: Current research, theory and practice*. Cresskill, NJ: Hampton Press.

Halpern, D. F. (2006). Halpern critical thinking assessment using everyday situations: Background and scoring standards. Unpublished report.

Halpern, D. F., & Riggio, H. R. (2002). Thinking critically about critical thinking: Workbook to accompany Thought and knowledge: An introduction to critical thinking. Fourth ed. Mahwah, NJ: Erlbaum.

Hart Research Associates. (2008). *How should colleges assess and improve student learning? Employers’ views on the accountability challenge*. *Association of American Colleges and Universities Report*. Washington, DC: American Association of Colleges and Universities and Hart Research Associates.

Hart Research Associates. (2013). *It Takes More Than a Major: Employer Priorities for College Learning and Student Success*. *Association of American Colleges and Universities Report*. Washington, DC: American Association of Colleges and Universities and Hart Research Associates.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77-89.

Jacobs, S. S. (1995). Technical characteristics and some correlates of the California critical thinking skills test, forms A and B. *Research in Higher Education*, *36*(1), 89-108. doi:10.1007/BF02207768.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment Facts and Fantasies. *Evaluation Review*, *31*(5), 415-439.

Klein, S. C., Liu, O. L. E., Sconing, J. A., Bolus, R. C., Bridgeman, B. E., Kugelmass, H. C. & Steedle, J. C. (2009). Test Validity Study (TVS) Report.

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

Ku, K. Y. L., & Ho, I. T. (2010a). Metacognitive strategies that enhance critical thinking. *Metacognition and Learning*, *5*(3), 251-267. doi:10.1007/s11409-010-9060-6.

Ku, K. Y. L., & Ho, I. T. (2010b). Dispositional factors predicting Chinese students' critical thinking performance. *Personality and Individual Differences*, *48*(1), 54-58. doi:10.1016/j.paid.2009.08.015.

Leppa, C. J. (1997). Standardized Measures of Critical Thinking: Experience with the California Critical Thinking Tests. *Nurse Educator*, *22*(5), 29.

Lesh, R. (1999). The development of representational abilities in middle school mathematics. In I. E. Sigel (Ed.), *Development of Mental Representation: Theories and Application* (pp. 323-350). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lesh, R. A., & Doerr, H. M. (2003). *Beyond Constructivism*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Doerr, H. M. (2000). Symbolizing, communicating, and mathematizing: Key components of models and modeling. In P. Cobb, E. Yackel & K. McClain (eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 361-384). Oxford: Routledge.

Marin, L. M., & Halpern, D. F. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity*, *6*(1), 1-13. doi:10.1016/j.tsc.2010.08.002.

MEDIA Project. (n.d.). Collaborative Research: Improving Engineering Students' Learning Strategies through Models and Modeling. Retrieved from http://www.modelsandmodeling.pitt.edu/

Moore, T., & Diefes-Dux, H. (2004). *Developing model-eliciting activities for undergraduate students based on advanced engineering content*. Presented at 34th ASEE/IEEE frontiers in education conference, Savannah, GA. doi: 10.1109/FIE.2004.1408557.

Nieto, A. M., & Saiz, C. (2008). Evaluation of Halpern's "structural component" for improving critical thinking. *The Spanish Journal of Psychology*, *11*(1), 266-274.

Norris, S. P. (1990). Effect of Eliciting Verbal Reports of Thinking on Critical Thinking Test Performance. *Journal of Educational Measurement*, *27*(1), 41-58. doi:10.1111/j.1745-3984.1990.tb00733.x.

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How Robust Are the Findings of Academically Adrift? *Change: The Magazine of Higher Learning*, *43*(3), 20-24. doi:10.1080/00091383.2011.568898.

Paul, R., & Elder, L. (2001). *The Miniature Guide to Critical Thinking*. Foundation for Critical Thinking. Retrieved from http://www.criticalthinking.org/

Paul, R., & Elder, L. (2005). *A guide for educators to critical thinking competency standards: Standards, principles, performance indicators, and outcomes with a critical thinking master rubric*. Foundation for Critical Thinking. Retrieved from http://www.criticalthinking.org/

Paul, R., & Elder, L. (2007). *Consequential validity: using assessment to drive instruction.* Foundation for Critical Thinking. Retrieved from http://www.criticalthinking.org/

Paul, R., & Elder, L. (2010). International Critical Thinking Test. Foundation for Critical Thinking.

Paul, R. (1993). The Logic of Creative and Critical Thinking. *American Behavioral Scientist*, *37*(1), 21-39. doi:10.1177/0002764293037001004.

Paul, R., & Elder, L. (1996). *Using Intellectual Standards to Assess Student Reasoning*. Foundation for Critical Thinking. Retrieved from http://www.criticalthinking.org/

Paul, R., Willsen, J., & Binker, A. J. A. (1993). *Critical thinking*. Santa Rosa, CA: Foundation for Critical Thinking.

Possin, K. (2013). *A Serious Flaw in the Collegiate Learning Assessment [CLA] Test* (pp. 1-15). Winona, MN: The Critical Thinking Lab.

Pusecker, K., Torres, R., Crawford, I., Levia, D., & Lehman, D. (2012). Increasing the Validity of Outcomes Assessment. *Peer Review*, *13/14*(4/1), 27-30.

Rashid, R. A., & Hashim, R. A. (2008). The Relationship between critical thinking and language proficiency of Malaysian and undergraduates. *Proceeding of the EDU-COM 2008 International Conference, Symposia and Campus Events*, 19-21 November 2008, Edith Cowan University, Perth Western Australia.

Rhodes, T. L. (2012). Emerging Evidence on Using Rubrics. *Peer Review*, *13/14*(4/1), 4-5.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68. doi:10.1037/0003-066X.55.1.68.

Schufried GmbH. (n.d.). *Vienna Test System*. schuhfried.com. Moedling, Austria: Schufried GmbH.

Self, B., Shuman, L. J., & Besterfield-Sacre, M. (2012). Model Eliciting Activities: Lessons Learned From a Five-Year, Seven-Institution Collaboration. Presented at the 6th International Technology, Education and Development Conference.

Shavelson, R. J. (2008). The collegiate learning assessment. In *Ford Policy Forum 2008: Forum for the Future of Higher Education*. Retrieved from http://net.educause.edu/ir/library/pdf/fp085.pdf

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Shuman, L. J. (2012). AC 2012-3847: CCLI: Model eliciting activities. Presented at the Proceedings of the ASEE Annual Conference.

Shuman, L. J., & Besterfield-Sacre, M. E. (2008). The model eliciting activity (MEA) construct: moving engineering education research into the classroom. Presented at the 9th Biennial ASME Conference on Engineering Systems Design and Analysis, Haifa, Israel.

Siefert, L. (2012). Assessing General Education Learning Outcomes. *Peer Review*, *13/14*(4/1), 9-11.

Solon, T. (2003). Teaching critical thinking: The more, the better. *The Community College Enterprise*, *9*(2), 25-38.

Steif, P. S., Lobue, J. M., Kara, L. B., & Fay, A. L. (2013). Improving Problem Solving Performance by Inducing Talk about Salient Problem Features. *Journal of Engineering Education*, *99*(2), 135-142. doi:10.1002/j.2168-9830.2010.tb01050.x.

Stein, B., & Haynes, A. (2011). Engaging Faculty in the Assessment and Improvement of Students' Critical Thinking Using the Critical Thinking Assessment Test. *Change: The Magazine of Higher Learning*, *43*(2), 44-49. doi:10.1080/00091383.2011.550254.

Stein, B., Haynes, A., & Redding, M. (2008). *Project CAT: Assessing Critical Thinking Skills Final Report*. Tennessee Tech University. Retrieved from http://www.tntech.edu/files/cat/reports/Project_CAT_Final_Report.pdf

Stein, B., Haynes, A., & Redding, M. (2006). *Project CAT: Assessing Critical Thinking Skills* (pp. 290-299). Presented at the Proceedings of the National STEM Assessment Conference, Washington, D.C.

Stein, B., Haynes, A., Redding, M., Ennis, T., & Cecil, M. (2007). Assessing critical thinking in STEM and beyond. In M. Iskander (ed.), *Innovations in E-learning, Instruction Technology, Assessment and Engineering Education* (pp. 7982). New York: Springer.

Stein, B., Haynes, A., & Unterstein, J. (2003). Assessing critical thinking skills. Presented at the SACS/COC Annual Meeting, Nashville.

Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *The Journal of General Education*, *46*(2), 129-164.

Tennessee Tech University. (n.d.). Successful Projects | Tennessee Tech University. Retrieved from http://www.tntech.edu/cat/links-to-successful-projects

Ungson, G. R., & Braunstein, D. N. (eds.). (1982). *Decision Making: An Interdisciplinary Inquiry*. Boston, MA: Kent Publishing Company.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement*, *52*(4), 1003-1017. doi:10.1177/0013164492052004025.

VALUE: Valid Assessment of Learning in Undergraduate Education. (n.d.). VALUE: Valid Assessment of Learning in Undergraduate Education. Association of American Colleges and Universities. Retrieved from http://www.aacu.org/value/rubrics/index_p.cfm?CFID=8319302&CFTOKEN=61210529

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London: Academic Press London.

Yildirim, T. P., Shuman, L., Besterfield-Sacre, M., & Yildirim, T. (2010). Model eliciting activities: assessing engineering student problem solving and skill integration processes. *International Journal of Engineering Education*, *26*(4), 831-845.

Zahner, D. (2013). Reliability and Validity – CLA. *Council for Aid to Education*. Retrieved from http://cae.org/images/uploads/pdf/Reliability_and_Validity_of_CLA_Plus.pdf

Higher Education Quality Council of Ontario

An agency of the Government of Ontario