Using Computer Simulated Science Laboratories: A Test of Pre-Laboratory Activities with the Learning Error and Formative Feedback Model

by

Man-Wai Chu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy

in

Measurement, Evaluation, and Cognition

Department of Educational Psychology
University of Alberta

Abstract

The use of computer simulated science laboratories (CSSL) to assess science knowledge and skills has become popular in recent years (Bennett, Persky, Weiss, & Jenkins, 2007; PhET, 2014). These digital environments are often superior to face-to-face environments in which traditional science laboratories are usually conducted. Traditional science laboratories have been criticized for providing a *recipe* of pre-determined linear steps for students to follow. In contrast, CSSLs encourage higher-order ideas such as scientific inquiry by allowing students to explore the laboratory (e.g., trying different procedures and making errors; Ma & Nickerson, 2006; Sahin, 2006). CSSLs have been shown to improve student achievement when used as a supplement to classroom activities (PhET, 2015; Scalise, Timms, Clark, & Moorjani, 2009; Quellmalz, Timms, & Schneider, 2009). This research study adds to this literature by investigating whether two interventions – a pre-laboratory activity and learning error intervention (LEI) – enhance students' learning and performance on a CSSL designed to measure students' science knowledge and skills. The use of pre-laboratory activities was selected for this study because many CSSLs tend to omit the use of a pre-laboratory activity, which are often used in traditional laboratories to cognitively prepare students for the experiment (Sahin, 2006; PheT, 2014). An LEI was used to encourage students to attempt multiple procedures while solving one problem during the CSSL; this differs from traditional laboratories which do not allow for much deviation from the linear steps (Bennett et al., 2007; Ma & Nickerson, 2006). These multiple trials permit the exploration of errors, which is an essential part of the learning process because it may inform future runs (Leighton, Chu, & Seitz, 2013). In order to investigate whether these interventions enhanced students' CSSL performance and associated learning competencies, a quasi-experimental design was used. The results indicated students who were

administered the pre-laboratory activity reported lower levels of test anxiety when compared with their peers who did not receive the activity. Furthermore, students who received the LEI scored significantly higher on two components of Problem 3 of a CSSL assessment and on a sub-section of a post-intervention survey measure. These findings are important because they provide evidence that both a pre-laboratory activity and a LEI can be beneficial in improving students' performance on an CSSL. Knowing the beneficial aspects of these kinds of interventions may help educators better utilize CSSLs in the classroom so that digital learning and associated assessment tools may be maximized.

Preface

This thesis is an original work by Man-Wai Chu. The research project, of which this dissertation is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name “Computer Simulated Science Laboratory Assessment,” No. Pro00040790, November 29, 2013.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor and mentor, Dr. Jacqueline P. Leighton, for her continued support of my Doctoral study and related research, and for her patience, guidance, and motivation. Her support helped me through the research and writing of this dissertation, as well as helped me develop the necessary skills needed to become a researcher.

I would also like to thank the many professors who helped me along the way, for their insightful comments and encouragements, but also for the hard questions which incented me to widen my research to other perspectives and areas.

I gratefully acknowledge the funding sources that made my Doctoral research possible. My research was supported by the Izaak Walton Killam Memorial Scholarship (2014-2016), Dorothy J Killam Memorial Graduate Prize (2014), Louise Svarich Memorial Graduate Award (2014), Queen Elizabeth II Graduate Scholarship (2013-2014), and Andrew Stewart Memorial Graduate Prize (2013).

Last, but not least, I would like to thank my family and friends for all their encouragement throughout the years.

Table of Contents

List of Tables

List of Figures

Introduction

Human Resources and Skills Development Canada recently posed the following question to the Council of Canadian Academies, an independent and not-for-profit organization that aims to facilitate evidence-based assessments to support public policy in Canada: How well is Canada prepared to meet future skills requirements in science, technology, engineering, and math (STEM)? The Council found that the national supply-and-demand for STEM skilled workers is balanced in the workforce, but were concerned about the quality and level of STEM skills held by these Canadians (Council of Canadian Academies, 2015). To address this concern, the Council indicated a need to develop STEM proficient students through high-quality programs during pre-primary education through to secondary school. Their hope is that initial investments into building fundamental STEM skills at a young age will develop higher quality STEM students; some of whom will continue onto STEM related careers. However, the types of educational programming needed to develop a high-quality STEM skilled population warrants investigation.

A source of concern about the quality of STEM skills developed by Canadian students comes in part from fluctuating international test results in their science and mathematics achievement (see Council of Ministers of Education Canada [CMEC], 2016; Statistics Canada, 2008). For example, Canadian students' science and mathematics performance on the Organization for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) has wavered over the past decade (OECD, 2016). PISA is a standardized test administered around the globe every three years to evaluate 15-year-old students on key academic competencies. For example, in 2006 Canadian students were ranked third internationally in science achievement (Statistics Canada, 2008). Only students from two

regions or countries – Finland and Hong Kong – performed better than Canadian students (Statistics Canada, 2008). In 2009, Canadian students dropped to a ranking of eighth place internationally in science achievement. Students from seven regions or countries – Shanghai, Finland, Hong Kong, Singapore, Japan, Korea, and New Zealand – outperformed Canadian students (Statistics Canada, 2010). The 2012 PISA results show Canadian students dropping to 10th place internationally in science achievement (CMEC, 2016). An improvement in Canadian students' science performance was reported in the most recent 2015 PISA results in which they ranked 7th internationally (OECD, 2016). The 2015 PISA results indicate that students from only six regions or countries – Singapore, Japan, Estonia, Chinese Taipei, Finland, Macao-China – performed better can Canadian students in estimated average scores. Although Canadian students' science achievement has been fluctuating over the past decade, they still score well above the OECD average.

A comparison between Canadian students' science PISA scores from 2006 and 2015 are not significantly different from a statistical perspective. However, some regions or countries have moved past Canada, which can make it look like Canadian students' science achievement is declining. This decline has prompted the Council of Canadian Academies, which includes educators, researchers and policymakers, to investigate the state of science education in Canada in order to find ways to improve its international competitiveness. Such investigations must include exploring and learning from other educational systems. However, it is necessary to begin by investigating gaps in Canada's teaching and assessment of science knowledge and skills.

Science Education: Purpose and Limitations

The purpose of science education, according to the National Research Council (2014), is to develop scientifically literate students who are capable of thinking critically and making

informed decisions on science and technology issues in order to function effectively in an increasingly complex society. As researchers and educators turn their focus toward STEM-related courses, some have argued that this goal will not be achieved because of the fundamental problem of *how* science is currently taught in schools (National Research Council, 2006). In particular, they criticize one key aspect of the science curriculum: the weak laboratory experiences encountered in classes (National Research Council, 2006, 2014). Laboratories are important to science education because they "enable students to interact *intellectually* as well as *physically*" in the classroom while observing evidence of the claims they generate about the world (Hofstein & Lunetta, 2003, pg. 49). However, critics have said the laboratories suffer a number of shortcomings, due to (1) a "disconnect… from the way science and engineering are practiced" because of a "lack of adequate instructional time and adequate space and equipment for investigation and experimentation" (National Research Council, 2014, pg. 13) and (2) the fact that "current large-scale assessments are not designed to accurately measure student attainment of the goals of laboratory experiences" (National Research Council, 2006, pg. 9). These gaps speak to both the teaching and assessment of science laboratories.

Correcting the first problem requires bringing real life examples, scenarios, and both hands-on and minds-on activities into the classroom to help students learn applicable knowledge and skills (Hofstein & Lunetta, 2003). *Hands-on* laboratories follow the traditional model of science laboratories, which emphasize a pre-specified procedure to be followed. By contrast, *minds-on* laboratories also include a component in which students are engaged in a dialogue involving the processes of scientific inquiry and the knowledge derived through these methods (National Research Council, 2006). The minds-on component is important because it encompasses one of the goals of science education, namely to provide students with

opportunities to learn the processes and dynamic relationships between empirical research and scientific discoveries (National Research Council, 2006). Learning these processes allows students to practice the scientific inquiry skills used in the real world, such as in the professions of science and engineering.

These hands-on and minds-on laboratories are challenging to design, however, because although many technical skills can be mimicked in a hands-on environment, some real-world minds-on skills are difficult to simulate. This difficulty stems in part from the lack of opportunities students have in a science laboratory to develop skills such as using the results of previous experiments to inform changes to future experiments. Real-world scientists have the opportunity to try out different approaches for implementing various scientific methods, which can entail making errors and repeating experiments with enhancements on each subsequent trial to refine and develop their final method. This lack of realism in hands-on science laboratories in the classroom may inadvertently cause students to view science laboratories as a closed-ended task with a single correct method that always leads to the right answer after the first attempt (Hofstein & Lunetta, 2003). That is, students typically try to follow what they believe to be the correct method, instead of using the laboratory as a way of exploring, gathering evidence for, and focusing on, the process of developing general scientific knowledge and skills. This idea of a single correct answer is further fuelled by the fact that many science laboratories in the classroom tend to be conducted to confirm known facts and theories taught in class instead of focusing on the scientific inquiry process used during laboratories.

Addressing the second problem, that large-scale assessments are not designed to accurately measure student attainment of laboratory experience goals, and require a dynamic large-scale assessment tool that can capture students' interactions during the laboratory and

assess their performance-based knowledge and skills (National Research Council, 2014). Currently, many classrooms assess students' science laboratory skills through careful teacher observations and checklists, as well as scoring post-laboratory write-ups. These forms of assessment are problematic because teachers may observe only a portion of each student's activities; that is, while they are monitoring several students at once, teachers may miss observing the full performance of any individual student. Moreover, the use of observations and checklists, especially when based on a portion of performance, can lead to teachers assigning different grades to the same full laboratory performance, thereby foregoing the consistency these instruments are designed to provide. These classroom-based assessment tools also require a considerable amount of resources to administer and score. Although laboratory assessments are difficult to assess on a large scale, they are vital to informing educators of student progress in terms of their understanding of the relationship between empirical research and development of scientific knowledge; in other words, student understanding of how the field of science generates knowledge. Consequently, there is a need to improve the tools used to assess student learning in laboratory experiences. Since science laboratories are performance-based tasks, it is important that their assessment includes a method to track student progress through the laboratory, as well as student understanding at the end of the experience (National Research Council, 2014). As such, new assessments of laboratory skills need to be dynamic so that student interactions may be captured and assessed throughout the task.

Computer Simulated Science Laboratories (CSSLs)

To overcome both of these problems with hands-on and minds-on science laboratories, *computer simulated science laboratories* (CSSLs) have been developed and are popular in science classrooms as a way to supplement science laboratories (Ma & Nickerson, 2006). CSSLs

are computer programs that contain models of hands-on science laboratories or scientific processes (DeJong & VanJoolingen, 1998). They bring a new level of dynamic interaction to the science classroom by mimicking science laboratories in a realistic way so that students may learn different laboratory knowledge and skills necessary in the real world through the use of a simulated environment (Sahin, 2006). CSSLs provide students with a digital medium to simulate the activities that take place during hands-on science laboratories, the latter of which may sometimes include elements that are considered impractical, expensive, impossible, or too dangerous to run in a science classroom (Strauss & Kinzie, 1994).

CSSLs have potential to help students develop an understanding of fundamental science knowledge and skills. Dwyer and Lopez (2001) indicate that CSSLs may be able to contribute to conceptual change while providing open-ended problem solving experiences and tools of scientific inquiry. Scientific inquiry, as the National Research Council (1996) notes, "refers to the diverse ways in which scientists study the natural world and propose explanations based on the evidence derived from their work" (pg. 23). Studying diverse types of scientific inquiry requires the ability to try different methods, such as using the results of previous trials to inform later attempts, which creates an iterative cycle.

This iterative process of scientific inquiry, in which students are encouraged to take the results of their first experiment to enhance their second experiment, is facilitated by the use of CSSLs in a way that may not be possible in regular classrooms. In regular classrooms where hands-on, science laboratories are performed, students seldom have the opportunities to repeat their experiments because of limited resources (e.g., equipment). In addition, CSSLs allow students to complete the laboratories relatively quickly because the time needed to gather and set-up the laboratory equipment is minimized. All the equipment may be already set up on a

CSSL or can be easily set up with a few mouse clicks so more time can be devoted to the process of scientific inquiry (National Research Council, 2014). Despite the many benefits of CSSLs, proper implementation of these tools is vital to maximize student learning using these resources (PhET, 2015).

Implementation of CSSLs as learning tools. The use of CSSLs as a supplement to hands-on laboratories appears to have many benefits, such as allowing students to explore and experiment with knowledge and skills in ways that would otherwise not be possible (PhET, 2015; Shute, 2013). Quasi-experimental research focusing on the use of simulations as a tool to support instruction, and on the instruction required to support the simulations has indicated that CSSLs can be useful in deepening students' understanding as measured by increased achievement test scores in experimental versus comparison groups (Barb, Dodge, Ingram-Goble, Pettyjohn, Peppler, Volk, & Solomou, 2010; Coller & Scotee, 2009). However, one criticism of these studies is that they approach the use of simulations without also considering the necessity and impact of supporting tools (e.g., teacher support or follow-up activities; Rutten, van Joolingen, & van der Veen, 2012).

Considering CSSLs have the potential to help teach students the scientific knowledge and skills underwriting scientific inquiry in a laboratory setting, it is important to set up students for success in learning these processes. In order for CSSLs to succeed as a tool to teach inquiry skills, which are considered a higher-level application of knowledge, students need to have a good understanding of the basic content domain knowledge before using CSSLs (Shute, 1993). For example, students need to understand the basic knowledge of motion if they are to apply it to a laboratory investigating acceleration. One way to enhance students' application of knowledge and skills is to have students review the necessary content prior to the CSSL or have access to it

during the CSSL. Traditional hands-on science laboratories tend to prepare students by reviewing the necessary content knowledge prior to the activity in the form of a *pre-laboratory activity*.

A pre-laboratory activity would normally require students to (a) think about the objectives, procedures, tools and goals of the laboratory before being confronted with the (simulated) laboratory environment, equipment and procedures and (b) review the basic content knowledge needed for the laboratory in order to prime students to make meaningful connections between content knowledge and the laboratory exercise (Wilkenson & Ward, 1997; Hodson, 2003). Although some CSSLs are able to provide students with on-demand access to content knowledge throughout the simulation in the form of a *help* or *glossary* button, it is also beneficial to have students review the necessary content knowledge prior to the laboratory so that they start the laboratory with requisite knowledge of the topic. Despite the benefits of using pre-laboratory activities in traditional hands-on laboratory settings, research studies using CSSLs seldom use these to prepare students to begin working in digital learning environments (PhET, 2015; Wilkenson & Ward, 1997). Hence, research into the use of pre-laboratory activities in conjunction with CSSLs should be investigated.

CSSLs as dynamic assessments. Developing a strong foundation of science content knowledge is important for success in the field, but equally important is an understanding of scientific inquiry, which explains the process of how scientists came to know these theories. While there are many tools to assess students' conceptual understanding of content knowledge, there are very few feasible tools to assess the process of science inquiry. Hence, there is a need for assessment tools that can capture the process of students completing a task designed to measure science inquiry.

Although one of the touted strengths of CSSLs is that they provide students with a digital learning environment to emulate the science inquiry process, it may also be beneficial to consider CSSLs as an assessment tool. CSSLs may be viewed as promising *hosts* for dynamic measurement tools designed to track what students are doing and thinking as they respond to the activities presented on the computer. In this way, CSSLs may bring greater alignment in the learning and assessment environments for students. Additionally, using simulations to teach and assess students is relatively common in post-secondary education and workplace accreditation, such as in the field of medicine (Scalese, Obeso, & Issenberg, 2008). Hence, using simulations in K-12 education could help prepare students for future assessments they may encounter in post-secondary and workplace environments (Cisco, Intel, & Microsoft, 2008).

Hofstein and Lunetta's (2003) review of hands-on science laboratories revealed that assessments of students' laboratory knowledge and skills were "seriously neglected" (pg. 47). Eleven years later, the National Research Council (2014) agreed that science laboratory assessments continue to be under-represented in the literature and struck a committee to develop an approach to science assessment that would support new curricular standards and frameworks (National Research Council, 2006, 2012; Next Generation Science Standards Lead States, 2013). The committee highlights the use of technological advances in performance-based science assessments because of their capabilities to provide different assessment formats (National Research Council, 2014; Scalise, 2009). Assessment formats such as CSSLs allow for *rich data* (i.e., complex user interactions captured using a range of digital technology devices) to be collected which could be used as evidence of students' knowledge and skill acquisition. CSSLs are generally carried out quickly and allow for multiple trials. These benefits, coupled with the assessment of students' abilities to plan and carry out an investigation and to analyse data, allow

for multiple pieces of evidence to be collected in a short amount of time. Using multiple sources of evidence would be expected to increase the reliability and validity of the evaluations; specifically, the development of more comprehensive score reports of student learning. For example, the CSSL score report could provide detailed information regarding students' graphing skills by indicating the number of attempts made, types of errors made during each attempt, and whether the correct graph was submitted as a final product.

CSSL assessment tools may be a significant departure from the traditional laboratory assessments to which students are accustomed. For example, the ability to repeat an experiment using different methods or use multiple trials for an experiment may be novel to students who are used to traditional hands-on experiments. Hence, there may be a need to explicitly educate students regarding scientific inquiry and the necessity of approaching problems using different methods and repeating experimental trials. One aspect of scientific inquiry is the necessity of *learning errors* during a laboratory experiment so that students learn from their previous trials, or errors, to inform necessary changes to subsequent trials (Firestein, 2016). In these cases, making an error or mistake is considered to be a natural part of the formative phase of learning that is encouraged as opposed to only focusing on a single right answer during exploratory training phases. Hence, an explicit learning error intervention may be needed when using CSSLs to encourage repeating experimental trials so that previous attempts may inform later trials.

Purpose of Study

Although science laboratories are viewed as a necessity in science education, the teaching and assessments associated with them require continued research and understanding (Hofstein & Lunetta, 2003; Ma & Nickerson, 2006; National Research Council, 2006, 2014). Hands-on science laboratories often miss the crucial aspect of the minds-on component in which students

reflect upon their actions and the authenticity in which students experience real-world science. The assessments of these laboratories traditionally include observations made (Ma & Nickerson, 2006) and/or write-ups scored by teachers (Doran, Boorman, Chan, & Hejaily, 1993; Tamir, Nussinovitz, & Friedler, 1982). However, this form of assessment can be problematic when considering the reliability of the scores generated, as teachers may use different criteria to evaluate observations and reports; thus, assigning different scores to similar performances. Hence, changes to the current approach toward teaching and assessing science laboratories may be needed to maximize laboratories as tools for learning. CSSLs may provide students with a flexible learning environment that allows engagement in scientific inquiry and with dynamic assessments designed to measure their acquisition of laboratory knowledge and skills. Although CSSLs have many benefits, the implementation of these tools requires investigation in order to maximize their potential.

The primary aim of the present research is to investigate the effects of two interventions – a pre-laboratory activity and a learning errors intervention – on students' socio-emotional experience with a CSSL, along with their corresponding science knowledge and skill acquisition as measured during the CSSL. This dissertation is divided into five chapters. First, an overview of science education and assessment is provided. Second, a review of computer-based or digital assessments used to measure performance-based knowledge and skills is presented. Additionally, CSSLs are presented as a potential medium to improve the implementation of hands-on and minds-on science laboratories and to dynamically measure performance-based knowledge and skills. Third, a detailed description of the method used to guide this study is provided. Fourth, the results of the study are presented. Last, a discussion of the results is provided.

Objectives of the Current Study

The primary objective of this dissertation study is to investigate two intervention treatments that may enhance the use of CSSLs as a teaching and assessment tool to measure students' science knowledge and skills. The CSSL that will be used in this research is called: The National Assessment of Educational Progress' (NAEP) Problem-Solving in a Technology Rich Environment (TRE) science laboratory (or NAEP TRESim for short), which will be described in detail later (Bennett, Persky, Weiss, & Jenkins, 2007). Three specific research questions guide this research:

(a) What are the effects of a pre-laboratory activity on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(b) What are the effects of a LEI on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(c) What are the interactions between the pre-laboratory activity and LEI on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

The students' socio-emotional experiences referred to in these research questions will be operationalized, in this study, using the variables school engagement, motivational goals, and student anxiety.

Literature Review

Science laboratories have been instrumental in the development of successful science education programs by providing students with opportunities to interact with natural phenomena or collect data in the real world using a variety of tools and techniques (National Research Council, 2006, 2014). Through these encounters, students may design an investigation, engage in scientific reasoning, manipulate equipment, record data, analyze results, and discuss findings as a means towards developing scientific inquiry skills. Scientific inquiry skills include not only asking questions and conducting experiments, but also considering the actions taken and reasoning throughout these experiences to understand the natural world; in other words, a *minds-on* approach (National Science Teachers Association, 2004). Although these laboratories are continually evolving in response to historical events and changing societal views, they have remained a vital and irreplaceable aspect of science education (National Research Council, 2006). However, the importance of science laboratories warrants investigation to continually enhance this educational tool so that students may maximize its benefits (Hofstein & Lunetta, 2004; Ma & Nickerson, 2006; National Research Council, 2006). This literature review explores the evolution of laboratories from physically tangible settings to digitally simulated environments while highlighting the shortcomings of each. The changes to these science laboratories are explored from the perspectives of viewing them as both a tool for learning and assessment.

This review is divided into three parts. First, the purpose of science education and the vital role science laboratories play are reviewed. Second, the challenges in terms of using both physical and simulated laboratories as learning and assessment tools are addressed. Third, this

review concludes with a rationale and proposal for two interventions that may help overcome the shortcomings of, and provide enhancement to, simulated science laboratories.

Purpose of Science Education

Many purposes of science education have arisen throughout history to fit with changing views of science and society's demands for science education. Prior to 1950, science was viewed as an inductive process (i.e., proceeding from observations to reach general conclusions); this led to facts and theories being taught to students through rote memorization and recitation of textbooks and lectures aimed at preparing graduates for science education in post-secondary institutions (National Research Council, 2006). After the two World Wars, there were concerns that science education was not rigorous enough to prepare future scientists and engineers to defend national interests. Hence, science education program were overhauled to focus on the process of scientific discovery, which was viewed as involving both inductive and deductive reasoning (i.e., developing specific inferences from known scientific facts and theories).

Beginning in the mid-1970s, there was a shift in the purpose of science education. Science education was deemed important for everyone and not just those wanting to become scientists. Instead of being a topic or subject for a select group, the purpose of science education was to develop scientific literacy in all students. To make informed decisions about the world around them, all students needed to increase their awareness and understanding of the natural world and develop their scientific reasoning (National Research Council, 2006; Rutherford & Ahlgren, 1990). More recently, the focus of science education has moved towards preparing a scientifically literate population that is proficient in science, technology, engineering, and mathematics (STEM) subjects (National Research Council, 2014).

A goal of STEM is to develop a scientifically literate population "who are better able to make decisions about personal health, energy efficiency, environmental quality, resource use, and national security" (Bybee, 2010, para. 5). To achieve these goals, STEM education is designed to approach the instruction of these four subject areas (i.e., science, technology, engineering and mathematics) holistically by focusing on the interdisciplinary and cohesive nature of the disciplines, while at the same time applying rigorous academic concepts in real-world contexts (Tsupros, Kohler, & Hallinen, 2009).

The holistic approach is important for teaching STEM subjects because these subjects are deeply intertwined in the real world and thus classroom-learning environments must reflect what occurs in everyday life. This need for a cross-curricular approach towards science education has led to the idea of a three-dimensional science curriculum centered around: (a) the *practice* through which scientists and engineers do their work, (b) the key *crosscutting concepts* that link the science disciplines (i.e., life science, physical science, earth, and space sciences, and engineering and technology), and (c) the *core ideas* of the different disciplines. These dimensions were developed by a committee of researchers and educators from the National Research Council, the National Science Teacher Association, the American Association for the Advancement of Science, and Achieve, as part of a two-step process. The first step involved the development of *A Framework for K-12 Science Education: Practice, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012), which is grounded in research on science and learning science aimed at identifying the science knowledge and skills K-12 students should know. This step laid the groundwork for the second step, which involved the development of *Next Generation Science Standards: For States, By States* (Next Generation Science Standards Lead States, 2013). The Next Generation Science Standards lists content and practice arranged in

a coherent manner across disciplines and grades to provide students with an internationally-benchmarked science education. To achieve the goals of these dimensions indicated in the framework and standards, the committee recommended the use of performance events, such as science laboratories, to illustrate how science is done in the real world (National Research Council, 2012).

Science Laboratories

Science laboratory experiences may involve a wide spectrum of activities for students, ranging from carrying out specified procedures, verifying established scientific knowledge to formulating questions, designing investigations, and creating and revising explanatory models. The definition of science laboratories is broad, as these experiences are expected to "provide opportunities for students to interact directly with the material world (or with data drawn from the material world), using the tools, data collection techniques, models, and theories of science" (National Research Council, 2006, pg. 3). This definition of laboratories focuses on *both* science process and content, two aspects that are critical to improving scientific literacy and preparing the next generation of scientists and engineers (National Research Council, 2006).

Many science laboratories tend to focus mainly on verifying known scientific facts and theories in which students perform specified procedures in *hands-on* environments designed to mimic real-life laboratories (Hofstein & Lunetta, 2004; Ma & Nickerson, 2006). These types of traditional laboratories tend to resemble "cookbook" activities in which students follow a series of prescribed steps in order to arrive at a correct answer (Chu, 2010; Domin, 1999). However, newer laboratories tend to include a *minds-on* aspect, which allows students to consider their actions and reasoning throughout the experience (National Research Council, 2006). This approach to developing scientific knowledge and skills, as previously noted, is called *scientific*

inquiry. This is important in science education because it focuses on *how* knowledge is created in the field of science. It is now widely established that science inquiry is a key skill in developing scientific literacy (National Research Council 1996). Science laboratories can provide hands-on investigations and minds-on reflections that "enable students to interact intellectually" (Hofstein & Lunetta, 2004, p. 49) and think critically to develop as scientifically literate citizens who can solve scientific problems. Although researchers agree on the importance of both science process and content for the development of science education, designing science laboratories that provide opportunities for students to experience both process and content has proven difficult (Hofstein & Lunetta, 2004; National Research Council, 2006).

Challenges of designing traditional hands-on and minds-on science laboratory experiences. As noted previously, laboratories are often designed to provide students with opportunities to mimic the knowledge and skills used outside the classroom. Although many technical skills can be mimicked in a hands-on science laboratory, some real-world minds-on skills are difficult to simulate; the difficulty of designing these laboratories stems in part from the lack of resources (e.g., equipment and time) and a focus on the correct *answer*. While the result of a laboratory is important, the process of arriving at that piece of knowledge is equally valuable. However, many of the traditional science laboratories used in classrooms today tend to focus more on the product than the process (Ackroyd et al., 2007).

Many real-world scientists have the opportunity to experiment with different approaches for implementing various scientific processes, which can entail making errors and repeating experiments with enhancements on each subsequent trial to refine and develop their final method (Firestein, 2016). However, hands-on science laboratories in the classroom often do not allow for this iterative process of inquiry; instead, they tend to follow the scientific method in a linear

sequence (Hodson, 1998, 2003). This sequence, for instance, often includes the following steps: (i) ask a question, (ii) do background research, (iii) identify a hypothesis, (iv) test the hypothesis by doing an experiment, (v) analyze the data and draw conclusion, and (vi) communicate results (Chalmer, 1999). The final step often requires students to think about the scientific method they used during the laboratory experience, thus reflecting on the process of inquiry and the objectives of a minds-on laboratory. However, this linear approach has been criticized as an inadequate reflection of what scientists actually do because it focuses on a single scientific route to reaching an answer, as opposed to the many routes or methods which might be considered and used to solve the same problem in the real world. Students are not given an opportunity to learn from their errors or attempt different methods during the experiment due to concerns about limited laboratory resources (e.g., solvents for a chemistry lab or organs for a biology dissection), time (i.e., most class periods do not allow students opportunities to repeat their experiment), or getting the wrong answer (i.e., in relation to the correct answer which is the known scientific fact they were expected to verify; Hodson, 1998; Hofstein & Lunetta, 2004).

Challenges of designing hands-on and minds-on science laboratory assessments. In addition to the challenges that are encountered during the design of hands-on and minds-on science laboratories, there are also challenges involved with the assessment of students in these laboratories (Hofstein & Lunetta, 2004; National Research Council, 2006, 2012, & 2014). Science laboratories tend to offer a classroom-based performance experience or opportunity in which the assessments are typically the responsibility of teachers. As such, teachers must assess both students' performance during the lab along with their conceptual understanding after the laboratory experience has taken place. During the laboratories, student performances are often assessed using teachers' observations with the help of a checklist or rubric listing the criteria of

skills in which students should be proficient (Ma & Nickerson, 2006). However, with average secondary-level class sizes of approximately 26 students (Albert's Commission on Learning, 2003), it is difficult for one teacher to observe the performance of all students reliably during the whole class period. Additionally, explicit observations of implicit skills such as minds-on understanding of scientific inquiry are difficult to record because these skills may not always show as external indicators of student mastery. After the laboratory is complete, students' performances are also often assessed using their written reports (e.g., laboratory write-up) aimed at measuring their understanding of the knowledge and skills they have gained (Doran, Boorman, Chan, & Hejaily, 1993). However, even written reports may be difficult for teachers to assess reliably if the laboratory has not provided the same opportunity for all students to engage in minds-on understanding, particularly when students complete laboratories within a group setting. In other words, both of these assessment methods, observations and written reports, are problematic as teachers may assign grades without opportunities to observe all students at similar time points or without ensuring that all students have had opportunities to conceptually consider the material. These assessment methods, thus, potentially compromise the reliability of the scoring and validity of inferences. The apparently subjective nature of the scoring involved with these assessments of students' laboratory performances warrants investigation so that more reliable and valid measures may be developed and used (Hofstein & Lunetta, 2004).

In addition to classroom-based assessments lacking strong reliability and validity, laboratory skills also present challenges for integration or inclusion on large-scale assessments, such as provincial or state achievement testing. Although laboratory skills are an important aspect of science education and the National Research Council (2006) emphasized their value to help students gain a deeper understanding of science knowledge and skills, large-scale

assessment of these skills is presently absent. Thus, not only are current classroom-based assessments not able to measure these skills reliably and validly, but large-scale assessments are not able to include measurement of these skills at all (National Research Council, 2014). For this reason, there is a need for study into the development of science laboratory assessments that can enhance the student learning experience, reliably capture students' performance and understanding of the laboratory, as well as allow for efficiency in administration. In order to determine an appropriate assessment method that is best able to satisfy these requirements, a review of two select performance assessments is presented in the next section.

Key Performance Assessments

Every year, digital educational assessments are refined and enhanced so that they are more effective in measuring students' learning and achievement. In particular, two forms of assessments – embedded assessment (EA) and evidence-centered game design (ECgD) – may offer particular benefits to measuring performance-based laboratory knowledge and skills. The next sub-sections offer a discussion of both forms of assessment, including their potential strengths for measuring laboratory performances.

Embedded assessment (EA). EA is a method designed to integrate the measurement of students' knowledge and skills within the learning environment (Wiliam, 2011). This method involves combining processes of teaching, learning, and assessment so that they are virtually indistinguishable (Wilson & Sloane, 2000). This integration requires assessment and instruction to be designed together so that each piece complements the other (Wilson & Sloane, 2000). Designing assessment activities during the initial planning of instruction is one way in which instruction can inform assessment and vice versa (William, 2011). In contrast, when assessment

activities are developed *at the end* of instruction planning, their meaningfulness is often lost (William, 2011).

An example of a course designed using an EA framework is the *Science Education for Public Understanding Project (SEPUP)*. The SEPUP is a science course designed to help junior-high level students (aged 12-15 years) focus on using science evidence to solve social and ethical issues (Wilson & Sloane, 2000). The activities, or EAs, used throughout the course were designed by the participating teachers and researchers to teach and assess five learning variables (or objectives): (1) understanding concepts; (2) designing and conducting investigations; (3) considering evidence and trade-offs; (4) communicating scientific information; and (5) participating in group interaction. Careful attention was paid to ensure the learning and assessment activities were designed together to prevent disconnection between the learning and assessment phases. This EA-based course example will be used throughout this section to highlight the different aspects of EA.

Although EAs aim to integrate instruction and measurement of learning, the use of EAs in the classroom is still in its infancy. Nonetheless, the expectation is that EAs will allow educators to measure and monitor students' learning using activities whose design is based on educational frameworks such as instructional scaffolding (i.e., a temporary framework aimed at promoting learning when knowledge and skills are first being introduced to students; Sawyer, 2006) and learning progressions (i.e., a framework for developing assessments aimed at moving students from novice toward expert understanding over time; Nichols, 2010). When EAs are developed and guided by these frameworks, educators are expected to have better assessment methods to monitor students' progress and performance during the learning process instead of at the end. These results may then be used to inform educators of students' learning and

achievement throughout the course (Wilson & Sloane, 2000). Tracking this progression of learning would be useful to ensure students have a strong core of knowledge as they build the higher-level thinking skills required for scientific inquiry.

Some similarities exist between EA and formative assessment; the latter of which has been defined by Popham (2008, p. 6) as "a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics." However, EA and formative assessment differ in how they may be applied to lessons and used in grading. First, formative assessments are typically administered during a lesson, but are not required to be seamlessly integrated into the learning experience (Shute, 2008; Wilson & Sloane, 2000). For example, a mid-term quiz may be administered in class as a formative assessment, but the teaching and learning are paused so that students may write the test. In contrast, EAs would be designed so that teaching and learning do not need to be paused for the assessment. For example, students may complete a two-day science laboratory in which the first day is used to develop the experimental plan while the second day is used to execute the plan. In between the two days, the teacher may assess students' plans and provide feedback; but from students' perspectives, their laboratory experiment was not paused for an explicit assessment. Second, students' performances on formative assessments are typically not used in the calculation of a final grade (O'Connor, 2010). This is distinct from EAs as the scores attained on EAs may be used to calculate students' grades (Office of Assessment Services, n.d.). EAs have specific strengths that may make them suitable for assessing students' science laboratory performances.

Strengths of embedded assessment (EA). The construct maps (e.g., test blue prints or specifications that guide the development of high quality assessments; Schmeiser & Welch,

2006) underlying the development of EAs typically reflect an ordering of qualitatively different levels of performance, focusing on increasing levels of knowledge structures or skills listed in the program of studies (Wilson, 2004). The expectation is that these construct maps are empirically grounded while at the same time being guided by the program of study outcomes (Wilson & Sloane, 2000). For example, during the initial planning stages of the SEPUP course, participating teachers and researchers mapped the five variables - (1) understanding concepts; (2) designing and conducting investigations; (3) considering evidence and trade-offs; (4) communicating scientific information; and (5) participating in group interaction - to the program of studies to ensure necessary elements of the state-mandated science program were taught. The participating teachers also broke down each of the five variables into smaller elements in an attempt to operationalize each one. For example, the *designing and conducting investigations* variable was broken down further into four elements (1) designing investigations; (2) selecting and recording procedures; (3) organizing data; and (4) analyzing and interpreting data. Each of these elements was measured using a variety of activities and scored using an analytic rubric so that teachers were able to track students' progress on each element throughout the course. Figure 1 shows a set of seven activities that were designed to assess the five variables listed previously and part of their corresponding elements. Some of the activities (e.g., John Snow and Search for Evidence) assessed two of the five variables at once (i.e., evidence and tradeoffs and communicating scientific information). These sets of activities were administered throughout the course so that each held a designated place in the instructional flow to allow students multiple opportunities throughout the year to provide evidence of their learning on each of the objectives or variables.

Activity	Variables and Elements				
	Designing and Conducting Investigations (DCI) *Designing Investigation *Selecting and Recording Procedures *Organizing Data *Analyzing and Interpreting Data	**Evidence and Tradeoffs (ET)** *Using Evidence *Using Evidence to Make Tradeoffs	**Understanding Concepts (UC)** *Recognizing Relevant Content *Applying Relevant Content	**Communicating Scientific Information (CM)** *Organization *Technical Aspects	**Group Interaction (GI)** *Time Management *Role Performance / Participation *Shared Opportunity
1 - Water Quality					
2 - Exploring Sensory Thresholds			√: Both Elements (Measurement and Scale)		
3 - Concentration			√: Applying Relevant Content		
4 - Mapping Death					√: Time Management; Shared Opportunity
5 - John Snow and Search for Evidence		A: Using Evidence		A: Both Elements	
6 - Contaminated Water	√: Designing Investigations				
7 - Chlorination	A: All Elements				

Figure 1. Partial assessment blueprint for SEPUP course. Adapted from "From Principles to Practice: An Embedded Assessment System," by M. Wilson and K. Sloane, 2000, *Applied Measurement in Education, 13*, p. 195. Copyright 2000 by Lawrence Erlbaum Associates, Inc. Reprinted with permission.

After the students responded to the activities, rubrics guided the rating and scoring of students' performances on a set of categorical outcomes (Wilson & Sloane, 2000). Each rubric score was linked to a specific activity that was related to a portion of the construct map. To increase the reliability of marking using these rubrics, participating teachers in the SEPUP course met periodically to score student work as a group. The rubrics used to rate and score students' laboratory performances informed the level of student achievement on a specific learning outcome. For example, a rubric that was used during the SEPUP course, shown in Figure 2,

indicates the performance students needed to demonstrate so that they can achieve a specific score on the variable *evidence and tradeoffs*. In the activity that made use of this rubric, students were asked to evaluate the advantages and disadvantages of different solutions to a problem based on available scientific evidence (Wilson & Sloane, 2000). Rubrics similar to this one were used throughout the course for many of the EAs.

Evidence and Tradeoffs (ET) Variables

Score	***Using Evidence:*** Response uses objective reason(s) based on relevant evidence to support choice.	***Using Evidence to Make Tradeoffs:*** Response recognizes multiple perspectives of issue and explains each perspective using objective reasons, supported by evidence, in order to make choice.
4	Response accomplishes Level 3 AND goes beyond in some significant way, such as questioning or justifying the source, validity, and/or quantity of evidence.	Response accomplishes Level 3 AND goes beyond in some significant way, such as suggesting additional evidence beyond the activity that would further influence choices in specific ways, OR questioning the source, validity, and/or quantity of evidence & explaining how it influences choice.
3	Response provides major objective reasons AND supports each with relevant & accurate evidence.	Response discusses at least two perspectives of issue AND provides objective reasons, supported by relevant & accurate evidence, for each perspective.
2	Response provides some objective reasons AND some supporting evidence, BUT at least one reason is missing and/or part of the evidence is incomplete.	Response states at least one perspective of issue AND provides some objective reasons using some relevant evidence BUT reasons are incomplete and/or part of the evidence is missing; OR only one complete & accurate perspective has been provided.
1	Response provides only subjective reasons (opinions) for choice and/or uses inaccurate or irrelevant evidence from the activity.	Responses states at least one perspective of issue BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence.
0	No response; illegible response; response offers no reasons AND no evidence to support choice made.	No response; illegible response; response lacks reasons AND offers no evidence to support decision made.
X	Student had no opportunity to respond	

Figure 2. An evidence and tradeoffs scoring guide for the SEPUP course. Adapted from "From Principles to Practice: An Embedded Assessment System," by M. Wilson and K. Sloane, 2000,

Applied Measurement in Education, 13, p. 193. Copyright 2000 by Lawrence Erlbaum Associates, Inc. Reprinted with permission.

After students' responses were marked using the rubrics, the scores from activities designed to measure similar variables were grouped together in a map that allowed teachers and students to identify their progress in a specific area. For example, the score report shown in Figure 3 displays the marks throughout the year for types of activities (e.g., pre-tests, A & B 1-12, and C 13-20) that measured the *evidence and tradeoffs* variable. Students' progression of learning based on these activities was also tracked using the *developmental levels* rubric (shown in the right-most column of Figure 3). The score reports represent student progress by indicating the skills they have mastered and those that need further refinement.

Pre-tests	Part 1: Water				Part 2: Materials Science			Part 3: Energy		Post-Tests	SEPUP Scale Score	Developmental Levels
	A & B 1-12	C 13-20	D 21-28	Link 1	A 29-38	B 39-46	Link 2	47-58	Link 3			
											2000	Level 4 *Goes beyond Level 3 in significant way*
											1950	
						16	12				1900	
				20				16				
			8								1850	
		8			12	15	11					
				19				15			1800	
16	12					14				20		
							10		12		1750	
		7	7	18	11	13		14				Level 3
15	11			17			9			○	1700	*Correct and Complete*
						12			11			
14	10			16	10	11	8	13		18	1650	
		6	6	15								
13					9	10	7	12	○	17	1600	
	9			14		9	○					
12		5		13	8	8		○		16	1550	
	8		5	12			5		9	15		
11		4		11	7	○	4	10		14	1500	Level 2
10	7		○	○		6		9	8	13		*Correct but important part missing*
9		3		9	6	5	3	8	7	12	1450	
	6			8	5	4				11		
8			3	7		3	2	7	6	10	1400	
7	5	2		6	○			6		9		
6	○		2	5	3	2	1	5	5	8	1350	
5		○		4				4	4	7		
○	3			3	2	1	0	3		6	1300	Level 1
3			1						3	5		*On task but incorrect*
2	2	0		2		0		2		4	1250	
					1				2	3		

1	1		0	1	0			1	1	2	1200	
0	0			0				0	0	1 0	1150 1100 1050	Level 0 *Off task or missing*

Figure 3. Score report for three units and associated activities measuring the *evidence and tradeoffs* variable. Adapted from "From Principles to Practice: An Embedded Assessment System," by M. Wilson and K. Sloane, 2000, *Applied Measurement in Education, 13*, p. 198. Copyright 2000 by Lawrence Erlbaum Associates, Inc. Reprinted with permission.

By using the activities as both a measure of student achievement and source of evidence to provide feedback, EAs may be considered an assessment format that satisfies both summative and formative objectives. EAs differ from ECgD, another assessment format that has strengths suitable for measuring science laboratory skills, as ECgD is considered a formative assessment method (Mislevy, et. al, 2014). However, it is important to bear in mind that any assessment tool can serve formative or summative purposes depending on how it is used.

Evidence-centered game design (ECgD). ECgD is the process of developing a digital game that can also function as a learning and assessment tool to measure skill-based competencies (e.g., collaboration, problem-solving, and communication; Mislevy, et. al, 2014). One aim of ECgD is to synthesize two design frameworks – game and assessment development – shown in Figure 4 into one unified process. On the right side of Figure 4, is the evidence centered design (ECD) framework often used to develop assessments based on evidentiary reasoning to make judgements on students' level of knowledge and skills (Mislevy, Almond, & Lukas, 2003). The ECD framework guides educators to articulate the inferences they wish to make about students and to decide on the evidence needed to support those inferences (Behrens, Mislevy, DiCerbo, & Levy, 2010). The five layers shown on the right side of Figure 4 represent

the different types of analyses and decisions made during the development and operation of an assessment system (see Mislevy et al., 2003 for more details regarding ECD). The left side of Figure 4 shows the design process typically used to guide the development of recreational video games. This game development process emphasizes quick implementation, testing, and enhancement of the product during what is called "the sprint" period. The majority of the enhancements of the game are done between and after the alpha- and beta-user test phases. These test phases are trial runs of the game administered to a pilot group of users so that their feedback maybe collected and used to enhance the game. By testing the product frequently, feedback from game testers is obtained to inform and outline usability, requirements, and constraints (Mislevy, et. al, 2014).

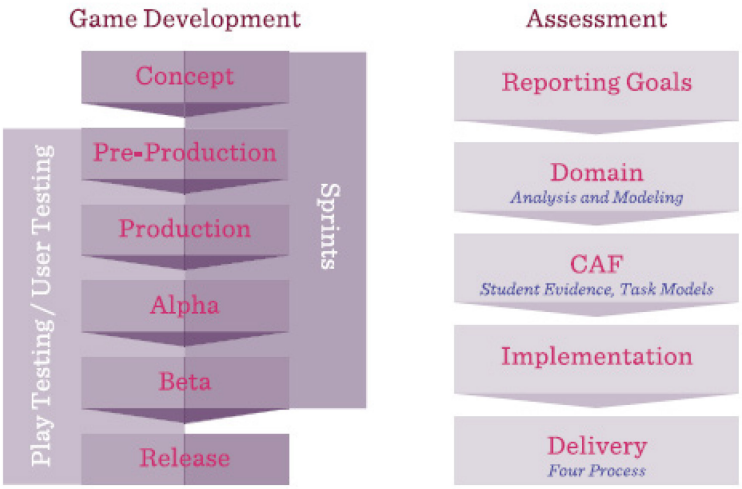


Figure 4. Design frameworks for games and assessments that are integrated using ECgD. Adapted from "Psychometric Considerations In Game-Cased Assessment," by R. J. Mislevy, A. Oranje, M. I. Bauer, A. von Davier, J. Hao, S. Corrigan, … M. John, 2014, *White Paper*, p. 135. Copyright 2014 by GlassLab. Reprinted with permission.

By unifying both of these frameworks, ECgD attempts to meaningfully integrate games and assessment, as shown in Figure 5. For example, Figure 5 illustrates the importance of developing an assessment product that has a meaningful context for students to learn specific knowledge and for educators to measure certain constructs. Once this *meaning* or macro-level defining stage is complete, micro-level designs follow to address the types of actions students need to perform during an activity to indicate whether they have provided sufficient evidence of mastering a construct. Considering the constellation of perspectives outlined in Figure 5 – meaning, construct, knowledge, actions, evidence, and activities – it is important to develop a product that adequately represents aspects of games, learning, and assessments that can collectively in a single tool evoke evidence of players' capabilities (Mislevy, et. al, 2014).

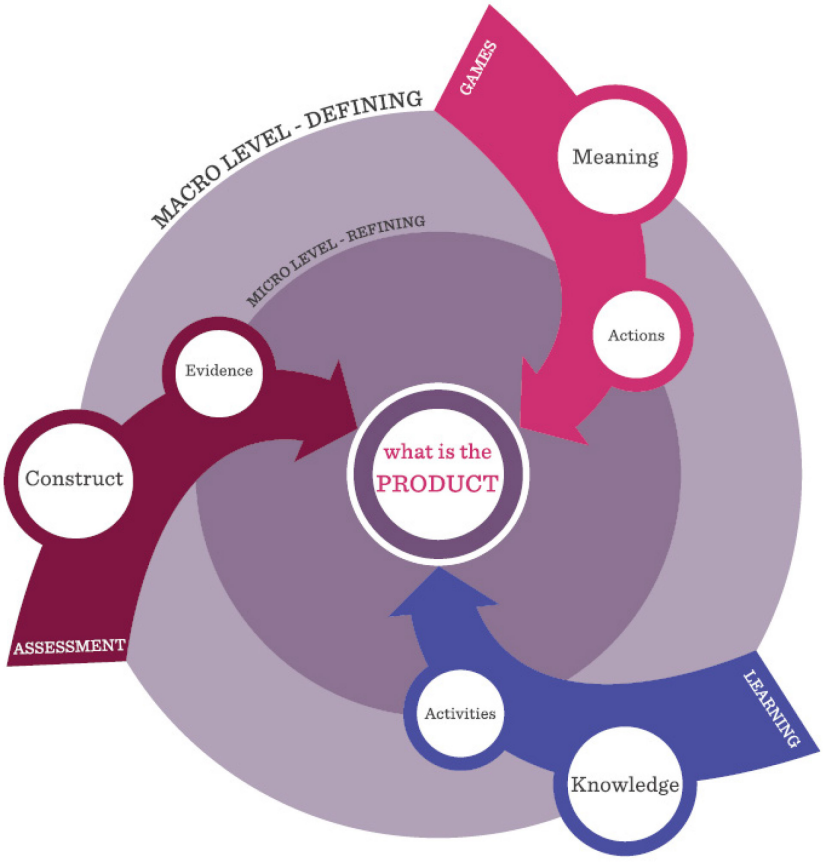


Figure 5. Model of unifying frameworks from the disciplines of games, assessment, and learning. Adapted from "Psychometric Considerations In Game-Cased Assessment," by R. J. Mislevy, A. Oranje, M. I. Bauer, A. von Davier, J. Hao, S. Corrigan, … M. John, 2014, *White Paper*, p. 136. Copyright 2014 by GlassLab. Reprinted with permission.

The integration of games and assessment leads to an ECgD framework that follows four phases (Mislevy, et. al, 2014, pg. 136):

1. Definition of competencies from a non-game realm.
2. A strategy for integrating externally-defined competency with gameplay competency.
3. A system for creating formative feedback that is integral with the game experience.
4. A method for iteration of the game design for fun, engagement, and deep learning, simultaneous with iteration of the assessment model for meaning and accuracy.

It is important to note that ECgD does not follow a retrospective development process, that is, it does not retrofit an existing game to an assessment to collect evidence of knowledge and skill competency. Examples of this type of retrofitting by researchers include the use of popular commercial video games such as *Portal 2* and *Lumosity* (designed for entertainment) to measure problem solving, spatial skills, and persistence (Shute, Ventura, & Ke, 2015). Retrofitting is problematic in ECgD because the types of observable evidence needed to make an inference about a specific skill may not have been designed in the original game. Therefore, making conclusions regarding students' skill levels based on the data collected from these games will invariably lead to weak and possibly inaccurate inferences. Instead, ECgD overcomes these issues by designing the game's mechanics to suit the assessment and learning needs of interest.

As such, it is important to consider the goals of games, assessments, and learning early during the initial planning stages; aspects that are reminiscent of EA.

However, ECgD builds upon the principles of EAs by situating assessment tasks within a digital game environment. ECgD assessments do not necessarily have to be seamlessly embedded into the learning environment; although this level of assessment and learning integration is becoming increasingly common (see Stealth Assessment; Shute & Ventura, 2013). Despite some ECgD assessments having explicit learning and assessment phases, the digital game environment tends to be highly immersive and engaging, thus helping to reduce test or evaluation anxiety (Shute, 2011; Shute, Hansen, & Almond, 2008). Part of this engagement is due to the real-time interactions between the user and the computer game, which is often viewed as feedback. This real-time feedback is made possible by using computers as a method for administering ECgD assessments. Although many, if not most, of the ECgD assessments are administered using computers (Rowe, Asbell-Clarke, & Baker, 2015; Rupp, Gushta, Mislevy & Shaffer, 2010), the framework itself does not mandate the use of digital technology.

Strength of evidence-centered game design (ECgD). The test blueprint that guides the development of an ECgD is called a competency model (Mislevy et al., 2014). These models are rooted in the principles of ECD and are developed from extensive literature reviews of specific constructs (Mislevy et al., 2003). Competency models are similar to the construct maps that are used to develop EAs. However, instead of reflecting an ordering of qualitatively different levels of performance, the competency models used in ECgD often show the links between latent and observable variables (Mislevy et al., 2003; Shute, 2011). Latent variables (e.g., creativity) are not directly observed, but are rather inferred from observable variables; while observable

variables are demonstrable knowledge and skills (e.g., number of agents used in a problem) within a content domain.

An ECgD-based assessment called *Physics Playground* (formerly known as Newton's Playground; Empirical Games, 2013) was developed to measure Newton's three laws of motion, along with creativity and persistence. An example of a partial competency model used in Physics Playground is shown in Figure 6 (Shute & Ventura, 2013). This partial competency model shows the levels of latent variables associated with creativity on the left side; these latent variables are linked to the observable variables on right side of the model. Although not shown in Figure 6, the latent and observable variables are linked to each other with conditional probabilities. These conditional probabilities are used to inform inferences about students' standing on the latent variables of interest based on their observed performance (Shute, 2011). For example, the researchers who developed this competency model indicate the *number of agents used in a problem* is an observable variable that represents evidence of students' familiarity, or *fluency*, with the game. The latent variable *fluency* is one of the three intermediate variables linked to the *cognitive skills* variable, which in turn is linked to the primary latent variable *creativity* (Shute & Ventura, 2013).

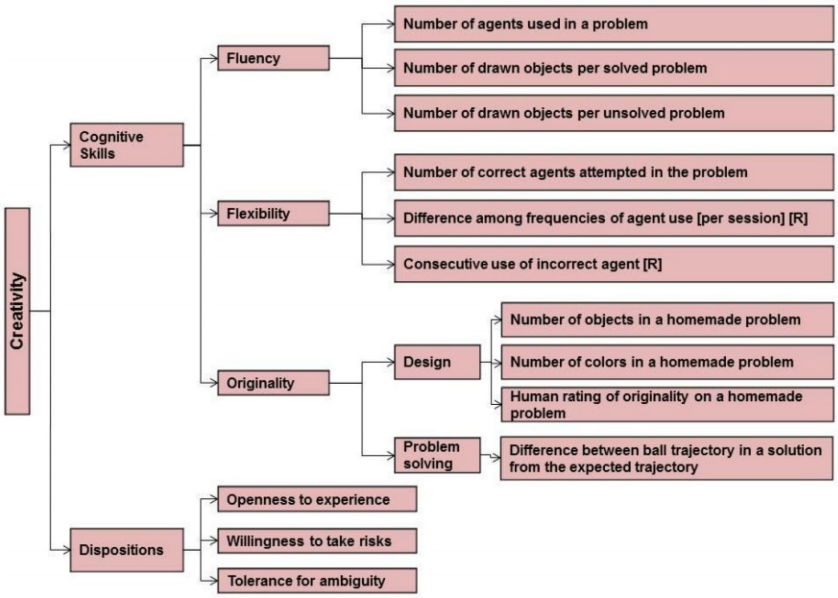


Figure 6. Competency model of creativity in Physics Playground. Adapted from "Stealth Assessment: Measuring and Supporting Learning in Video Games" by V. Shute and M. Ventura, 2013, p. 50. Copyright 2013 Massachusetts Institute of Technology. Reprinted with permission.

Both EA and ECgD assessments hold great potential for measuring performance-based skills in an unobtrusive format. Combining the strengths of these assessments could result in an assessment format that is stronger than each one individually. For example, the use of a curricular-informed construct map (i.e., those used in EA) as an underlying model to develop assessments would ensure the measures are relevant to the classroom, as well as produce results that are an accurate representation of students' knowledge and skills as outlined in the program of studies. Additionally, using open-ended tasks that are embedded seamlessly into a digital learning environment (e.g., those developed using ECgD) could allow for increased student engagement during the learning without being hindered by an explicit assessment phase (Wilson & Sloane, 2000).

Using a combination of EA and ECgD assessments could also enhance efforts to reliably and validly measure science laboratory performances in a digital learning environment. This would allow the capture of students' problem-solving processes in addition to their final responses. Capturing the process of problem solving is superior to many traditional assessment formats (e.g., multiple choice items) that tend to only capture the final response. Capturing the problem-solving process is as important as capturing the product in science education because a significant part of scientific inquiry is premised on the reasoning behind a scientific discovery. Process data may be more easily collected using a computer program than paper-based, performance tasks, as students can interact with the activities and their responses are recorded instantaneously. These digital environments, which can include simulations, may allow educators to replicate how science is done in the real world within a classroom setting. The idea of measuring response processes in the context of mimicking real-world science practice is alluring, but not all digital environments are capable of measuring process as educators would expect. The next section explores one environment, computer simulations, which provide a possible platform to capture process data and replicate real-world science in assessments. Computer simulations utilize the ideas of EA to capture student performance and the framework of ECgD to develop simulation-based assessment.

Computer Simulations as an Educational Tool: Rationale and Proposal for Research

Computer simulations are programs that run on a single computer, or on a network of computers, to mimic an abstract model of a particular system (Strogatz, 2007). Computer simulations have been used for educational purposes for many years. For example, Link Flight Simulators used in 1934 facilitated flight training for US army officers after several deaths (Rosen, 2008). Considering computers have become faster at processing information, the

complexity and realism of the simulations have also increased. Another example is NASA's extensive use of computer simulations to aid in the retrieval and repair of shuttles. In this respect, the military has been a major force in promoting and advancing the technology needed to build computer simulations; for example, accounting for 80% of all modeling and computer simulations (Baker, Niemi, & Chung, 2008). The health sciences have also promoted the use of computer simulations in the form of simulated scenarios to aid with training (e.g., response training in an emergency room; see Rosen, 2008). Although military and medical computer simulation designers have been responsible for a great deal of this innovation, the digital gaming industry has in fact been a recent leader in furthering this advancement (Mizuko, 2009; Squire & Patterson, 2010). These advancements have led to an increased utilization of simulations in primary and secondary education contexts (Shute, 2013).

As digital simulations become more sophisticated, educators recognize the potential to capitalize on these innovations by adding instructive components, such as learning and assessment features, to enhance student understanding. Simulations have been shown to improve learning in science by facilitating knowledge integration and deepening understanding of complex topics, such as genetics and physics (Quellmalz, Timms, & Buckley, 2009). These interactive digital simulations have also proven beneficial in terms of assessment because they provide a method for collecting more and distinct types of data (e.g., process data) than could be done with a paper-and-pencil test (Institute of Education Sciences, 2006). Another benefit of incorporating simulations within classrooms is the ability to replicate complex, dynamic environments, such as how science occurs in the real world; especially in a laboratory setting where students need to interact with real data and conduct experiments using equipment (e.g.,

gas chromatography-mass spectrometry). Digital simulations broaden accessibility to tools and experiences that are difficult to create in real life, inside a classroom.

Computer Simulated Science Laboratories (CSSL) As a Learning Tool

Computer Simulated Science Laboratories or CSSLs are computer programs that can be designed to model hands-on and minds-on science laboratories and scientific processes. Without access to digital simulations in the classroom, it would be difficult in some cases to mimic real-life scenarios and contexts for student learning. For example, one of the requirements in Alberta's Grade 11 physics curriculum is for students to investigate the motion of various objects, such as cars and planes (Alberta Education, 2014a). Digital simulations offer a way for students to further their understanding of the motion of cars and planes without having to leave the classroom. Another example is the Physics Education Technology (PhET) simulation created by a non-profit organization to provide opportunities for research-based science and mathematics interactive experiences. PhET allows students to investigate and experiment with digital replicas of real objects (Baker, Niemi, & Chung, 2008; PhET, 2015). One benefit of using digitally-simulated objects is that they allow students to place greater focus on higher-level thinking processes instead of spending too much time on the technical skills normally required in a laboratory (Sahin, 2006). In fact, some researchers argue that computer simulations are superior to real-world experiments because they allow students to interact with all aspects of experiments as opposed to working with partial data or selected steps of an experiment (Sahin, 2006; PhET, 2015). For example, gas expands under the application of heat as evidenced by increased pressure inside a container; however, in a simulation, students may view the invisible and underlying causes of these effects (e.g., particles vibrate faster as the heat causes them to collide into each other more violently, resulting in increased pressure). In other words, the simulation is

able to offer an environment in which students see and interact with the experiment at a fine-grained level that is not readily transparent in a traditional laboratory (PhET, 2015).

Simulations used in the classroom have also been shown to improve students' knowledge integration skills, which in turn facilitate a deeper understanding of complex topics (Gobert, O'Dwyer, Horwitz, Buckley, Levy, & Wilensky, 2011). For example, in a study focused on human systems simulations, Ioannidou, Repenning, Webb, Keyser, Luhn, and Daetwyler (2010) found significant improvements in student learning about each system's facts, connections to adjacent systems, and ability to apply knowledge about the relationships between each system in different situations. By interacting with individual human systems, students were able to advance their understanding of how each system worked in conjunction with other systems and to understand the intricacies of the entire human body (Ioannidou et al., 2010). In addition to developing deeper understanding, simulations have also been found to influence the manner in which students solve problems (Quellmalz et al., 2009). For example, Stieff and Wilensky (2003) found that students who used *NetLogo* to learn about chemical equilibrium tended to apply higher-order thinking skills (e.g., using a variety of conceptual strategies instead of relying on algorithmic approaches or rote facts when solving problems) than students who used traditional science tools. Thus, when simulations are used appropriately under the right conditions, for example, as supplements to other classroom activities, some research studies have reported overall gains in student achievement, based on pre- and post- test scores (PhET, 2015; Scalise, Timms, Moorjani, Clark, Holtermann, & Irvin 2011; Quellmalz et al., 2009). While simulations may be used as a learning resource, they may also be used as an assessment tool.

Computer-simulated science laboratories (CSSL) as assessment tools. Digital test environments that mirror the real world are important to consider, as students will likely

encounter non-paper-based assessments beyond the classroom. Although students continue to be assessed in classrooms with traditional test formats (e.g., paper-and-pencil), most students experience evaluative situations and feedback in different forms in their everyday lives. For example, the most common forms of assessment that they may encounter outside of school are informal performance-based formats where actually solving a problem is the criterion by which performance is deemed successful (Shute & Becker, 2010). Tools designed to measure this kind of performance and learning beyond the classroom may help teachers prepare students for future endeavors. In addition to designing assessments that mirror real-world activities, it is also important to ensure they are reliable and valid.

Tasked with the problem of narrowing the gap between classroom assessments and real-world problems, performance-based tasks are viewed as promising because their goals and features – at least in principle – appear to align well with problems encountered in everyday life (Shute, 2011). As previously mentioned, performance-based assessments "require students to create an answer or product that demonstrates their knowledge and skills" in a real-life task (U. S. Congress, Office of Technology Assessment, 1992, pg. 5); teachers can then evaluate the process and/or product of the response. CSSLs, like other digitally integrated assessments, are performance-based; they are able to capture both the process and product of students' responses by tracking their actions while they work dynamically through the task-based simulations. The features of CSSLs may facilitate their utility in measuring science skills, such as scientific inquiry in a laboratory environment (Domin, 1999).

CSSLs are also appealing because they may allow for large-scale assessment. As has already been mentioned, science laboratory skills such as scientific inquiry are traditionally assessed by teachers through classroom observations, checklists, and written reports, which can

pose problems for reliability, validity, and integration in large-scale testing (Hofstein & Lunetta, 2004). The reliability of these traditional assessments tends to be lower for a variety of reasons, including because teachers may assign different grades to the same laboratory performance (reflecting the subjective nature of scoring) given that only a portion of the students' activities were actually observed. However, large-scale testing of laboratory skills may be possible with CSSLs as they can standardize the administration of tasks, activities, and resources for laboratories, including the scoring of tasks. The administration of traditional science laboratories appears to be too cumbersome and idiosyncratic to permit this level of access, efficiency and standardization. Objectively measuring and scoring students' performance on CSSLs using the same score metric would be expected to result in more accurate scores and comparisons than has traditionally been found with teacher observations, checklists, and reports. Considering the potential of standardizing CSSLs as large-scale measures of science laboratory skills, there is a need for further investigation to determine whether CSSLs are appropriate measures of scientific laboratory skills. A prime example is the NAEP CSSL known as the *Problem Solving in Technology-Rich Environments* (TRE) assessment described in the following section.

An Example of a Computer Simulated Science Laboratory (CSSL): Problem Solving in a Technology-Rich Environment (TRE)

The National Assessment of Educational Progress (NAEP) created the *problem solving in TRE project* to explore the use of new technologies for improving exam administrations and designing enhanced item formats (Bennett, Persky, Weiss, & Jenkins, 2007). For this project, a CSSL labelled TRE Simulation (TRESim) was created to explore the innovative use of computers for developing, administering, scoring, and analyzing science performance results. TRESim requires students to perform a series of experiments to answer three complex problems:

(a) how do different payload masses affect the altitude of a helium balloon? (b) how do different amounts of helium affect the balloon's altitude? and (c) how do the amount of helium and payload mass together affect the altitude of a helium balloon? (Bennett et al., 2007).

NAEP's Technology-rich environment simulation (TRESim). TRESim is a good example of a CSSL because it provides students with opportunities to experience a *hands-on* and *minds-on* simulated laboratory environment that is designed to mimic how real-world scientists run experiments. TRESim is also a good example of EA and ECgD because its pedagogical objectives and design were considered alongside tasks and activities designed to assess scientific knowledge and skills; therefore, it is not surprising to find many components of EA and ECgD within the TRESim. It provides students with a simulated science laboratory environment in which they can showcase their knowledge and skills. It offers students a series of instructions and a guided practice experiment, before presenting questions associated with three increasingly complex problems.

Upon opening the TRESim program, students are presented with a series of screens introducing the goals of the simulation – to conduct a series of experiments to respond to the three problems regarding the relationship between mass, altitude, and volume in a helium balloon example. Next, there is a practice experiment with information about where all the necessary resources (e.g., problem to be solved, button to run experiment, and glossary help) are located; students are guided by prompts to explore and click each of the resource buttons. As students click through each of the buttons, they are provided with an explanation of what each button does and the resource it provides. A screenshot of the practice experiment, shown in Figure 7, reveals a problem to be solved – located in the top right corner of the screen – and the resource buttons – located along the top and bottom right of the screen. Within the practice

experiment, students encounter the practice question (which also happens to be the first question of the TRESim): "*How do different payload masses affect the altitude of a helium balloon?*" As they try to solve this problem, the simulation prompts students to manipulate different payload masses attached to the balloon to see the various altitudes.

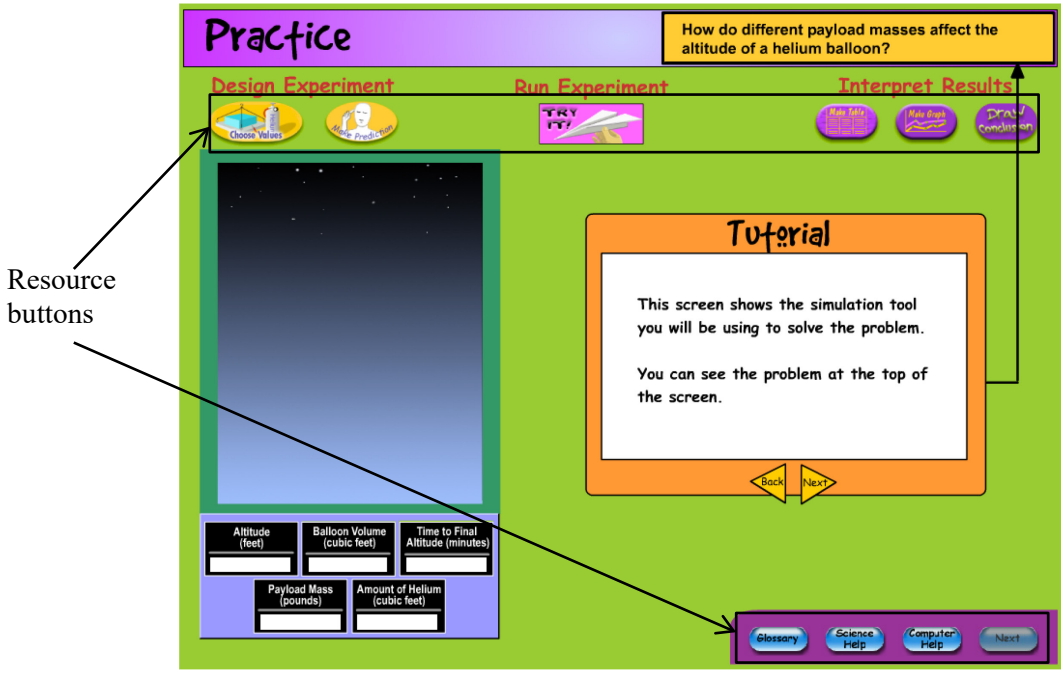


Figure 7. Screenshot of TRE Sim practice experiment. Adapted from "Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project," by R. E. Bennett, H. Persky, A. R. Weiss, and F. Jenkins, 2007, *U.S. Department of Education (NCES 2007–466)*, p. 13. Copyright 2007 by the National Center for Education Statistics. Reprinted with permission.

After completion of the practice experiment, the simulation continues. Students now have three complex problems to solve: (a) how do different payload masses affect the altitude of a helium balloon (same as practice problem)? (b) how do different amounts of helium affect the balloon's altitude? and (c) how do the amount of helium and payload mass together affect the

altitude of a helium balloon? To answer each problem, students are able to manipulate independent variables, make predictions, and run various experiments. After a series of completed trials, students have an option to make a table and/or graph to help them interpret the data and draw conclusions about the relationships between the given variables. Each problem ends with multiple-choice and short-response questions that allow students to indicate and display their findings and generate conclusions. These questions exemplify efforts to embed assessments, which allow educators to measure students' achievement, during a learning activity.

The TRESim possesses many characteristics of both EA and ECgD. For example, the basis of the test blueprint underlying the TRESim is a competency model, similar to those that guide the development of an ECgD. Competency models (e.g., see Figure 6) include well-researched links between latent and observable variables (Mislevy et al., 2003; Shute, 2011). The competency model that guided the development of TRESim is shown in Figure 8 (Bennett et al., 2007).

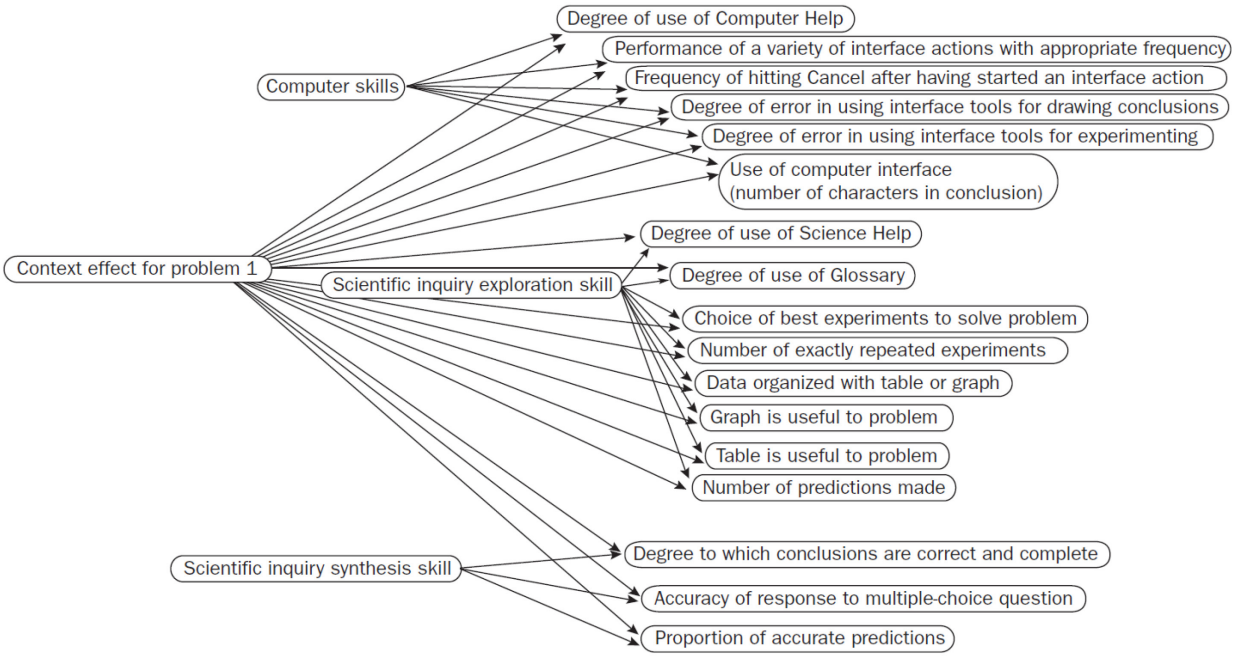


Figure 8. Observable and latent variables of Problem 1 in TRESim. Adapted from "Problem Solving in Technology-Rich Environments: A Report From the NAEP Technology-Based Assessment Project," by R. E. Bennett, H. Persky, A. R. Weiss, and F. Jenkins, 2007, *U.S. Department of Education (NCES 2007–466)*, p. 33. Copyright 2007 by the National Center for Education Statistics. Reprinted with permission.

In Figure 8, the 17 variables on the right represent the observable variables, while the three on the left (i.e., computer skills, scientific inquiry exploration skill, and scientific inquiry synthesis skill) represent the latent variables. For example, the researchers who developed this competency model indicate the 15th observable variable *degree of which conclusions are correct and complete* is evidence of the latent variable *scientific inquiry synthesis skill* (Bennett et al., 2007). Although this model does not possess the curricular alignment of EA construct maps, the knowledge and skills measured by the items can be linked to learner outcomes in the Alberta program of study (see Method section). The TRESim includes a variety of task formats, which is also characteristic of both EA and ECgD assessments. For example, during the experimental phase of the TRESim, seamlessly embedded tasks – those that are characteristic of ECgD assessments – are administered to students (e.g., organizing data into a table and/or graph). These data may be assessed in an unobtrusive way because students are allowed to create tables and/or graphs at any moment during problem solving to help inform their future trials, which also provides an opportunity to integrate assessment tasks within the simulation. Although the TRESim does not provide students with explicit feedback on these tables and/or graphs, students may self-assess the completeness of them by judging whether or not more trials are required to understand the relationship between the variables. After these tasks are completed, students are prompted to answer multiple-choice and open-ended questions designed to measure the

knowledge they have generated, in the form of conclusions, about relationships between the variables. Explicit assessments such as these are characteristic of EA activities.

In addition to its design as an assessment of scientific skills (Bennett et al., 2007), NAEP's TRESim may also have the potential to be a teaching and learning tool. The select feedback (e.g., steps of an experiment) students receive during and after portions of the TRESim may help improve understanding and future scientific inquiry tasks. Because this simulation stands as a significant departure from traditional classroom laboratories, certain enhancements during its administration may help maximize its potential as an educational tool. The next section presents two hypothesized interventions designed to improve the administration of the TRESim: a pre-laboratory activity and a learning error intervention.

Objective of Present Study: Enhancements to CSSLs

The objective of the present study is to investigate two interventions that are hypothesized to enhance the use of CSSLs, namely, the NAEP TRESim. TRESim is an ideal CSSL because it provides a platform to test these interventions. As already mentioned, the purpose of CSSLs is to teach and assess higher-level scientific thinking skills that require the application of scientific knowledge. However, students need a basic level of understanding before they can apply their content knowledge to solve higher-level science tasks. Many traditional science laboratories have a *pre-laboratory activity* designed to prepare students by having them review the requisite knowledge needed. By approaching the laboratory with adequate background knowledge, students are able to focus on higher-level skills, such as knowledge application and scientific inquiry. Many of the studies that investigate simulated laboratories, however, do not administer these pre-laboratories (Gobert et al., 2007; Sahin, 2006; PhET, 2015; Quellmalz et al., 2009). These studies may not have included the use of pre-

laboratory activities because the simulations were a supplement to in-class activities. However, because CSSLs aim to mimic real-life laboratory environments, investigating the efficacy of pre-laboratory activities in enhancing students' performance during the laboratory may enhance their ecological validity (Hofstein & Lunetta, 2004; Ma & Nickerson, 2006).

In addition to the pre-laboratory activity, another real-world aspect of scientific thinking and practice is learning from training errors (Firestein, 2016). CSSLs may provide an excellent test ground for investigating how students learn from the errors they make during science laboratories. Advances in simulation technologies have greatly improved the real-world feel of CSSLs; however, they present a relatively large departure from the traditional hands-on laboratories students have come to expect. One difference is that simulated environments often allow, and encourage, students to solve problems by using different methods and repeating experimental trials. Traditional laboratories do not encourage such deviations in practice, as they tend to follow a linear sequence of scientific inquiry. Because CSSLs allow, and encourage, students to solve problems by using different methods of scientific inquiry, including repeating experimental trials, students are provided an opportunity to use their errors from previous training attempts as learning opportunities to inform later trials. The idea of viewing errors as a positive and natural aspect of learning, especially during training phases, may be new to students who feel embarrassed and ashamed when they make them. Despite the body of research behind the pedagogical value of errors during the learning process (e.g., Firestein, 2016; Ohlsson, 1996), there is a lack of research of how students can be encouraged to view their errors as a learning tool to improve their understanding (Leighton, Chu, Seitz, 2013). The next section presents the two hypothesized interventions in detail.

Pre-laboratory activity. The NAEP TRESim presents a guided problem-based laboratory environment in which students apply their knowledge and skills to solve problems that do not have clearly defined solutions at the outset (Domin, 1999). The ability to solve problems that do not have clearly defined solutions at the outset is a higher-level skill and often requires students to apply their background knowledge strategically. To help students focus on the appropriate content knowledge to use in completing the laboratory activity, a preparatory activity is often recommended (Wilkenson & Ward, 1997). The reasons for this are manifold. First, by drawing upon their background knowledge, students can focus and think critically about their actions as opposed to mindlessly following algorithmic instructions (Hodson, 2003). Second, by orienting students to the basic principles required for sound research and problem solving, educators can in effect provide students with advance organizers to increase their learning and retention of the laboratory material and experience (Cheronis, 1962, pg. 105; Domin, 1999). For example, Cheronis (1962) argues that not preparing students with the requisite background knowledge will hinder their ability to apply this knowledge to new problems, learn from the process, and complete an activity in a meaningful way. Including pre-laboratory activities may be especially relevant for simulated environments, where students may become easily distracted and confused by the graphics and digital interactions of the activity. In order for CSSLs to be usefully implemented as a learning and assessment environment for laboratory skills, it is important that they be treated similarly to traditional hands-on laboratory activities, which are often preceded with a pre-laboratory. Thus, the utility of a preparatory laboratory activity that precedes the simulation is worthy of investigation to ensure that students are ready for the CSSL. These pre-laboratory activities also follow the Principles of Fair

Assessment Practices (1993) by helping students prepare for the assessment they will encounter in the CSSL.

Learning error intervention. Learning errors often occur during the formative (training) stages of acquiring new knowledge and skills. Errors are important sources of information to consider in learning complex material, as they indicate misunderstandings. Thus, using errors constructively could advance student learning during laboratory experiences. One way to enhance the use of CSSLs is to hold a class discussion regarding the importance of exploring learning errors during the simulation. According to the Learning Errors and Formative Feedback (LEAFF) model (Leighton, Chu & Seitz, 2013), encouraging discussion about the value of learning from mistakes is hypothesized to enhance performance, especially within domains such as mathematics and science, where students have been taught that a single correct answer exists but are often afraid to generate incorrect responses. The LEAFF model outlines that a learning environment deemed emotionally safe by students allows them to feel at ease revealing their misconceptions and learning from their errors during the formative phases of learning. When students feel at ease revealing what they do not understand and thus share their misunderstandings, instructors can help correct these misconceptions by providing relevant formative feedback that is specifically targeted to the errors revealed, and in turn students are expected to be more receptive to this feedback than they would otherwise (please see Leighton et al.'s 2013 chapter for details). The LEAFF model, shown in Figure 9, involves three parts.

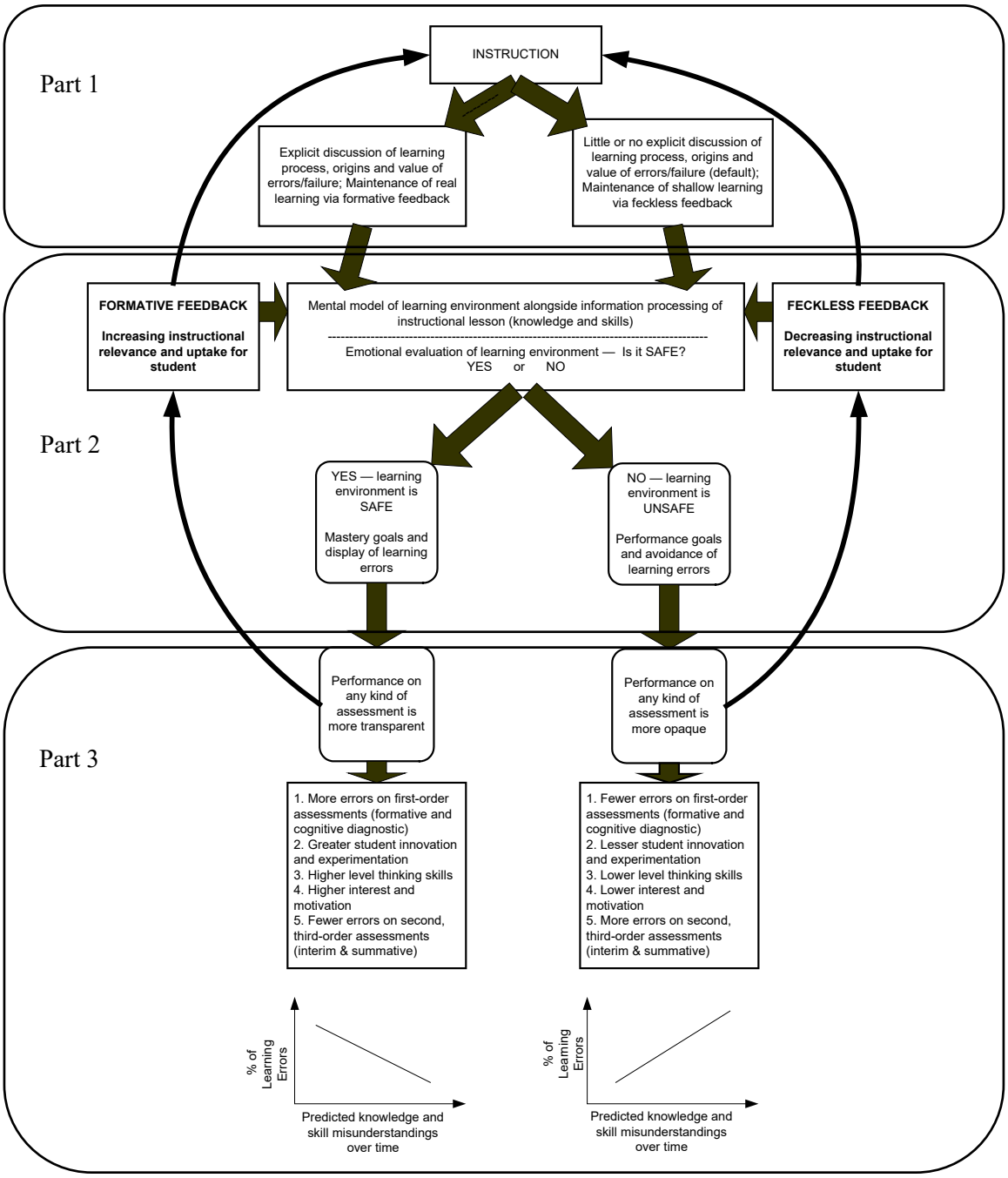


Figure 9. The Learning Errors and Formative Feedback (LEAFF) model. Adapted from "Errors in Student Learning and Assessment: The Learning Errors and Formative Feedback (LEAFF) Model" by J. P. Leighton, M-W. Chu, and P. Seitz, 2013, *In R. Lissitz (Ed.), Informing the Practice of Teaching Using Formative and Interim Assessment: A Systems Approach*, p. 197.

The first part of the LEAFF model focuses on the instructional climate within classrooms, where instructors engage in pedagogical behaviors (e.g., nodding with approval when students make a mistake during the training of a new skill) that either explicitly or implicitly promote safety or risk for learners experimenting with new knowledge and skills. To promote safety, educators are encouraged to explicitly discuss and convey to students that errors are a natural and necessary part of learning complex material (Firestein, 2016). Students who are receptive to these ideas are expected to view their classrooms as emotionally safe (i.e., viewing errors as a learning tool) and, consequently, show and discuss their training mistakes in order to overcome misconceptions and experience greater learning.

The second part of the LEAFF model focuses on students' mental models of the classroom environment. Mental models (Johnson-Laird, 1983) are internal representations that reflect an individual's perception and understanding of the world around them for the purposes of reasoning and problem solving. Student who view the classroom or learning environment as emotionally safe are likely to possess mental models that provides an internal sense of ease, allowing them to demonstrate misunderstandings on formative assessments and interpret formative feedback as more relevant and useful in guiding their learning.

The final part of the LEAFF model focuses on student performance. Students who feel at ease within their learning environment are likely to make more errors during the training phase of learning because they feel safe taking intellectual risks, and gaining a deeper understanding of the content. However, as a result of this early intellectual risk-taking and opportunity for formative feedback, students who feel safe in their learning environments are likely to make fewer errors during post-training on summative assessments. Over time, students who feel safe in revealing what they do not understand are expected to exhibit enhanced socio-emotional

experiences and improved academic performance. The student socio-emotional experiences investigated in this study will include the variables such as school engagement, motivational goals, and test anxiety.

Leighton et al. (2013) outlined the importance of creating a safe psycho-social environment for learning and assessment. One way begins with a discussion of errors. By having the teacher discuss the formative value and necessity of errors during the training phase of learning, he or she in effect attempts to remove the stigma associated with making errors for students, and helps promote safety in having students reveal misconceptions in their newfound knowledge and skills. The LEAFF model, and specifically a discussion of learning errors, aligns well with the objectives of CSSLs as these digital science environments provide opportunities for experimentation of scientific inquiry skills and, by extension, learning through errors.

It is hypothesized that these two interventions – a pre-laboratory activity and a learning error class discussion – should improve students' performance on the TRESim in ways that extend what is already afforded by the digital learning environment resembling a real-world science laboratory. This hypothesis needs support with empirical evidence. This study focuses on investigating whether these two interventions has an effect not only on student performance on the NAEP TRESim but also students' socio-emotional experiences related to the TRESim.

Method

The objective of this dissertation research was to investigate the use of computer simulated science laboratories (CSSLs) as an assessment tool to measure students' science knowledge and skills given weaknesses in traditional science laboratory assessments, namely, their limited scope and static nature. The CSSL that was used in this research is called the *National Assessment of Educational Progress (NAEP) Problem-Solving in a Technology Rich Environment Science Laboratory* (from hereon NAEP TRESim or simply TRESim; see NAEP, 2007; Bennett, Persky, Weiss, & Jenkins, 2007). To investigate this NAEP TRESim as an assessment tool, three research questions, previously presented and listed below, guided the study.

(a) What are the effects of a pre-laboratory activity on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(b) What are the effects of a LEI on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(c) What are the interactions between the pre-laboratory activity and LEI on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

A research design was developed to answer these specific questions. Students' socio-emotional experiences will be operationalized using the variables school engagement, motivational goals, and student anxiety.

A brief description of the design is provided in the next section with details to follow in subsequent sections.

Overview of Research Design

Before describing the research design developed to answer the three research questions outlined, the process for obtaining ethics approval from the University of Alberta and the appropriate school boards is presented. This process is presented at the outset as it is an important requirement to ensure ethical considerations are followed at all stages of the research.

Ethics. Normal protocols, as specified by the University of Alberta Research Ethics Board 2, were undertaken. The ethics proposal, which outlined the research process, was reviewed and approved by the University of Alberta. In addition to seeking ethics approval from the University, ethical protocols were also undertaken for each of the school boards identified by the author as potential candidates for inclusion in the research. The author chose school boards in two large urban school districts in Alberta given their proximity for data collection. By the time data collection was scheduled to start, only one school board had approved this study. Hence only principals, teachers, and students from schools within that district were invited to participate. The ethical protocols of particular interest to both the University of Alberta and school district ethics boards were the following: (1) informed consent letters for both students and their parents or guardians to indicate voluntary involvement in the research, (2) use of computer-generated codes and student-made identification codes to protect the privacy of participants, and (3) use of pseudonyms in the reporting of findings to protect participants' identities.

Design of a quasi-experimental study with two interventions. A 2 (Pre-Lab Activity versus no Pre-Lab Activity) × 2 (LEI versus no LEI) quasi-experimental design was used to

evaluate Grade 8 students' socio-emotional experience, and science knowledge and skills based on their NAEP TRESim performance. Given the design, the students took part in one of four treatments, which reflected a combination of two independent variable manipulations, namely, interventions of a pre-laboratory activity and LEI as shown in Table 1. It is important to note that students were not randomly assigned to treatments, given that random assignment would have disrupted normal classroom activity. Instead, participating schools were assigned to one of the four treatments.

Table 1

Design of Quasi-Experimental Study with Two Interventions

Interventions	Pre-Lab Activity	No Pre-Lab Activity
LEI	School A	School C
No LEI	School B	School D

Students in all four treatments, generated by crossing the two interventions shown in Table 1, were administered three tasks - a pre-intervention survey measure, a post-intervention survey measure, and the NAEP TRESim assessment. Although the pre- and post-intervention survey measures probed background characteristics and student socio-emotional experiences given the NAEP TREsim, performance on the NAEP TRESim was the main outcome variable in the present study. The pre- and post-intervention survey measures included sub-scales specifically chosen to measure socio-emotional variables associated with different parts of the LEAFF model and other variables relevant to the study. These variables will be elaborated upon later in this chapter. The completion of these three tasks and the administration of the applicable treatments were estimated to take 130 minutes. However, the length of each science class period in the participating schools was only 60 minutes. The teachers and author decided to administer the

materials over two days to minimize the disturbance to students' schedules in courses that preceded or followed the science classes. On Day 1, the science class was shortened to 35 minutes and on Day 2, the class was prolonged to 95 minutes. Agreements were made with the teachers of students whose classes preceded or followed the science classes to accommodate this schedule. The schedule for administering the materials in each school is shown in Table 2.

Table 2

Schedule of Materials Administered to Students in Each School

School	Day 1	Day 2
A	Pre-intervention survey measure	LEI
	Pre-laboratory activity	TRESim
		Post-intervention survey measure
B	Pre-intervention survey measure	No LEI
	Pre-laboratory activity	TRESim
		Post-intervention survey measure
C	Pre-intervention survey measure	LEI
	No pre-laboratory activity	TRESim
		Post-intervention survey measure
D	Pre-intervention survey measure	No LEI
	No pre-laboratory activity	TRESim
		Post-intervention survey measure

Detailed information regarding the two interventions and administration schedule will be presented later in this chapter. The next section describes the characteristics of the participants.

Participants

Students and teachers of Grade 8 science classes were invited to participate in this study. The sample of students and teachers who participated in this study is considered a sample of convenience because all of the participants came from one school district. Only one school district (or board) superintendent approved the study in time for data collection. After approval of the present study, the district's science consultant recommended the schools best suited for participation in the study. The district's science consultant recommended approaching large junior high schools with similar levels of academic achievement, as measured by provincial standardized tests known as the provincial achievement test (PAT). Large junior high schools were able to accommodate the large sample of students planned for the study while similar academic achievement would ensure a large pool of comparable participants, at least on the variable of achievement. For example, if previous academic achievement was not considered or controlled during the sampling, then differential performance between schools on the outcome variables of interest (i.e., TRESim) could be attributed to previous academic achievement and not to the intervention implemented in the study. Controlling other demographic variables such as socioeconomic status (SES), which previous research suggests accounts for academic performance differences (Lytton & Pyryt, 1998), was also considered in the present study.

Once the schools were identified by the consultant, the schools' principals, science teachers, and science students were invited to participate in the present study. The four schools that participated in the study were considered larger high junior high schools within the district, but they were not well matched as expected in terms of their PAT achievement scores and/or their SES. The SES information was based on median household income of the community/area in which the school was located (City of Calgary, 2016). For example, schools A and D were

located in areas with higher SES (i.e., average median household income = $82, 308) relative to school B and C, which were located in areas with lower SES (i.e., average median household income = $59, 837). The differential profiles of participating schools are elaborated later in the chapter in light of efforts to control their potential effects on the results.

In total, 298 students from 14 classes, as well as their 10 teachers participated in the study. The 10 teachers included seven males and three females, but no additional demographic information was collected from them. One hundred forty-one students self-identified as male (47.3%), 121 as female (40.6%), and 36 did not disclose gender (12.1%). The students represented more than 11 ethnicities with a majority of them indicating they were Caucasian (33.9%). The students' ages ranged from 12 to 15 years with 99.2% of them indicating they were between 13 and 14 years of age at the time of data collection. The students also self-disclosed that 262 (87.9%) of them had access to a computer at home. A majority of the students (61.7%) learned to use the computer on their own. The demographic composition of students in each of the four treatments is presented in Table 3.

Informed consent. Since the participants of this study included Grade 8 students, who are typically 13 years of age and below the age of consent, parental or guardian consent was also needed before students could participate in the study. One week before the study was administered, parental/guardian consent forms were sent home with students so that approval could be received in time for students to participate in the study. A copy of the information letters and consent forms for parents, teachers, and students are shown in Appendix A, B, and C respectively. Students who did not return their consent forms in time for the study obtained verbal consent from their parents/guardians. To obtain verbal consent, the students' science teachers called their homes to explain the study to their parents/guardians and asked them for

permission to have students participate in the study. Verbal consent is a relatively common approach used by the participating teachers to obtain permission for their students to participate in different events throughout the school year. In fact, the method of obtaining verbal consent was recommended by the participating teachers to the author. The script used by the teachers to obtain verbal consent is presented in Appendix D.

Table 3

Demographic Composition of Students in the Four Schools

	School A		School B		School C		School D	
	Number of Students	Percent of Students (%)	Number of Students	Percent of Students (%)	Number of Students	Percent of Students (%)	Number of Students	Percent of Students (%)
Number of students	108	36.2	73	24.5	69	23.2	48	16.1
*Gender								
Male	62	57.4	25	34.2	29	42.0	25	52.1
Female	41	40.0	27	37.0	32	46.4	21	43.8
Ethnicity								
Caucasian	53	49.1	9	12.3	13	18.8	26	54.2
African American	5	4.6	10	13.7	7	10.1	5	10.4
Filipino	23	21.3	19	26.0	21	30.4	2	4.2
Latin American	6	5.6	7	9.6	10	14.5	5	10.4
Other	16	14.8	8	11.0	11	15.9	9	18.8
Age								
13	42	38.9	25	34.2	31	44.9	18	37.5
14	60	55.6	26	35.6	30	43.5	26	54.2
Have access to computer at home	102	94.4	61	83.6	54	78.3	45	93.8
Learned majority of their computer skills by themselves	78	72.2	34	46.6	45	65.2	27	56.3

Rationale for recruiting Grade 8 science students. Grade 8 science was chosen for the present study because the NAEP TRESim was originally designed to be cognitively appropriate for students at this level, who are typically 13 years of age (NAEP, 2007; see also Bennett et al., 2007). Although the content knowledge used in the TRESim was not explicitly listed in the Alberta program of study (i.e., the provincial document that regulates the learning outcomes taught in Alberta classrooms), the contents of the TRESim were nonetheless highly relevant to the Alberta science program. For example, the TRESim focuses on the science behind the buoyancy of a helium balloon, and utilizes the ideas of the particle model of matter, which is covered in the Grade 8 science unit on *Mix and Flow of Matter* (Alberta Education, 2014b). The NAEP TRESim also includes aerodynamics, which is taught in the Grade 6 science units *Air and Aerodynamics* and *Flight* (Alberta Education, 1996). The laboratory skills required in the TRESim, such as performing and recording as well as analyzing and interpreting skills, correspond to one of the four foundational pillars of the junior high (i.e., grades 7-9) science Alberta program of study (Alberta Education, 2014b). Thus, administering the TRESim to Grade 8 students provided them with a review of Grade 6 science content while also measuring their application of Grade 8 science content knowledge and skills.

Procedure: Interventions (Pre-Laboratory Activity and Learning Errors Intervention [LEI])

The two interventions (independent variables) – pre-laboratory activity and LEI – were manipulated simultaneously to create four treatments in the 2×2 quasi-experimental design introduced earlier. As mentioned previously, the present study followed a quasi-experimental design because students were not randomly assigned to one of the four treatments created from the combination of two levels of each of the independent variables. Random assignment of

students to treatments was not possible because it was considered problematic to implement within the existing classrooms; for example, students within the classroom would have been expected to discuss the presence or absence of the pre-laboratory activity and LEI with their peers and thus potentially bias the results. Consequently, a decision was made to minimize the potential for students to discuss the treatments with each other by assigning all students from each of the junior high schools to only one of the four treatments. Although this process of assignment effectively confounds students at a given school with a specific treatment, preliminary information, such as demographic and prior-knowledge data, was collected from the students using the pre-intervention survey measure to account and control for pre-existing differences among students participating in the four treatments. More information regarding the prior-knowledge measure questions are discussed later. The interventions – pre-laboratory activity and LEI – are described next.

Intervention 1: Pre-laboratory activity. This intervention was completed by students in schools A and B only. The pre-laboratory activity was administered after students completed the consent form and pre-intervention survey measure. The activity took approximately 15 minutes to administer on Day 1 of data collection. The students assigned to treatments that did not involve a pre-laboratory activity were given 15 minutes at the end of class time to work on homework instead of completing the activity. The pre-laboratory activity was developed by the author, and is shown in Appendix E. The pre-laboratory activity was designed to have students review the first problem of the TRESim and thus give them an opportunity to review the basic concepts related to the scientific processes required to solve the three problems presented in the NAEP TRESim. The use of a pre-laboratory activity before a traditional hands-on laboratory is a relatively common practice, but this idea has not translated over to digital laboratories, such as

CSSLs (Scalise et al., 2011). Hence, in this study, the use of a basic pre-laboratory activity – essentially one that replicated the first problem of the TRESim was used – as an initial attempt to investigate how such an activity might cue students to review and prepare for the knowledge and skills needed for the CSSL.

The pre-laboratory activity comprised the first problem of the TRESim (i.e., How do different payload masses affect the altitude of a helium balloon?). This problem was designed to cue the foundational scientific inquiry skills students needed to solve the three TRESim problems, which involved items that probed identification of key variables; for example, the manipulated, responding, and controlling variables required to answer a research problem. Students who were administered the pre-laboratory activity were thus exposed to the first TRESim problem twice, once during the pre-laboratory activity and once during the TRESim assessment. Students who were not administered the pre-laboratory activity were exposed to this problem only once, during the actual TRESim administration. Although the same problem was used during the pre-laboratory activity and TRESim assessment, a different approach was used each time. The pre-laboratory activity focused on the planning stages (e.g., listing materials needed for and identifying the different variables of the experiment) while the TRESim assessment required students to execute their experimental design (e.g., selecting different masses and running the experiments) to solve the problem.

After the pre-laboratory activity was completed, the author collected and marked all the activities so that they could be returned to students during the next science class, which happened to be the following day. The marked activities included only a numerical score for each of three sections; no additional feedback was provided to control for amount and ensure consistency of feedback given to students. The feedback provided did not include the correct answer for the

items administered in the pre-laboratory activity. Although the feedback provided was consistent among the students who received this intervention (i.e., students in schools A and B), a limitation with the pre-laboratory activity and feedback was that it was not elaborative or personalized feedback. Elaborative and personalized feedback, which highlights each students' learning errors, has been deemed useful in improving students' understanding because it targets their specific areas of weaknesses so that they may focus on improving these areas (Shute, 2008).

On Day 2 of data collection, the author returned the marked pre-laboratory activities to students and explicitly asked them to reflect on the mistakes for learning purposes. The author indicated "please look over the pre-laboratory activity and focus on the mistakes that were made. How would you fix the mistake so that you could get full marks next time?" Students who completed the pre-laboratory activity were given approximately 5 minutes to review their mistakes on Day 2 before beginning the TRESim assessment. Students assigned to treatments that did not involve a pre-laboratory activity were given this review time (5 minutes) to work on their homework after they completed the TRESim assessment and post-intervention survey measure.

Intervention 2: Learning error intervention (LEI). The LEI intervention was only administered to students in schools A and C. The intervention consisted of a script that led students through a brief but targeted discussion of the learning process; specifically highlighting the necessity of making mistakes and learning from mistakes. The discussion was designed to explicitly inform students that mistakes or learning errors are not only an important aspect of the learning or training process but often necessary to encourage exploration of different methods of problem solving (Firestein, 2016). This discussion was guided by a five-slide PowerPoint presentation, shown in Appendix F, which took approximately 5 minutes to administer at the

beginning of the class on Day 2. During the PowerPoint presentation (i.e., at the end of slide 3) students were explicitly encouraged by the author to think of an experience in which they could recognize the value of making mistakes while learning a new skill and how those mistakes helped them to learn. Students who did not receive the LEI were provided with 5 minutes to work on their homework at the end of class time after they completed the TRESim and post-intervention survey measure.

Experimental Material

As mentioned previously, two levels of the interventions were administered to students in a 2×2 design as shown in Table 1. In what follows, the temporal and material details of the study, including the introduction (common to all treatments), interventions and administration of the pre- and post-intervention survey measures and TRESim experienced by all students at each school are described.

Procedure/materials common to all treatments. After consent was obtained from teachers and parents/guardians, as previously mentioned, the author again explained the study to the students on Day 1 so that their consent could be affirmed. This explanation and affirmation of consent took approximately 5 minutes. The author took 5 minutes to verbally explain the study by reading the Student Information Letter and Consent Form to students and emphasized that their participation in the study was voluntary and they could withdraw their data from the study at any time until one month after the data collection was complete. After this verbal explanation, the author administered student consent forms and asked them to sign the forms. In order to protect students' identities, they were asked to make up a *student code* on a separate sheet that would be written on their consent forms and used as their "names" during this study. The author checked the student-generated codes to ensure they were unique. Students were then

asked to write their student codes on the consent forms, which also contained their real names to allow the researcher to remove their data if they wished to withdraw from the study during the specified time. On Day 2, after the appropriate treatments were administered to students at each school, the consent forms were returned to students so they could write down their CSSL codes, which were generated randomly by the TRESim assessment. The two codes written on the consent form, the student-generated and CSSL codes, were used to link students' treatment group with their outcome data (i.e., TRESim, pre- and post-intervention survey measure performance). The following outlines unique procedures followed for each school depending on their specific treatment:

> **School A:** The students from school A (n=108) completed the pre-intervention survey measure (approximately 20 minutes) and pre-laboratory activity (approximately 15 minutes) on Day 1. The items administered as part of the pre-intervention survey measure also included ten *prior-knowledge questions* to account and control for pre-existing differences among students participating in the four treatments. These items are described in the next section and are shown in Appendix G. On Day 2, these students were presented with a 5-minute PowerPoint LEI discussion at the beginning of the class period. Following this intervention, students' marked pre-laboratory activities were returned and they were instructed to review their errors, which took approximately 5 minutes. After students reviewed their errors, the author collected the pre-laboratory activities so that students did not reference them during the TRESim assessment. The collection of pre-laboratory activities ensured all students in the study would have access to the same materials *during* the TRESim assessment (i.e., it prevented students in schools A and B from having an advantage over their peers in schools C and D by having

access to their pre-laboratory activity *during* the TRESim). The author then instructed students to complete the TRESim (approximately 60 minutes) and post-intervention survey measure (approximately 25 minutes). The items administered as part of the post-intervention survey measure included a post-intervention assessment question to evaluate students' abilities to design an experimental method to solve a science problem related to the TRESim. The items from the post-intervention survey measure are shown in appendix H. A detailed description of the questions and survey items administered during the two days of data collection are described in a later section.

School B: The students from school B (n=73) also completed the pre-intervention survey measure (approximately 20 minutes) and pre-laboratory activity (approximately 15 minutes) on Day 1. On Day 2, these students did not receive the 5-minute LEI but did receive their marked pre-laboratory activities and were instructed to review their errors for approximately 5 minutes. After students reviewed their errors, the author collected the pre-laboratory activities, instructed students to complete the TRESim (approximately 60 minutes) and then the post-intervention survey measure (approximately 25 minutes). Since these students were not administered the 5-minute LEI PowerPoint presentation, they were given these extra 5 minutes to work on their homework after they completed the post-intervention survey measure.

School C: The students from school C (n=69) completed the pre-intervention survey measure (approximately 20 minutes) on Day 1. However, the pre-laboratory activity was not administered, and instead they were given the 15 minutes to spend on homework after they completed the pre-intervention survey measure. On Day 2, these students were presented with a 5-minute LEI PowerPoint discussion, the TRESim (approximately 60

minutes), and the post-intervention survey measure (approximately 25 minutes). Since these students did not complete the pre-laboratory activity, they were not provided with 5 minutes to review errors from the activity. Hence, they were instructed to use that time to work on their homework after they completed the post-intervention survey measure.

School D: The students from school D (*n*=48) completed the pre-intervention survey measure (approximately 20 minutes) on Day 1. However, the pre-laboratory activity was not administered, and instead they were given the 15 minutes to work on their homework after they completed the pre-intervention survey measure. On Day 2, these students were not presented with the 5-minute LEI PowerPoint presentation, but were directly presented with the TRESim (approximately 60 minutes) and the post-intervention survey measure (approximately 25 minutes). Since these students did not complete the pre-laboratory activity (i.e., did not require 5 minutes to review errors) or have a 5-minute LEI presentation, they were given these extra 10 minutes to work on their homework after they completed the post-intervention survey measure.

Although students in each of the schools were administered a combination of the two interventions – pre-laboratory activity and LEI – all students were administered the same pre- and post-intervention survey measures and TRESim assessment. These are described in detail in the next section.

Pre-intervention survey measure. A survey booklet, found in Appendix G, was developed by the author of this study to measure key socio-emotional variables, such as achievement orientation, motivation, engagement and also knowledge variables of relevance to the present study. These variables measured were specifically chosen to address specific aspects of the LEAFF model and background information (e.g., use of computer technology during

science class and students' prior science knowledge). These data were collected to control for pre-exiting differences that could be expected to influence how students engaged with and reacted to the interventions. It is important to note that multiple scales were used to capture different aspects of these socio-emotional constructs. In addition, the background information collected provided the author with a baseline measure of students' use of technology in the science classroom and prior-science knowledge. These baseline measures were considered necessary to help understand potential differences in students' TRESim assessment performance. A summary of the sub-scales administered during the pre-intervention survey measure and the rationale for inclusion are provided in Table 4.

Table 4

Summary of Subscales Included and Reasons for their Inclusion in the Pre-Intervention Survey Measure

Subscales	Source of Original Instrument	Number of Items in Survey	Reasons for Using These Items
Mastery goal orientation	Patterns of adaptive learning scales (Midgley et al., 2000)	Survey 1: Items #1-5	Baseline measure of students' goal orientation for learning to inform part two of the LEAFF model
Performance-approach goal orientation	Patterns of adaptive learning scales (Midgley et al., 2000)	Survey 1: Items #6-10	Baseline measure of students' goal orientation for learning to inform part two of the LEAFF model
Performance-avoid goal orientation	Patterns of adaptive learning scales (Midgley et al., 2000)	Survey 1: Items #11-14	Baseline measure of students' goal orientation for learning to inform part two of the LEAFF model
Intrinsic goal orientation	Motivated strategies for learning questionnaire (Pintrich, Smith, Garcia, & McKeachie, 1991)	Survey 2: Item #1-4	Baseline measure of students' goal orientation for learning to inform part two of the LEAFF model
Extrinsic goal orientation	Motivated strategies for learning questionnaire (Pintrich et al., 1991)	Survey 2: Item #5-8	Baseline measure of students' goal orientation

			for learning to inform part two of the LEAFF model
Learning strategies: critical thinking	Motivated strategies for learning questionnaire (Pintrich et al., 1991)	Survey 2: Item #9-13	Baseline measure of students' learning strategies to inform part two of the LEAFF model
Frequency of scientific methods used in class	TRESim background questionnaire (Bennett et al., 2007)	Survey 3: Item #1-4	Baseline measure of students' frequency of performing different activities in the science classroom
Frequency of computer use in science class	TRESim background questionnaire (Bennett et al., 2007)	Survey 3: Item #5-9	Baseline measure of students' use of computers to complete different activities in the science classroom
Prior science knowledge	Prior-knowledge questions (Bennett et al., 2007)	Prior-knowledge questions: Item #1-10	Baseline measure of students' prior science knowledge

The booklet consisted of three surveys compiled with items from three pre-existing scales with adequate reliability and validity (see Bennett et al., 2007; Midgley et al., 2000; Pintrich, Smith, Garcia, & McKeachie, 1991), and designed to measure goal orientations, motivational learning strategies, and use of scientific process. Where appropriate, internal consistency or alpha values (α) for the scale items are presented in the Results section. Details of the surveys are described in the following subsections. In addition to the three surveys, 10 prior-knowledge questions, developed by NAEP, were included in the pre-intervention survey measure.

Survey 1: Subscales from the Patterns of adaptive learning scale (PALS; Midgley et al., 2000). Items designed to measure students' mastery, performance-approach, and performance-avoid goal orientations were administered. As part of the PALS (Midgley et al., 2000), these items are designed to measure the relationship between students' learning environment and motivation, affective disposition, and behaviour, which are

variables related to the LEAFF model and therefore relevant to the study. Students responded to 14 of the 94 original PALS items using a 5-point Likert-type scale ranging from 1 – *Not true at all*, 3 – *Somewhat true*, to 5 – *Very true*. Items #1-5 measured *mastery goal orientation* and targeted the extent to which students are focused on developing their competency and extending their mastery and understanding in achievement settings (e.g., It's important to me that I learn a lot of new concepts this year). Items #6-10 measured *performance-approach goal orientation* and targeted the extent to which students are focused on demonstrating their competence (e.g., It's important to me that other students in my class think I am good at my class work). Items #11-14 measured *performance-avoid goal orientation* and targeted the extent to which students are focused on avoiding the demonstration of incompetence (e.g., It's important to me that I don't look stupid in class). The instruction, when completing this scale, was for students to think about their classes in general because the items generally reflected long-term affective dispositions.

Survey 2: Motivated strategies for learning questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1991). Items designed to measure students' motivational orientations, but also including use of different learning strategies, were administered. The items were taken from the MSLQ (Pintrich et al., 1991; see Duncan & McKeachie, 2005, for actual items). Students responded to only 13 of the full set of 81 MSLQ items using a 7-point Likert-type scale ranging from 1 – *Not at all true of me* to 7 – *Very true of me*. The 13 items were chosen specifically to measure students' motivation and learning strategies related to the LEAFF model. Items #1-4 measured *intrinsic goal orientation* and targeted students' perceptions of the reasons they engage in learning tasks, including

reasons for challenge, curiosity, and mastery (e.g., I prefer class material that really challenges me so I can learn new things). Items #5-8 measured *extrinsic goal orientation* and targeted students' perception of their engagement in learning tasks based on external factors such as grades, rewards, performance, evaluation by others, and competition (e.g., Getting a good grade in the class is the most satisfying thing for me right now). Items #9-13 measured *learning strategies*, for example, the critical thinking associated with applying previous knowledge to new situations or making critical evaluations of ideas when learning (e.g., I often find myself questioning things I hear or read in class to decide if I find them convincing). The instructions, when completing this survey, was for students to think about their classes in general because the items generally reflected long-term affective dispositions.

Survey 3: NAEP TRESim background questionnaire (Bennett et al., 2007). Items designed to measure students' frequency of implementing different aspects of the scientific method using computers in the classroom were administered. Specifically, 9 items were compiled from the NAEP TRESim background questionnaire (Bennett et al., 2007). Students responded to the items using a 3-point Likert-type scale ranging from 1 – *Never*, 2 – *Sometimes, but less than once a month*, to 3 – *Once a month or more*. This 3-point Likert-type scale was slightly modified from the original 4-point Likert-type scale, which included a fourth response option of *Not taking science*. This response option was removed because all students in our sample were currently enrolled in a science class. Items #1-4 queried students about different aspects of science activities in the classroom; in particular, about how often they performed certain activities such as *design your own science experiment or investigation.* Items #5-9 queried students about their use of a

computer to perform certain activities such as c*ollect data using lab equipment that interfaces with computers (for example, probes)*. The instruction, when completing this scale, was for students to think about their science classes in particular because the items reflected activities performed during science class.

NAEP prior-knowledge questions. In addition to the three surveys, 10 questions designed to measure students' existing science knowledge and skills were administered; these questions were developed by NAEP testing specialists (Bennett et al., 2007). The responses to these questions were used as covariates in data analysis to help control for students' prior knowledge and skills across the different treatments.

Pre-laboratory activity. The pre-laboratory activity (see Appendix E) was developed by the author, based on the NAEP TRESim, to introduce and probe students' review of concepts related to scientific inquiry in a laboratory situation. The activity required students to apply their science knowledge and skills to a novel situation. This activity was completed by students in schools A and B only. The activity presented the first task of the TRESim and included three items. The first item required students to develop a hypothesis. The second item required students to explain the scientific method they intended to use when solving the given problem. This second item was designed to scaffold students' responses as it was divided into four sub-items, requiring students to list the materials needed for an experiment, develop the steps required for the experiment, indicate the recording/measurement tools that would be used, and describe a method to organize the data collected. The third item required students to list the manipulated, response, and control variables. The overarching problem presented in this activity, was to find the relationship between the amount of mass hanging from a balloon and its altitude.

NAEP Technology-rich environment simulation (TRESim). This computer simulated assessment was developed by NAEP testing specialists to measure Grade 8 students' scientific problem solving knowledge and skills in a technology-rich environment (NAEP, 2007). This simulation was first administered to 1,946 Grade 8 American students in 2003 (NAEP, 2007), and consists of three problems:

1. How do different payload masses affect the altitude of a helium balloon?
2. How do different amounts of helium affect the balloon's altitude? and
3. How do the amount of helium and payload mass together affect the altitude of a helium balloon?

This simulation was designed to measure students' skill as related to the following constructs: (a) scientific exploration, (b) scientific synthesis, and (c) computer skills. Each area can be viewed as comprising a latent variable linked to a series of observable variables. The TRESim competency model for Problem (task) 1, which was described in the previous chapter, is shown in Figure 8. The model shown in Figure 8 comprises 17 observable variables for problem 1 that were used to measure students' skill in scientific exploration, scientific synthesis, and computer skills. These 17 observable variables were used to measure performance during Problems 2 and 3 as well. In addition to the 51 observables variables across problems 1 through 3 (17 observable variables × 3 problems), 1 observable variable (i.e., a series of concluding multiple-choice items administered after Problem 3) was incorporated to measure students' overall knowledge and skills regarding the content of the three TRESim problems (NAEP, 2007). The TRESim was designed to measure all 52 of these observable variables.

Analyses of the 2003 data, by the NAEP team, indicated that many of the observable variables contributed very little to explaining students' TRESim performance and they were

removed (Bennett et al., 2007). Hence, students' total score on the 2003 TRESim administration consisted only of 28 observable variables associated with the three problems. These observable variables were divided into three scores: a scientific exploration score, consisting of 11 variables (original α=0.78; see Bennett et al., 2007); a scientific synthesis score, consisting of 8 variables (original α=0.73; Bennett et al., 2007); and a computer skills score, consisting of 9 variables (original α=0.74; Bennett et al., 2007). Overall, these observable variables had an overall alpha value of 0.89 (see Bennett et al., 2007).

For this present study, 38 observable variables were collected and used in the data analyses. First, four observable variables were added (*number, range, and distribution of experiment* for Problem 1; *number, range, and distribution of experiment* for Problem 2; as well as *controlling variables* and *number, range, and distribution of experiment* for Problem 3) that were not part of the original 52 designed for data capture in the TRESim log files. Thus, these four variables were targeted specifically for this study. The author of this study considered these four variables necessary to include in order to obtain a fuller picture of students' performance on the TRESim assessment as these variables measure different aspects of developing experimental designs, which is part of scientific inquiry (Hofstein & Lunetta, 2003). Although these four variables were not part of the original 52 designed for data capture in the TRESim log files, the author developed a parsing program to pull these data to track students' performance. The data on these four variables were expected to enhance understanding of students' development and application of experimental designs; for example, the 'controlling variable' captured students' data on whether they kept one of the task variables (e.g., mass) constant while manipulating the other task variables (e.g., volume) for three or more trials. Student data on the 'number, range, and distribution of experiments observable variables' were captured by measuring whether the

trials students ran reflected a good representation of the manipulated variable (e.g., in Problem 1, did students manipulate the mass to run only 10 lb and 20 lb OR did they manipulate the mass to run 10 lb, 30 lb, 60 lb, and 90 lb?). These variables were designed to measure whether students manipulated variables in such a way that a good spread of instances was tested.

The remaining 34 variables measured in this present study came from the 52 observable variables originally designed for the TRESim. Fourteen of the 34 observable variables included a focus on science exploration (i.e., degree of science help [for Problems 1, 2, and 3], degree of use of glossary [for Problems 2, and 3], data organized with table or graph [for Problems 1, 2, and 3], number of predictions made [for Problems 1 and 2]), scientific synthesis (i.e., proportion of accurate predictions [for Problems 1 and 3]), and computer skills (i.e., degree of computer help [Problems 2 and 3]). The remaining 20 observed variables were chosen from the refined analyses presented in Bennett et al. (2007). These 20 observable variables (i.e., 11 from science exploration + 7 from scientific synthesis + 1 from computer skills + 1 from conclusion) were used to place an emphasis on science exploration and scientific synthesis, and to a lesser extent on computer skill. To focus on the area of science exploration, there were 11 observable variables (i.e., degree of use of glossary [Problem 1], choice of best experiment to solve problems [Problems 1, 2, and 3], graph is useful to [solving] problem [Problems 1, 2, and 3], table is useful to [solving] problem [Problems 1, 2, and 3], and number of predictions made [Problem 3]). The focus on the area of scientific synthesis used seven observable variables (i.e., degree to which conclusions are correct and complete [Problems 1, 2, and 3], accuracy of response to multiple-choice question [Problems 1, 2, and 3], and proportion of accurate predictions [Problem 2]). In addition to the emphasized areas, the present study also examined

the area of computer skills with one observable variable (i.e., degree of computer help [Problem 1]), and concluded with a multiple-choice item comprising one observable variable.

These 38 observed variables were split so that 12 were measured in each of Problems 1 and 2 during the TRESim, 13 measured in Problem 3, and one measured during the conclusion section (i.e., a series of multiple choice questions). Table 5 shows the list of observable variables that was used in the original TRESim and the present study. Detailed explanations of how the 34 observable variables used in the present study (taken from the original 52) were scored can be found in the original TRESim (NAEP, 2007; see also Bennett et al., 2007). Appendix I shows a list of the observable variables, their operationalization, and scoring during the simulation.

Table 5

Types of Observable Variables Measured in the Original TRESim and the TRESim used in this Present Study for all Three Problems

Types of variables listed in competency model	Problem 1/ ORIG	Problem 1/PS	Problem 2/ORIG	Problem 2/PS	Problem 3/ORIG	Problem 3/PS
Science Exploration						
1. Degree of science help		X		X		X
2. Degree of use of glossary	X	X		X		X
3. Choice of best experiment to solve problems	X	X	X	X	X	X
4. Number of exactly repeated experiments						
5. Data organized with table or graph		X		X		X
6. Graph is useful to problem	X	X	X	X	X	X
7. Table is useful to problem	X	X	X	X	X	X
8. Number of predictions made		X		X	X	X
9. Controlling variables[+]						X
10. Number, range, and distribution of experiments[+]		X		X		X
Scientific Synthesis						
11. Degree to which conclusions are correct and complete	X	X	X	X	X	X
12. Accuracy of response to multiple-choice question	X	X	X	X	X	X
13. Proportion of accurate predictions		X	X	X		X
Computer Skills						
14. Degree of computer help	X	X		X		X
15. Performance of a variety of interface actions with appropriate frequency						
16. Frequency of hitting Cancel after having started an interface action						
17. Degree of error in using interface tools for drawing conclusions	X		X		X	
18. Degree of error in using interface tools for experimenting	X					
19. Use of computer interface (number of characters in conclusion)	X		X		X	

*Three multiple-choice items administered during the conclusion portion of TRESim were measured by both the original TRESim and the TRESim used in this study.

[+]These variables were not part of the original TRESim competency model, but were included in the TRESim used in this study to better understand students' experimental designs.

Post-intervention survey measure. This measure, shown in Appendix H, was designed to be administered after the TRESim assessment. The measure included a single-item summative-type assessment question, developed by the author of this study, designed to evaluate students' abilities to choose the best experiment to solve a problem. Based on the TRESim design, this item focused on investigating the buoyancy of a helium balloon. Specifically, students who responded to this item were expected to solve a problem that involved four variables (i.e., How do the amount of helium, payload mass, and temperature together affect the altitude of a helium balloon?). Through this paper-and-pencil item, students had the opportunity to showcase their skills by developing the best scientific method to solve the problem. This item was used as a dependent variable representing student science knowledge and skills, and it was used as a comparison to their TRESim performance. Following this item, students were asked to complete a survey designed to measure engagement with the TRESim, test anxiety, use of computer technology, and other demographic variables associated with this study. Some of these variables (e.g., engagement) are related to the LEAFF model and focus on students'socio-emotional experiences as per the three research questions guiding this study. Survey items were compiled from pre-existing instruments found in the research literature, all of which had acceptable levels of reliability and validity as illustrated in the Results section. A summary of the sub-scales administered during the post-intervention survey measure and the rationale for inclusion are provided in Table 6.

Table 6

Summary of Subscales Included and Rationale for their Inclusion in the Post-Intervention Survey Measure

Subscales	Source of Original Instrument	Number of Items in Survey	Reasons for Using These Items
Emotional engagement with TRESim	School engagement scale (Fredericks, Blumenfeld, Friedel, & Paris, 2005)	Survey 1: Items #1-6	Measure of students' emotional engagement with the TRESim assessment that may influence their mental model of the learning environment.
Cognitive engagement with TRESim	School engagement scale (Fredericks et al., 2005)	Survey 1: Items #7-11	Measure of students' cognitive engagement with the TRESim assessment that may influence their mental model of the learning environment.
General test anxiety	Motivated strategies for learning questionnaire (Pintrich et al., 1991)	Survey 2: Item #1-5	Measure of students' test anxiety related to the mental models they develop of their learning environments, including traditional assessments.
Specific TRESim anxiety	Motivated strategies for learning questionnaire (Pintrich et al., 1991)	Survey 3: Item #1-5	Measure of students' anxiety with TRESim related to the mental models they develop of their learning environments, including the simulated laboratory assessment.
Range of activities performed using computers	NAEP TRESim background questionnaire (Bennett at al., 2007)	Survey 4: Item #1-7	Baseline measure of students' use of computers to complete different activities in the science classroom. This provides a partial measure of students' proficiency with using computers.
Frequency of general computer use	NAEP TRESim background questionnaire (Bennett at al., 2007)	Survey 4: Item #8-10	Baseline measure of frequency of student use of computers. This also provides a partial measure of students' proficiency with using computers.
Motivation to use computers	NAEP TRESim background questionnaire (Bennett at al., 2007)	Survey 4: Item #11-13	Measure of students' motivation to learn and complete schoolwork when using a computer.

The surveys are described as follows:

Survey 1: School engagement scale – Behavioral, emotional, and cognitive engagement (Fredericks, Blumenfeld, Friedel, & Paris, 2005). This survey was administered

following completion of the TRESim assessment. The survey included adapted items from the School Engagement Scale, which includes Behavioral, Emotional, and Cognitive Engagement subscales (Fredericks et al., 2005). Although the original scale was developed to measure students' classroom and school engagement, for the purposes of this study, only the emotional and cognitive subscales were adapted to measure engagement during the simulated laboratory by changing the words 'classroom' and 'school' to 'simulated science laboratory.' The behavioral subscale was not adapted because it was not considered fully relevant to the study, as it reflected long-term behavioural dispositions, which was not a focus of the study, and the items were not easily modified for a simulated laboratory environment. Thus, students responded to only 11 of the full set of 15 items using a 5-point Likert-type scale ranging from 1 – *Never*, 2 – *On occasion*, 3 – *Some of the time*, 4 – *Most of the time*, to 5 – *All of the time*. Items #1-6 were adapted from the original emotional engagement subscale to measure students' emotional engagement during a school activity (e.g., I feel happy when using the simulated science lab) and items #7-11 were adapted from the original cognitive engagement subscale to measure students' cognitive engagement in a classroom activity (e.g., When I read the instructions and post-lab conclusions, I ask myself questions to make sure I understand what it is about). The instructions, when completing this scale, were for students to think about their experiences during the CSSL because the items focused on their engagement during the TRESim assessment.

Surveys 2 & 3: Motivated strategies for learning questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1991). These surveys included 10 items adapted from the MSLQ (Pintrich et al., 1991; full scale includes 81 items) and used to measure anxiety in

regular testing situations (survey 2) and also specifically during the CSSL (survey 3). Students responded using a 7-point Likert-type scale, ranging from 1 – *Not at all true of me* to 7 – *Very true of me*. Only items from the MSLQ that focused on test anxiety were included, as this was one of the dispositions of interest in this study. The items in survey 2 measured students' perceived levels of psychological over-arousal, feelings of worry and dread, self-deprecating thoughts, and tension that might occur during regular test situations (e.g., When I take a test I think about how poorly I am doing compared with other students; see also Duncan & McKeachie, 2005; Zeidner, 1998). The instruction, when completing this scale, was for students to think about tests in general because the items reflected their experiences with tests in the past.

Items in survey 3 were identical to survey 2 except they were adapted to specifically measure anxiety during the CSSL (e.g., When I did the simulated science lab I thought about how poorly I was doing compared with other students). The survey items were adapted by changing the word *test* with *simulated science laboratory*. The instruction, when completing this scale, was for students to think about their experiences of the CSSL because the intent was to measure students' levels of anxiety when using a simulation as an assessment, instead of their past experiences with other simulation programs.

Survey 4: NAEP TRESim Background Questions (Bennett et al., 2007). This survey consisted of demographic items from the NAEP TRESim background questionnaire. The survey measured students' background variables, as well as their frequency and reasons for using computers. Items #1-7 were designed to measure the type of activities for which students generally used computers (e.g., play computer games) using a 4-point Likert-

type scale ranging from 1 – *Not at all*, 2 – *Small extent*, 3 – *Moderate extent*, and 4 – *Large extent*. All the items and their Likert-type scales in this survey were retained in their original wording format. Items #8-10 were designed to measure the frequency with which students used computers (e.g., How often do you use a computer at school?) using a 5-point Likert-type scale ranging from 1 – *Never or hardly ever*, 2 – *Once every few weeks*, 3 – *About once a week*, 4 – *Two or three times a week*, to 5 – *Every day*. Items #11-13 were designed to measure students' motivation for using a computer for school work (e.g., I am more motivated to get started doing my schoolwork when I use a computer) using a 5-point Likert-type scale ranging from 1 – *I never use a computer*, 2 – *Strongly disagree,* 3 – *Disagree*, 4 – *Agree*, to 5 – *Strongly agree*. Items #14-15 were designed to measure where students learned their computer skills and whether they had access to a computer at home. Items #16-18 involved questions related to background (demographic) variables such as gender, age, and ethnicity. These items were administered in the post-intervention survey measure to distribute the questions over time so students would not become fatigued with too many survey questions at the beginning of the study. All materials used in the presented study were administered to students in a specific order in order to standardize the procedure.

Data Analysis

The data analyses conducted in this study were aimed at answering the research questions previously mentioned, namely, whether or not the two interventions – pre-laboratory and LEI – had an effect on students' socio-emotional experiences and science knowledge and problem-solving skills. The data analyses involved considering (1) missing data, (2) pre-intervention

covariates, (3) TRESim observable variables, and (4) post-intervention subscales. First, an analysis of the missing data was performed to investigate the best method to handle missing data.

Second, a multivariate analysis of variance (MANOVA) was used to determine whether there were any pre-existing differences among the students in the four schools using items from 13 subscales administered in the pre- and post-intervention surveys (e.g. goal orientations, learning strategies, and prior knowledge). Although it is not usual to include measures for pre-existing differences in a post-survey, the reason two of the subscales for pre-existing differences were administered in the post-intervention survey measure was to balance out the number of items included for data collection. The scales that were added to the post-intervention survey measure were carefully selected to request information about past behaviors that were less likely to be influenced by the interventions. That is, the computer-related subscale administered in the post-intervention survey measure requested students to indicate the *types of activities they completed using a computer* (e.g., to what extent do you write using a word processing program on a computer?) and the *frequency of using a computer* (e.g., how often do you use a computer at school?). Pre-existing differences among the students were investigated, given that they were not randomly assigned to treatment conditions; survey responses that indicated significant differences among students were incorporated into the analyses in the form of covariates.

Third, principal component analysis was performed on the TRESim data as the number of observable variables was too large for a multivariate analysis of covariance (MANCOVA). Thus, a principal component analysis was first performed to reduce the number of observed variables into fewer components. Then, a MANCOVA was used to determine whether differences existed among the four schools on the variables of interest. Following the results of the MANCOVA, a discriminant function analysis was performed to investigate the TRESim variable components

that reflected the most notable differences among treatment conditions. Since discriminant function analysis does not permit for the inclusion of covariates, the covariate information from the MANCOVA was considered in light of the discriminant function results. Fourth, the post-intervention survey measure responses were analysed using a MANCOVA to determine differences among the schools. The results of these analyses are presented in the next chapter.

Results

As outlined in the Methods Section, a quasi-experimental study was conducted in which two interventions, a pre-laboratory activity and learning error intervention (LEI), were manipulated to evaluate their effect on Grade 8 students' socio-emotional experiences, science knowledge, and skills based on their NAEP TRESim performance. Specifically, three research questions were investigated in this study:

(a) What are the effects of a pre-laboratory activity on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(b) What are the effects of a LEI on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(c) What are the interactions between the pre-laboratory activity and LEI on students' socio-emotional experiences, as well as on understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

A quasi-experimental design was used to investigate whether these two interventions – pre-laboratory activity and LEI – had an effect on students' socio-emotional experiences and academic performance as measured by a CSSL assessment called the NAEP TRESim. All students were administered a pre-intervention survey measure, the NAEP TRESim assessment, and a post-intervention survey measure. The NAEP TRESim science laboratory assessment was administered to the following four treatment groups of Grade 8 science students: (1) Pre-laboratory activity and LEI, (2) No Pre-laboratory activity and LEI, (3) Pre-laboratory activity and no LEI, and (4) No pre-laboratory activity and no LEI. This chapter presents the results from

analyses of the pre- and post-intervention survey measures and the NAEP TRESim assessment across the four treatments. All analyses were performed using SPSS Version 21.0.

Data analyses are presented in four sections. First, missing data were analyzed, including possible reasons for missing data; the use of listwise deletion to address missing data throughout the analyses is explained. Second, the pre- and post-intervention survey measures, as well as prior-knowledge questions were analysed for initial group differences so that any identified differences could be used as covariates in the present analyses and/or considered in the interpretation of results. Third, the TRESim assessment data were analysed to evaluate whether the two interventions, and any possible interaction, had any effects. Fourth, the post-intervention survey measure data were analysed for students' socio-emotional experiences with the NAEP TRESim science laboratory, as well as any associations with student performance on the TRESim.

Missing Data

In total, 298 students participated in the study. The percentage of total missing data was 7.75%, which was made up of 5.11% due to absences/incorrect student codes and 2.64% due to random non-responding. Investigation of the reasons for missing data revealed that most of it was a result of students not completing one of the two (i.e., pre- or post-) intervention survey measures, and/or the TRESim. The main reason some students did not complete one of the intervention survey measures and/or the TRESim was because they missed one of the two days of data collection. The pre-intervention survey measure was administered on Day 1 of data collection while the post-intervention measure and TRESim were administered on Day 2. Hence, students who were absent for one of the two days could have missed completing the pre-intervention survey measure, the post-intervention survey measure, and/or the TRESim. In

addition, some students' data indicated they did not complete the TRESim but had completed the post-intervention survey measure, both of which were administered on Day 2 of data collection. This outcome occurred when students failed to write down their TRESim code on the post-intervention survey measure, which prevented the researcher from linking their TRESim action log to their student code and surveys. The missing data that resulted from these absences (i.e., missing the entire pre-intervention survey measure, post-intervention survey measure, and/or TRESim) was 5.11% of the full data set. Table 7 presents a breakdown of the number of students who missed each of the pre-intervention survey measure, the post-intervention survey measure, and/or TRESim.

Table 7

Number of Students with Missing Data from the Pre-Intervention, Post-Intervention, and/or TRESim

	Pre-Intervention	Post-Intervention	TRESim
Number of Students (Percent of Students)	22 (7.38%)	28 (9.40%)	27 (9.06%)

Further analysis revealed that there was some overlap among students in terms of their missing data. For example, of the 22 students who did not complete the pre-intervention survey measure and 27 students who did not complete the TRESim, two of these students (0.67%) did not complete both the pre-intervention survey measure and TRESim. Similarly, of the 22 students who did not complete the pre-intervention survey measure and 28 who did not complete the post-intervention survey measure, one student (0.34%) did not complete either survey. Additionally, of the 27 students who did not complete the TRESim and 28 students who did not complete the post-intervention survey measure, 17 students (5.70%) did not complete both the

TRESim and the post-intervention survey measure. These 17 students most likely missed the second day of data collection during which both the TRESim and post-intervention survey measure were administered. Only 19 students (6.38%) missed only the pre-intervention survey measure, 8 students (2.68%) missed only the TRESim, and 10 students (3.36%) missed only the post-intervention survey measure. Two approaches were considered, replacement of missing data with an imputed value and listwise deletion, to address the problem of missing data in subsequent analyses. Replacement of the missing data with an imputed value was determined to be an inadequate solution because of the large proportion (5.11%) of missing data associated with, or originating from, specific students (Gravetter & Wallnau, 2009, Pigott, 2001).

Listwise deletion was considered as the defensible method to handle the missing data because the total percent of random missing data (e.g., missing one or two items in a survey) was only 2.64% (recall that 7.75% of the total missing data involved 5.11% due to absences/incorrect student codes and 2.64% due to random omissions). More specifically, the percentages of random missing data were 0.30% from the pre-intervention survey measure, 0.00% from the TRESim (the computer-generated logs were not programmed to miss any data points), and 0.63% from the post-intervention survey measure. Considering the low proportion of random missing data (i.e., less than 5%) listwise deletion is normally recommended (Acock, 2005; Gravetter & Wallnau, 2009). Thus, listwise deletion was used to handle all missing data for subsequent univariate and multivariate analyses.

This section focused on identifying the missing data in the study and how the missing data were addressed. The next sections are guided by the three research questions presented in the *Method* section of this dissertation, which required an exploratory analysis of the data to determine whether the two interventions – pre-laboratory activity and LEI – had significant

effects on students' socio-emotional experiences, as well as their science knowledge and skills as measured by the NAEP TRESim. This exploratory analysis was split into three sections so that each portion of the data could be analysed in a logical order. Considering the exploratory nature of the analyses, which resulted in 19 statistical tests conducted, a more conservative alpha of α=0.01, rather than α=0.05, was used to determine whether a statistical test led to a significant result.

Part 1: Before the TRESim Assessment

In order to determine whether there were any pre-existing differences among students in each of the four treatment groups, a MANOVA of the items in the pre-intervention survey measure and select items from the post-intervention survey measure was conducted. Descriptive statistics for each of the subscales are presented in Table 8.

Table 8

Descriptive Statistics of the Possible Covariate Subscales Based on Each Schools' Treatment

Subscale	Mean (Standard Deviation)			
	School A: Pre-lab & LEI (n=93)	School B: Pre-lab (n=50)	School C: LEI (n=61)	School D: Comparison (n=44)
1. Mastery goal orientation	4.08 (0.80)	4.27 (0.59)	4.20 (0.53)	4.17 (0.62)
2. Performance-approach goal orientation	2.65 (0.98)	2.58 (0.85)	2.77 (0.96)	2.91 (1.13)
3. Performance-avoid goal orientation	2.94 (0.96)	2.56 (0.93)	3.09 (1.02)	3.17 (0.97)
4. Intrinsic goal orientation	4.23 (1.23)	4.50 (1.21)	4.48 (1.04)	4.06 (1.02)
5. Extrinsic goal orientation	5.44 (1.37)	5.11 (1.36)	5.81 (1.03)	5.40 (1.31)
6. Learning strategies: critical thinking	4.03 (1.31)	4.65 (1.04)	4.72 (1.14)	4.09 (1.23)
7. Frequency of scientific methods used in class	1.68 (0.54)	1.94 (0.55)	1.64 (0.51)	1.59 (0.58)

8. Frequency of computer use in science class	1.76 (0.47)	1.83 (0.50)	1.88 (0.59)	1.78 (0.54)
9. Range of activities performed using computers	2.66 (0.58)	2.58 (0.63)	2.85 (0.45)	2.71 (0.52)
10. Frequency of general computer use	3.28 (0.79)	3.54 (0.89)	3.52 (0.86)	3.79 (0.83)
11. Prior-knowledge question	5.34 (2.04)	3.96 (1.47)	4.52 (1.86)	5.32 (2.10)

Notes: Boxed subscale items were administered during the post-intervention survey measure; all other subscale items were administered during the pre-intervention survey measure. School sample size is denoted using *n*.

Pillai's trace criterion was used to evaluate group differences because it is considered robust in the presence of unequal sample sizes, and violations of homogeneity of variance-covariance and normality (Tabachnik & Fidell, 2007). Violation of the assumption of homogeneity of variance-covariance was confirmed by a significant Box's M test of equality of covariance matrices, $F(198, 83099)=1.478$, $p<0.01$; Box's M=320.679. However, because Box's M test is highly sensitive to violations of normality (Tabachnik & Fidell, 2007), the normality of the data was also analysed using the Kolmogorov–Smirnov (KS) test. The KS test indicated that only two variables satisfied the normality assumption (i.e., performance-approach goal orientation, KS=0.057, $p>0.01$; and learning strategies: critical thinking, KS=0.052, $p>0.01$) and the remaining variables did not; thus providing an explanation for the results of Box's M test. Nine of the 11 possible covariates were not normally distributed because many students rated the items from these subscales very highly; this was not surprising given the nature of the student sample. A MANOVA of the subscales indicated that there was a statistically significant difference among the four groups in their responses, $F(33, 708)=3.506$, $p<0.01$; Pillai's trace=0.421, partial eta squared=0.140.

Since the multivariate test was significant, univariate tests were performed to investigate which subscales showed significant differences among the four groups. The univariate tests

indicated that five subscales showed significant response differences among the four groups, namely: (1) performance-avoid goal orientation, (2) learning strategies: critical thinking, (3) frequency of scientific methods used in class, (4) frequency of general computer use, and (5) prior-knowledge questions. The univariate tests and the internal consistency for each subscale are shown in Table 9.

Table 9

Results of ANOVA to Assess Pre-Existing Group Differences and Internal Consistency of Each Subscale

Subscale	Type III Sum of Squares	Mean Square	F	*p* value	Partial Eta Squared	Internal Consistency
1. Mastery goal orientation	1.387	0.462	1.027	0.381	0.012	0.848
2. Performance-approach goal orientation	3.298	1.099	1.150	0.330	0.014	0.867
3. Performance-avoid goal orientation	11.202	3.734	3.966	0.009*	0.046	0.733
4. Intrinsic goal orientation	7.064	2.355	1.789	0.150	0.022	0.701
5. Extrinsic goal orientation	14.011	4.670	2.848	0.038	0.034	0.790
6. Learning strategies: critical thinking	25.281	8.427	5.819	0.001*	0.067	0.799
7. Frequency of scientific methods used in class	3.601	1.200	4.062	0.008*	0.048	0.788
8. Frequency of computer use in science class	0.606	0.202	0.751	0.523	0.009	0.756
9. Range of activities performed using computers	2.255	0.752	2.482	0.062	0.523	0.672
10. Frequency of general computer use	8.315	2.772	3.988	0.008*	0.062	0.290

11. Prior-knowledge questions	78.393	26.131	7.215	0.000*	0.081	0.623

Notes. df_{group}=3, df_{error}=244; *denotes statistically significant univariate test result at α=0.01.

The significant group differences shown in Table 9 warranted post-hoc comparisons using the Tukey-Kramer test. This post-hoc test allows for multiple pairwise comparisons of group means without inflating the type 1 error, and is considered appropriate when unequal group sizes are present. The group means for each subscale were previously presented in Table 8. The Tukey-Kramer test revealed no significant differences among the four groups on the sub-scales *performance-avoid goal orientation* and *frequency of scientific methods use*. Although these results might appear inconsistent with the univariate findings, which indicated group differences in these two subscales, it is important to note that the univariate findings consider all possible comparisons of the groups, even, for example, combinations of group comparisons that may not be meaningful (e.g., groups 1 and 2 vs. 3 and 4). However, when the Tukey-Kramer test was used to evaluate study-relevant comparisons, the results indicated none of the four groups, when compared to each other, were significantly different on the two sub-scales.

The Tukey-Kramer test did reveal significant group differences on the other three sub-scales. First, the Tukey-Kramer test indicated that students from School C (i.e., receiving only LEI intervention) rated themselves significantly higher than their peers in School A (i.e., receiving both pre-laboratory and LEI interventions) on the *learning strategies: critical thinking* subscale; no other significant differences were found. Second, the Tukey-Kramer test showed that students from School D (i.e., control, receiving no interventions) rated themselves significantly higher than their peers in School A (i.e., receiving both pre-laboratory and LEI interventions) on the *frequency of general computer use* subscale; no other significant differences were found. Third, the Tukey-Kramer test revealed that students from School A (i.e., receiving both pre-laboratory and LEI interventions) and D (i.e., control, receiving no

interventions) scored significantly higher than their peers in School B (e.g., receiving only pre-laboratory intervention) on the *prior knowledge question* scale; students from School C (i.e., receiving only LEI intervention) were not significantly different from the other three groups, scoring in the middle of the *prior knowledge question* scale.

Although the Tukey-Kramer test revealed specific group differences in three of the sub-scales, the univariate tests did reveal group differences in five of the subscales. Hence, the five subscales (i.e., *performance-avoid goal orientation*, *learning strategies: critical thinking*, *frequency of scientific methods used in class*, *frequency of general computer use*, and *prior-knowledge questions*) were identified as possible covariates when analyzing TRESim performance and the remaining post-intervention survey measure data. These five subscales were identified as possible covariates because another criterion of whether a variable should be used as a covariate is its correlation with the dependent variables. Hence, the correlation of each of these five possible covariates and the dependent variable was considered prior to its use in the analyses.

The next part of the analyses was designed to determine whether the two interventions – pre-laboratory activity and LEI – had an effect on students' performance on the TRESim. A series of principal component, multivariate, and discriminant function analyses was conducted to investigate the effects of the interventions. The principal component analysis (PCA) was used to reduce the number of observable, dependent variables into fewer components. This was followed by a multivariate analysis of the components to identify whether significant differences existed among the four groups. Following a significant multivariate analysis test, a discriminant function analysis was used to indicate how the groups differed on a composite of the components. The following are the findings from these analyses.

Part 2: TRESim Assessment

Group results on the NAEP TRESim are presented by the three problems included in the assessment as follows:

Problem 1: How do different payload masses affect the altitude of a helium balloon?

TRESim Problem 1 included 12 observable variables, which is considered too many for a two-way multivariate analysis (Tabachnik & Fidell, 2007). Hence, a PCA was first performed to reduce the observable variables into fewer components. Although Problem 1 included 12 observable variables, only ten observable variables were used in the PCA because two of the variables were found to be unsuitable for inclusion since they were not continuous. Specifically, the first variable, *use of table or graph* was a categorical variable that indicated whether students used a table, graph, or both to organize their data. The second variable, *concluding results*, required an open-ended response from students and was therefore not numerically scored. Responses were not scored and used because the TRESim program only captured the first portion of students' open-ended responses. More specifically, the log files of students' responses to the open-ended items only included 266 characters (including spaces). This limitation was not known to either the researcher or students during data collection; hence, many students wrote relatively lengthy responses to the open-ended questions but information past the 266th character was not captured. The remaining ten variables used in the analyses reflected continuous scales. Descriptive statistics for these ten variables are shown in Table 10.

Table 10

Descriptive Statistics of Problem 1 Observable Variables Based on Each School's Treatment

Observable Variable	Mean (Standard Deviation)			
	School A: Pre-lab & LEI (*n*=55)	School B: Pre-lab (*n*=34)	School C: LEI (*n*=33)	School D: Comparison (*n*=36)

1. Frequency of using computer help button	0.13 (0.43)	0.26 (0.75)	0.18 (0.58)	0.11 (0.40)
2. Frequency of using glossary button	0.31 (0.77)	0.71 (1.70)	1.03 (2.10)	0.28 (0.85)
3. Frequency of using science help button	0.18 (0.84)	0.56 (1.05)	0.39 (1.14)	0.19 (0.47)
4. Choice of best experiment to solve problems	2.04 (0.74)	1.85 (0.70)	1.94 (0.79)	2.03 (0.61)
5. Number, range, and distribution of experiments	2.20 (1.25)	2.00 (1.28)	1.94 (1.35)	2.33 (1.15)
6. Graph is useful to solving problem	1.95 (1.19)	1.79 (1.15)	1.82 (1.16)	2.22 (1.05)
7. Table is useful to solving problem	1.47 (1.07)	0.85 (1.11)	1.15 (1.00)	1.39 (1.08)
8. Number of predictions made	4.00 (2.90)	4.35 (3.27)	3.76 (3.00)	3.36 (2.38)
9. Number of correct predictions made	2.56 (2.43)	2.35 (2.39)	2.55 (2.22)	2.03 (1.96)
10. Score of multiple-choice conclusion item	0.62 (0.49)	0.38 (0.49)	0.42 (0.50)	0.47 (0.51)

Note: School sample size is denoted using *n*.

It should be noted that the total sample size used in the analyses of Problem 1 was 158 because the coding of two of the ten variables resulted in many missing values. In particular, two variables (i.e., *graph is useful to solving problem* and *table is useful to solving problem*), which were coded to reflect the quality of students' responses (i.e., responses were ranked using the computer-based rubrics shown in Appendix I), revealed a number of students who did not use either a graph or a table. Although a value of -9 was assigned to students who did not create a graph or table, this value is categorical and does not reflect a meaningful, continuous value. This categorical code could also not be considered a low ranking, because the TRESim did not prompt students to make both a graph and table, so treating the -9 as a low ranking on the assessment would have penalized students who did not make either. Thus, the categorical code of -9 had to

be considered a missing value; this method was considered the most defensible treatment of the -9 value. Consequently, this treatment of the -9 values increased the number of missing values in the dataset. Since, listwise deletion was used for the analyses of the data, the sample size decreased for Problem 1 analyses.

Next, a PCA extraction was completed using the *oblimin with Kaiser normalization* rotation; this analysis indicated a reduction of the 10 variables into three components as shown in Table 11. The use of the *direct oblimin with Kaiser normalization* rotation using a delta of 0 was justified because the correlation among the ten variables ranged from -0.172 to 0.838, as shown in Table 12. This PCA resulted in a simple loading structure in which three components were found. These components were labelled, based on the content of the variables: *prediction and method*; *use of help button*; and *data organization and conclusion*. Only one PCA was conducted for the whole sample (n=158) because conducting separate PCAs for each of the four schools would have resulted in sample sizes that were too small (n=33-55) for ten variables (Osborne & Castello, 2004).

Table 11

Pattern Matrix of Problem 1 Principal Component Analysis

	TRESim Component Label		
TRESim Observable Variables	Prediction and Method	Use of Help Button	Data Organization and Conclusion
8. Number of predictions made	0.928		
9. Number of correct predictions made	0.846		
5. Number, range, and distribution of experiments	0.799		
4. Choice of best experiment to solve problems	0.702		
3. Frequency of using science help button		0.894	
1. Frequency of using computer help button		0.860	
2. Frequency of using glossary button		0.842	

7. Table is useful to solving problem	0.820
6. Graph is useful to solving problem	0.733
10. Score of multiple-choice conclusion item	0.535

Notes: For ease of comparison, variable numbers correspond to original listing as shown in Table 8. This factor loading converged after 4 iterations.

Table 12

Correlation of the Ten Observable Variables from Problem 1

	Frequency of using computer help button	Frequency of using glossary button	Frequency of using science help button	Choice of best experiment to solve problems	Number, range, and distribution of experiments	Graph is useful to solving problem	Table is useful to solving problem	Number of predictions made	Number of correct predictions made	Score of multiple-choice conclusion item
1. Frequency of using computer help button	1.000									
2. Frequency of using glossary button	.546	1.000								
3. Frequency of using science help button	.725	.652	1.000							
4. Choice of best experiment to solve problems	.035	.027	.136	1.000						
5. Number, range, and distribution of experiments	-.035	.007	.022	.807	1.000					
6. Graph is useful to solving problem	-.161	-.057	-.099	.327	.297	1.000				
7. Table is useful to solving problem	-.024	.006	-.027	.312	.290	.324	1.000			
8. Number of predictions made	-.021	.070	.030	.495	.574	.196	.216	1.000		
9. Number of correct predictions made	-.004	.081	.025	.430	.509	.243	.250	.833	1.000	
10. Score of multiple-choice conclusion item	-.047	-.111	-.154	.337	.269	.278	.214	.191	.224	1.000

The mean of each component was calculated so that each component was represented by one value. The means and standard deviations of the components are shown in Table 13.

Table 13

Descriptive Statistics of Problem 1 Components Based on Each School's Treatment

TRESim Component	Mean (Standard Deviation) by School			
	School A: Pre-lab & LEI (*n*=100)	School B: Pre-lab (*n*=63)	School C: LEI (*n*=61)	School D: Comparison (*n*=47)
Prediction and Method	2.40 (1.53)	2.19 (1.55)	2.43 (1.56)	2.29 (1.21)
Use of Help Button	0.20 (0.72)	0.31 (0.78)	0.38 (0.92)	0.16 (0.42)
Data Organization and Conclusion	1.19 (0.71)	0.87 (0.75)	1.07 (0.63)	1.22 (0.66)

Note: School sample size is denoted using *n*.

The homogeneity of variance and normality assumptions were tested for the components shown in Table 13. First, the assumption of homogeneity of variance was violated as indicated by Box's M test (Box's M[18, 157724]=39.191, sig.=0.004); however, this test is highly sensitive to departures from normality. As expected, violations of normality were indicated by significant Kolmogorov–Smirnov (KS) values for all three components. Although the data did not meet the normality assumption, large sample sizes of 100 to 200 have been found to render such violation less problematic for analysis (see Tabachnik & Fidell, 2007; Waternaux, 1976, 1984); thus, a MANCOVA was implemented as planned.

Before the MANCOVA was conducted, correlations between the three components and five possible covariates (see Part 1: Before the TRESim Assessment) were calculated to determine inclusion of covariates in further analyses. The results of the correlation analysis are shown in Table 14.

Table 14

Correlation of the Three Components of Problem 1 and Five Possible Covariates

	Performance-avoid goal orientation	Learning strategies: critical thinking	Frequency of scientific methods used in class	Frequency of general computer use	Prior-knowledge questions	Prediction and method	Use of help button	Data organization and conclusion
Performance-avoid goal orientation	1.000							
Learning strategies: critical thinking	0.267*	1.000						
Frequency of scientific methods used in class	0.100	0.357*	1.00					
Frequency of general computer use	-0.004	-0.006	0.097	1.000				
Prior-knowledge questions	0.082	0.106	-0.039	-0.012	1.000			
Prediction and method	0.041	0.101	-0.032	0.020	0.142	1.000		
Use of help button	0.058	-0.034	0.063	-0.076	-0.065	0.047	1.000	
Data organization and conclusion	0.069	0.055	-0.046	-0.010	0.243*	0.554*	-0.043	1.000

Note: the * denotes the correlation is statistically significant at the 0.01 level.

The results indicated that none of the five possible covariates were associated with the three components[1]. Hence, a two-way MANOVA was conducted and the results indicated no significant differences among the groups. Since no differences were observed, no further analyses were conducted for Problem 1 of the TRESim assessment.

Problem 2: How do different amounts of helium affect the balloon's altitude? With TRESim Problem 2, a similar approach was used to investigate group differences. This time,

[1] The prior-knowledge questions covariate was statistically significantly correlated to the data organization and conclusion component, but the correlation was 0.243, which is considered a weak correlation because it is less than 0.3 (Gravetter & Wallnau, 2009). As only weak correlations were obtained between the observed variables and the covariates, no covariates were used for the subsequent MANOVA analysis.

only nine of the 12 observable variables were used in a PCA because two of the observable variables (i.e., *use of table or graph* and *concluding results*) were not continuous as mentioned previously and therefore unsuitable for the analysis; in addition, one variable – *frequency of using science help button* – displayed no variance, as none of the students in any of the four groups pushed the *science help* button while solving Problem 2. Descriptive statistics for the nine variables are shown in Table 15. Similar to Problem 1, the sample size for the analyses was reduced, this time to 109 students because the coding of -9 for two variables (i.e., *graph is useful to solving problem* and *table is useful to solving problem*) was considered a missing value.

Table 15

Descriptive Statistics of Problem 2 Observable Variables Based on Each Schools' Treatment

Observable Variable	Mean (Standard Deviation)			
	School A: Pre-lab & LEI (*n*=39)	School B: Pre-lab (*n*=25)	School C: LEI (*n*=24)	School D: Comparison (*n*=21)
1. Frequency of using computer help button	0.00 (0.00)	0.16 (0.62)	0.00 (0.00)	0.00 (0.00)
2. Frequency of using glossary button	0.00 (0.00)	0.12 (0.44)	0.08 (0.28)	0.10 (0.30)
4. Choice of best experiment to solve problems	1.95 (0.65)	1.84 (0.75)	1.67 (0.76)	2.24 (0.77)
5. Number, range, and distribution of experiments	1.74 (1.16)	1.56 (1.19)	1.38 (1.35)	1.95 (1.24)
6. Graph is useful to solving problem	2.13 (1.30)	0.96 (1.37)	1.83 (1.47)	1.71 (1.31)
7. Table is useful to solving problem	1.64 (0.93)	0.92 (1.19)	1.42 (1.02)	1.52 (0.87)
8. Number of predictions made	4.56 (3.53)	3.76 (2.92)	3.54 (3.04)	4.00 (3.87)
9. Number of correct predictions made	1.56 (1.80)	1.28 (1.67)	1.33 (1.95)	1.71 (2.47)
10. Score of multiple-choice conclusion item	0.36 (0.49)	0.12 (0.33)	0.29 (0.46)	0.10 (0.30)

Note: School sample size is denoted using *n*.

Next, the result of the PCA extraction and *direct oblimin with Kaiser normalization* rotation using a delta of 0 indicated a simple loading of the nine variables onto three components as shown in Table 16. The use of the *direct oblimin with Kaiser normalization* rotation using a

delta of 0 was justified because of the range of correlations from -0.207 to 0.737 among the nine observable variables as shown in Table 17. The three components found for Problem 2 were labelled based on the content of the variables: method, data organization and conclusion; use of help button; and prediction.

Table 16

Pattern Matrix of Problem 2 Principal Component Analysis

	TRESim Component		
	Method, Data Organization, and Conclusion	Use of Help Button	Prediction
5. Number, range, and distribution of experiments	0.757		
6. Graph is useful to solving problem	0.736		
4. Choice of best experiment to solve problems	0.670		
7. Table is useful to solving problem	0.543		
10. Score of multiple-choice conclusion item	0.507		
1. Frequency of using computer help button		0.929	
2. Frequency of using glossary button		0.920	
8. Number of predictions made			-0.930
9. Number of correct predictions made			-0.927

Notes. This factor loading converged after 6 iterations. For ease of comparison, variable numbers correspond to original listing as shown in Table 12.

Table 17

Correlation of Problem 2 Observable Variables

	Frequency of using computer help button	Frequency of using glossary button	Choice of best experiment to solve problems	Number, range, and distribution of experiments	Graph is useful to solving problem	Table is useful to solving problem	Number of predictions made	Number of correct predictions made	Score of multiple-choice conclusion item
1. Frequency of using computer help button	1.000								
2. Frequency of using glossary button	.737	1.000							
4. Choice of best experiment to solve problems	.140	.130	1.000						
5. Number, range, and distribution of experiments	.034	.037	.661	1.000					
6. Graph is useful to solving problem	-.132	-.207	.280	.360	1.000				
7. Table is useful to solving problem	.060	.029	.298	.310	.291	1.000			
8. Number of predictions made	-.020	-.008	.468	.404	.072	.186	1.000		
9. Number of correct predictions made	-.035	-.066	.347	.351	.018	.189	.773	1.000	
10. Score of multiple-choice conclusion item	-.059	-.033	.212	.338	.296	-.005	.137	.097	1.000

The mean of each of the three components was calculated. The component means and standard deviations are shown in Table 18.

Table 18

Descriptive Statistics of Problem 2 Components Based on Each School's Treatment

TRESim Component	Mean (Standard Deviation)			
	School A: Pre-lab & LEI (*n*=100)	School B: Pre-lab (*n*=63)	School C: LEI (*n*=61)	School D: Comparison (*n*=47)
Method, Data Organization, and Conclusion	1.32 (0.73)	0.92 (0.75)	1.14 (0.72)	1.20 (0.62)
Use of Help Button	0.03 (0.15)	0.06 (0.34)	0.10 (0.64)	0.12 (0.59)
Prediction	2.91 (2.56)	1.99 (1.99)	2.43 (2.21)	2.32 (2.59)

Note: School sample size is denoted using *n*.

Next, the homogeneity of variance and normality assumptions were tested. The homogeneity of variance assumption was tested using the Box's M test. The results indicated that the homogeneity of variance assumption was violated, Box's M[18, 157724]=190.020, sig.=0.000. Normality was violated as indicated by significant KS tests on the means of the three components used. Although the data did not meet the normality assumption, large sample sizes of 100 to 200 have been found to render such violation less problematic for analysis (see Tabachnik & Fidell, 2007; Waternaux, 1976, 1984); thus, the MANCOVA was implemented next. Before the MANCOVA was conducted, a correlation analysis between the three components and five possible covariates was analysed. The results of this correlation are shown in Table 19.

Table 19

Correlation of the Three Components of Problem 2 and the Five Possible Covariates

	Performance-avoid goal orientation	Learning strategies: critical thinking	Frequency of scientific methods used in class	Frequency of general computer use	Prior-knowledge questions	Method, data organization, and conclusion	Use of help button	Prediction

Performance-avoid goal orientation	1.000							
Learning strategies: critical thinking	0.267*	1.000						
Frequency of scientific methods used in class	0.100	0.357*	1.00					
Frequency of general computer use	-0.004	-0.006	0.097	1.000				
Prior-knowledge questions	0.082	0.106	-0.039	-0.012	1.000			
Method, data organization, and conclusion	0.047	0.095	-0.070	0.000	0.198*	1.000		
Use of help button	0.023	0.030	0.009	-0.006	-0.065	0.015	1.000	
Prediction	0.076	0.097	-0.019	-0.002	0.145	0.319*	-0.031	1.000

Note: the * denotes the correlation is statistically significant at the 0.01 level.

The results of this correlation indicated that none of the five possible covariates were suitable to be covariates for these three components[2]. As such, a MANOVA was conducted which included no covariates. Again, the results showed no significant differences among the groups in terms of main effects or interaction. Since no significant differences were observed, no further analyses were conducted for Problem 2 of the TRESim assessment.

Problem 3: How do the amount of helium and payload mass together affect the altitude of a helium balloon? The third TRESim problem had 13 observable variables, but only ten were used. The ten observable variables included the nine used from Problem 2, plus one variable that was exclusive to Problem 3, the *manipulating one variable while controlling others*. The latter variable was used to measure students' ability to manipulate and control variables when three variables are present. The descriptive statistics for the ten observable variables are

[2] The prior-knowledge questions covariate was statistically significantly correlated to the method, data organization and conclusion component, but the correlation was 0.198 which is considered a weak correlation because it is less than 0.3 (Gravetter & Wallnau, 2009). Hence, no covariates were used in the subsequent analysis.

presented in Table 20. Again, the sample size for this analysis was reduced to 102 because of the -9 missing value associated with the variables *graph is useful to solving problem* and *table is useful to solving problem*.

Table 20

Descriptive Statistics of Problem 3 Observable Variables Based on Each School's Treatment

Variable	Mean (Standard Deviation)			
	School A: Pre-lab & LEI (n=44)	School B: Pre-lab (n=21)	School C: LEI (n=23)	School D: Comparison (n=14)
1. Frequency of using computer help button	0.11 (0.39)	0.00 (0.00)	0.00 (0.00)	0.21 (0.58)
2. Frequency of using glossary button	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.21 (0.43)
4. Choice of best experiment to solve problems	2.02 (0.76)	1.43 (0.81)	1.87 (0.82)	1.93 (1.00)
5. Number, range, and distribution of experiments	0.41 (0.79)	0.19 (0.51)	0.35 (0.78)	0.71 (0.99)
11. Manipulating one variable while controlling others	0.36 (0.61)	0.14 (0.36)	0.39 (0.72)	0.43 (0.65)
6. Graph is useful to solving problem	2.57 (0.79)	2.43 (0.81)	2.48 (0.90)	2.43 (0.76)
7. Table is useful to solving problem	1.89 (0.87)	0.67 (0.91)	1.43 (1.16)	1.71 (0.68)
8. Number of predictions made	2.41 (1.92)	1.29 (0.72)	2.96 (2.87)	1.50 (1.45)
9. Number of correct predictions made	0.95 (1.12)	0.43 (0.81)	1.04 (1.69)	0.29 (0.73
10. Score of multiple-choice conclusion item	0.30 (0.46)	0.14 (0.36)	0.22 (0.42)	0.14 (0.36)

Note: School sample size is denoted using *n*.

The results of the PCA extraction and *direct oblimin with Kaiser normalization* rotation using a delta of 0 indicated a simple loading of the ten variables onto four components as shown in Table 21. The *direct oblimin with Kaiser normalization* rotation was used because of the correlations ranging from -0.067 to 0.685 among the ten observable variables, as shown in Table 22. The four components from Problem 3 were labelled based on the content of the variables: *method*; *use of help button*; *prediction*; and *data organization and conclusion*.

Table 21

Pattern Matrix of Problem 3 Principal Component Analysis

	TRESim Component			
	Method	Use of Help Button	Prediction	Data Organization & Conclusion
11. Manipulating one variable while controlling others	.923			
5. Number, range, and distribution of experiments	.854			
4. Choice of best experiment to solve problems	.828			
1. Frequency of using computer help button		.856		
2. Frequency of using glossary button		.843		
8. Number of predictions made			.940	
9. Number of correct predictions made			.943	
7. Table is useful to solving problem				.751
6. Graph is useful to solving problem				.737
10. Score of multiple-choice conclusion item				.547

Note. This factor loading converged after 7 iterations.

Table 22

Correlation of Problem 3 Observable Variables

	Frequency of using computer help button	Frequency of using glossary button	Choice of best experiment to solve problems	Number, range, and distribution of experiments	Manipulating one variable while controlling others	Graph is useful to solving problem	Table is useful to solving problem	Number of predictions made	Number of correct predictions made	Score of multiple-choice conclusion item
1. Frequency of using computer help button	1.000									
2. Frequency of using glossary button	.514	1.000								
4. Choice of best experiment to solve problems	.061	.050	1.000							
5. Number, range, and distribution of experiments	.187	.109	.501	1.000						
11. Manipulating one variable while controlling others	.125	.071	.685	.651	1.000					
6. Graph is useful to solving problem	.005	-.029	.135	.064	.077	1.000				
7. Table is useful to solving problem	.101	.021	.120	.009	.134	.272	1.000			
8. Number of predictions made	.054	-.057	.437	.138	.188	.095	-.042	1.000		
9. Number of correct predictions made	.068	-.067	.269	.075	.183	.176	.008	.739	1.000	
10. Score of multiple-choice conclusion item	-.026	-.048	.220	.104	.171	.181	.170	.075	.017	1.000

The means and standard deviations of the four components are shown in Table 23.

Table 23

Descriptive Statistics of Problem 3 Components Based on Each School's Treatment

Component	Mean (Standard Deviation)			
	School A: Pre-lab & LEI (*n*=100)	School B: Pre-lab (*n*=63)	School C: LEI (*n*=61)	School D: Comparison (*n*=47)
Method	0.76 (0.55)	0.48 (0.49)	0.67 (0.58)	0.74 (0.66)
Use of Help Button	0.06 (0.23)	0.03 (0.20)	0.00 (0.00	0.07 (0.28)
Prediction	1.31 (1.24)	0.80 (0.84)	1.39 (1.59)	1.00 (1.10)
Data Organization and Conclusion	1.13 (0.75)	0.63 (0.70)	0.97 (0.75)	0.76 (0.73)

Note: School sample size is denoted using *n*.

Next, the homogeneity of variance and normality assumptions were tested. The assumption of homogeneity of variance was violated as indicated by Box's M test (Box's M[20, 84176]=33.776, sig.=0.036); however, this test is highly sensitive to departures from normality. Next, the KS test was performed to test the normality assumption. Normality was violated as indicated by significant KS tests for all four of the components. Although the data did not meet the normality assumption, large sample sizes of 100 to 200 have been found to render such violation less problematic for analysis (see Tabachnik & Fidell, 2007; Waternaux, 1976, 1984); thus, a MANCOVA was planned as before. Before the MANCOVA was conducted, however, a correlation was performed between the four components and five possible covariates to determine whether each covariate was suitable to include in the analysis. The results of the correlations are shown in Table 24.

Table 24

Correlation of the Four Components of Problem 3 and the Five Possible Covariates

	Performance -avoid goal orientation	Learning strategies	Frequenc y of scientific	Frequenc y of general	Prior-knowledg	Metho d	Use of help	Predictio n	Data organizatio

		: critical thinking	methods used in class	computer use	e questions		butto n		n and conclusion
Performanc e-avoid goal orientation	1.000								
Learning strategies: critical thinking	0.267*	1.000							
Frequency of scientific methods used in class	0.100	0.357*	1.00						
Frequency of general computer use	-0.004	-0.006	0.097	1.000					
Prior-knowledge questions	0.082	0.106	-0.039	-0.012	1.000				
Method	-0.019	-0.073	-0.025	0.064	0.200*	1.000			
Use of help button	-0.002	-0.090	0.002	0.000	-0.009	0.129	1.000		
Prediction	0.057	0.034	-0.016	-0.026	0.093	0.309*	0.007	1.000	
Data organization and conclusion	0.022	0.049	-0.065	0.069	0.198*	0.405*	0.062	0.225*	1.000

Note: the * denotes the correlation is statistically significant at the 0.01 level.

The results of the correlations indicated that none of the five possible covariates were suitable to include in a MANCOVA[3]. Hence, a two-way MANOVA was conducted instead. The results indicated that the pre-laboratory main effect was not significant, but the LEI main effect was statistically significant, $F(4, 264)=5.358$, $p<0.01$, Pillai's trace=0.075, partial eta squared=0.075. There were no significant interaction effects. The significant LEI effect indicated that students who received the LEI intervention (i.e., Schools A and C) performed significantly better than students who did not receive the LEI (i.e., Schools B and D). Thus, a discriminant function analysis was performed to further analyze component differences between these groups.

[3]The prior-knowledge questions covariate was statistically significantly correlated to the method, as well as the data organization and conclusion components, but the correlations are weak (i.e., it is less than 0.3; Gravetter & Wallnau, 2009). Hence, no covariates were used in the subsequent analyses.

In order to further investigate component differences between the groups receiving the LEI intervention (Schools A and C) and groups not receiving the LEI intervention (Schools B and D), a discriminant function analysis was performed. Although discriminant function analysis does not allow for covariates, the correlation analysis indicated no significant differences among the groups on the covariates. As shown in Table 25, the discriminant function analysis revealed significant differences between the LEI schools and non-LEI schools based on the first function or separation among groups, chi-square=34.032, Wilk's lambda=0.880, $p<0.01$. Parenthetically, although Pillai's trace criterion is normally used to report significant group differences when data violate the homogeneity of variance assumption, Wilk's lambda is the only coefficient produced in a discriminant function analysis. For this reason, Wilk's lambda is used to report significant differences. As shown in Table 26, the first function explained 71.0% of the total variance, which is relatively large.

Table 25

Wilk's Lambda Results of the Three Discriminant Function Components

Test of Functions	Wilks' Lambda	Chi-square	df	Significance
1	.880	34.032	12	.001
2	.963	10.112	6	.120
3	.990	2.727	2	.256

Table 26

Eigenvalues of the Three Discriminant Function Components

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.094	71.0	71.0	.293
2	.028	21.2	92.2	.165
3	.010	7.8	100.0	.101

Table 27

Structure Matrix of Problem 3 Discriminant Function Analysis

	Function		
	1	2	3
Data Organization & Conclusion	.902	.137	-.277
Prediction	.593	-.174	.485
Use of Help Button	.009	.758	-.384
Method	.441	.624	.599

Illustrated in Table 27 are the components that contributed most to the first function. These components included *data organization and conclusion*, *prediction*, and *method*. The *use of the help button* did not contribute much to the first function. The next step was to analyze the groups of students who received the LEI (School A and C) and did not receive the LEI (School B and D) with regard to the first discriminant function and, specifically, the three components that contributed most to the first function. The means and standard deviations of the first discriminant function and three components are shown in Table 28.

Table 28

Means, Standard Deviations, and Univariate MANOVA Results of First Discriminant Function and Three Components Based on the LEI Intervention

	Mean (SD)		Type III Sum of Squares	F	*p*	Partial Eta Squared
	Schools A and C: LEI (*n*=161)	Schools B and D: No LEI (*n*=110)				
Discriminant Function 1	0.243 (1.045)	-0.355 (0.928)	23.352	23.371	0.000*	0.080
Data Organization and Conclusion	1.070 (0.753)	0.683 (0.708)	9.791	18.124	0.000*	0.063
Prediction	1.335 (1.378)	0.882 (0.960)	13.445	8.945	0.003*	0.032
Method	0.708 (0.565)	0.591 (0.581)	0.897	2.746	0.099	0.010

Notes: School sample size is denoted using *n*, while * denotes significant univariate tests using α=0.01.

The assumptions of homogeneity of variance and normality were both violated as indicated by the Box's M test (Box's M[10, 257120]=28.340, sig.=0.002) and the KS test

respectively. A correlation was conducted using the three components of the first discriminant function and the five possible covariates to determine whether each covariate was suitable for further analysis. The results of the correlations are shown in Table 29.

Table 29

Correlation of the Three Components of the First Discriminant Function and Three Components and the Five Possible Covariates

	Performance-avoid goal orientation	Learning strategies: critical thinking	Frequency of scientific methods used in class	Frequency of general computer use	Prior-knowledge questions	Discriminant Function 1	Data organization and conclusion	Prediction	Method
Performance-avoid goal orientation	1.000								
Learning strategies: critical thinking	0.267*	1.000							
Frequency of scientific methods used in class	0.100	0.357*	1.00						
Frequency of general computer use	-0.004	-0.006	0.097	1.000					
Prior-knowledge questions	0.082	0.106	-0.039	-0.012	1.000				
Discriminant Function 1	0.022	0.059	-0.060	0.056	0.219*	1.000			
Data organization and conclusion	0.022	0.049	-0.065	0.069	0.198*	0.909*	1.000		
Prediction	0.057	0.034	-0.016	-0.026	0.093	0.609*	0.225*	1.000	
Method	-0.019	-0.073	-0.025	0.064	0.200*	0.454*	0.405*	0.309*	1.000

Note: the * denotes the correlation is statistically significant at the 0.01 level.

The results of the correlations indicated that none of the five possible covariates were suitable for inclusion in further analyses. Hence, a MANOVA was conducted to determine whether the three components of the first discriminant function were significantly different between the two groups. The results of this MANOVA indicated statistically significant differences, $F(4, 266)=5.930$, $p<0.00$; Pillai's trace=0.082, partial eta squared=0.082, between the two groups of

students, namely, those receiving the LEI versus those that did not. Next, the univariate results were analyzed to determine which of the dependent variables (i.e., three components of the first discriminant function) were significantly different between the two groups. The results, as shown in Table 28, indicate students who received the LEI performed better than students who did not receive the LEI on the first discriminant function and only two of the three components - *data organization and conclusion* and *prediction* but not *method*. The next part of the analyses focuses on investigating differences among the four groups in the post-intervention survey measure.

Part 3: After the TRESim Assessment

Shown in Table 30 are the descriptive statistics of the post-intervention survey measures, including overall group performance on each of the surveys (subscales).

Table 30

Descriptive Statistics and Alpha Coefficients of the Post-Intervention Survey Measure Subscales Based on Each School Treatment

Subscale	Mean (Standard Deviation)				Internal Consistency
	School A: Pre-lab & LEI (*n*=106)	School B: Pre-lab (*n*=53)	School C: LEI (*n*=62)	School D: Comparison (*n*=47)	
Post-Intervention Question Score	7.30 (3.24)	5.44 (2.47)	6.04 (3.15)	5.59 (3.44)	0.545
Emotional Engagement with TRESim	3.37 (1.00)	3.64 (0.78)	3.52 (0.79)	3.57 (0.93)	0.911
Cognitive Engagement with TRESim	2.43 (0.93)	2.95 (0.77)	2.67 (0.78)	2.36 (0.82)	0.809
Specific TRESim Anxiety	2.19 (1.19)	2.89 (1.42)	2.72 (1.40)	2.44 (1.14)	0.837
General Test Anxiety	3.45 (0.16)	3.34 (0.22)	4.41 (0.20)	3.01 (0.25)	0.869
Motivation to use Computers	3.88 (0.08)	4.11 (0.10)	4.03 (0.08)	3.92 (0.08)	0.759

Note: School sample size is denoted using *n*.

Two-way ANCOVAs were performed to investigate whether the two interventions (i.e., pre-laboratory activity and LEI) had any significant effects on each of the post-intervention survey measures. First, the homogeneity of variance and normality assumptions were tested. Second, a correlation analysis was conducted to determine whether each of the five possible covariates previously identified in *Part 1: Before the TRESim Assessment* (i.e., performance-avoid goal orientation, learning strategies: critical thinking, frequency of scientific methods used in class, frequency of general computer use, and prior-knowledge questions) were statistically significant. Only covariates that were significantly correlated with the post-intervention subscales were retained in the analyses.

Post-intervention question score(s). This summative assessment question was designed to evaluate students' abilities to choose the best experiment to solve a problem after the TRESim. The purpose of this question was to better understand whether students were able to utilize the skills they had learned during the TRESim on a new problem. The question required students to solve a problem that involved four variables (i.e., How do the amount of helium, payload mass, and temperature together affect the altitude of a helium balloon?). This question was split into four sub-items that guided students through different areas of problem solving. Students' mean performance across the treatment groups on the question and its four sub-items are shown in Table 31.

Table 31

Descriptive Statistics for Performance on Post-Intervention Question and its Four Sub-Items by School

Component	Mean (Standard Deviation)			
	School A: Pre-lab & LEI (*n*=106)	School B: Pre-lab (*n*=54)	School C: LEI (*n*=63)	School D: Comparison (*n*=47)

How do the amount of helium, payload mass, and temperature together affect the altitude of a helium balloon?	7.30 (3.24)	5.44 (2.47)	6.01 (3.10)	5.59 (3.44)
Sub-items				
1. List the materials needed for your experiment	2.12 (1.13)	1.48 (0.77)	1.84 (1.07)	1.62 (1.05)
2. Steps of your experimental design	3.78 (2.11)	2.83 (1.78)	2.89 (2.03)	2.68 (2.26)
3. What tools will you use to record your data	0.43 (0.55)	0.35 (0.44)	0.34 (0.59)	0.33 (0.53)
4. How will you organize your data	0.97 (0.46)	0.77 (0.32)	0.94 (0.53)	0.96 (0.36)

Note: School sample size is denoted using *n*.

The assumption of homogeneity of variance was tested using Levene's test, which indicated that the post-survey intervention question and its four sub-items satisfied the assumption as shown in Table 32. However, the assumption of normality was violated by all five variables when using the KS test. Although the data did not meet the normality assumption, large samples have been found to render such violations less problematic for analysis (see Tabachnik & Fidell, 2007; Waternaux, 1976, 1984); thus, the ANCOVA was implemented as planned.

Table 32

Levene's Test of Homogeneity of Variance for the Post-Intervention Question and its Four Sub-Items

Component	Mean (Standard Deviation)	
	Levene Statistic	Significance
How do the amount of helium, payload mass, and temperature together affect the altitude of a helium balloon?	2.003	0.114
Sub-items		
1. List the materials needed for your experiment	2.448	0.064

2. Steps of your experimental design	0.928	0.427
3. What tools will you use to record your data	1.155	0.327
4. How will you organize your data	1.786	0.150

Note: $df_1=3$ and $df_2=266$

The two-way ANCOVA included only the prior-knowledge question as a covariate as it was significantly correlated with post-intervention question scores (Pearson Correlation=0.354, $p<0.01$). The ANCOVA analysis revealed that the interaction between the pre-laboratory activity and LEI did not lead to significant effects on the post-intervention measures. As well, the pre-laboratory activity main effect was not significant. However, the LEI main effect was significant, $F(1, 244)=8.084$, $p<0.01$, partial eta squared=0.032, with students in schools receiving the LEI scoring higher than non-LEI schools. The ANCOVA revealed the prior-knowledge question covariate was significantly different between the groups, $F(1, 244)=30.257$, $p<0.01$, partial eta squared = 0.110. Table 33 shows the means of the post-intervention question were higher for the students who received the LEI than the students who did not receive the LEI.

Table 33

Mean and Standard Deviation of Total Post-Intervention Question Score Based on LEI Intervention

	N	Mean (SD)
LEI (Schools A & C)	169	6.820 (3.239)
No LEI (Schools B & D)	101	5.505 (2.946)

The ANCOVA indicated the post-intervention question scores were statistically different between the students who received the LEI and the students who did not receive the LEI.

However, because the post-intervention question included sub-scores, these sub-scores were analyzed (Babenko & Rogers, 2014). Individual ANCOVA analyses were performed on each sub-score to investigate performance differences between students who received the LEI and students who did not receive the LEI. The reason for performing individual ANCOVAs was because the internal consistency (alpha coefficient) of the four items that made up the post-intervention question score was 0.545 (as shown in Table 30). An internal consistency value of 0.545 is relatively low; as well, the correlations among the four sub-scores were low to moderate, as shown in Table 34. Since the internal consistency coefficient and correlations among the four sub-scores were moderately low, it suggested the possibility that there were four different constructs underlying the total post-intervention question score. Hence, four ANCOVA analyses of the sub-scores were undertaken to further investigate the differences among the groups.

Table 34

Correlations of Post-Intervention Question Sub-Scores

	List the materials needed for your experiment?	Steps of your experimental design?	What tools will you use to record your data?	How will you organize your data?
1. List the materials needed for your experiment?	1.000			
2. Steps of your experimental design?	0.555	1.000		
3. What tools will you use to record your data?	0.333	0.246	1.000	
4. How will you organize your data?	0.251	0.272	0.080	1.000

The correlations among the five possible covariates and the four post-intervention question sub-scores, as shown in Table 35, indicate that the only significant covariate was the prior-knowledge question. The prior-knowledge question covariate and one of the four post-intervention questions

- *steps of your experimental design item* – shared a correlation of 0.3, which is considered moderate (Gravetter & Wallnau, 2009).

Table 35

Correlation of the Four Post-Intervention Question Sub-Scores and Five Possible Covariates

	Performance-avoid goal orientation	Learning strategies: critical thinking	Frequency of scientific methods used in class	Frequency of general computer use	Prior-knowledge questions	List the materials needed for your experiment?	Steps of your experimental design?	What tools will you use to record your data?	How will you organize your data?
Performance-avoid goal orientation	1.000								
Learning strategies: critical thinking	0.267*	1.000							
Frequency of scientific methods used in class	0.100	0.357*	1.00						
Frequency of general computer use	-0.004	-0.006	0.097	1.000					
Prior-knowledge questions	0.082	0.106	-0.039	-0.012	1.000				
List the materials needed for your experiment?	0.088	0.074	0.047	0.050	0.296*	1.000			
Steps of your experimental design?	-0.015	0.027	0.014	-0.031	0.300*	0.550*	1.000		
What tools will you use to record your data?	-0.076	-.033	0.071	0.026	0.141	0.333*	0.246*	1.000	
How will you organize your data?	0.052	0.012	0.038	0.018	0.254*	0.251*	0.272*	0.080	1.000

Note: the * denotes the correlation is statistically significant at the 0.01 level.

Hence, only the prior-knowledge question covariate was used in the ANCOVA. As illustrated in Table 36, the results of the four ANCOVAs on the sub-scores indicate that students' performance on *listing materials needed for the experiment* was significantly different, $F(1, 246)=11.972$, $p<0.01$, partial eta squared=0.046, between the students who received the LEI and the students who did not receive the LEI. Specifically, the students who received the LEI statistically outperformed the students who did not receive the LEI on this sub-score. The prior-knowledge question covariate was also significant, $F(1, 246)=21.055$, $p<0.01$, partial eta squared=0.079.

Table 36

ANCOVA Results for the Four Post-Intervention Question Sub-Scores

	Mean (SD)		Type III Sum of Squares	F	*p*	Partial Eta Squared
	Schools A and C: LEI (n=154)	Schools B and D: No LEI (n=94)				
Listing materials needed for the experiment	2.05 (1.110)	1.53 (0.924)	12.059	11.972	0.001*	0.046
Steps of your experimental design	3.52 (2.121)	2.80 (1.960)	20.157	5.150	0.024	0.021
What tools will you use to record your data	0.399 (0.570)	0.330 (0.484)	0.172	0.602	0.438	0.002
How will you organize your data	0.965 (0.499)	0.867 (0.345)	0.316	1.682	0.196	0.007

Notes: School sample size is denoted using *n*, while * denotes significant univariate tests using α=0.01.

Emotional engagement with TRESim. The assumption of homogeneity of variance and normality were violated, as indicated by the Levene and KS tests, respectively. However, the large sample size renders this violation less problematic for analysis, and the ANCOVA was conducted (Gravetter & Wallnau, 2009). The initial two-way ANCOVA, which used the *learning strategies: critical thinking* (Pearson Correlation=0.316, $p<0.01$) covariate, revealed the

covariate was significantly different among the groups, $F(1, 241)=16.771$, $p<0.01$, partial eta=0.065. This initial two-way ANCOVA also indicated that the pre-laboratory activity, as well as the LEI, and their interaction had no significant effects on students' emotional engagement with the TRESim.

Cognitive engagement with TRESim. The assumption of homogeneity of variance was satisfied (Levene's Statistic[3, 265]=1.652, $p=0.178$), but normality was violated as indicated by the KS test. Again, the large sample size renders this violation less problematic for the ANCOVA analysis (Gravetter & Wallnau, 2009). A two-way ANCOVA was conducted next. The ANCOVA, which included the covariates, *learning strategies: critical thinking* (Pearson Correlation=0.447, $p<0.01$) and *frequency of scientific methods used in class* (Pearson Correlation=0.408, $p<0.01$), revealed that both these covariates were significantly different among the groups; for *learning strategies: critical thinking*, $F(1, 241)=29.915$, $p<0.01$, partial eta squared=0.110 and for *frequency of scientific methods used in class*, $F(1, 241)=18.948$, $p<0.01$, partial eta squared=0.073. However, the ANCOVA also indicated that neither the pre-laboratory activity nor the LEI, or their interaction had effects on students' cognitive engagement with the TRESim.

Specific TRESim Anxiety. The assumption of homogeneity of variance was satisfied (Levene's Statistic[3, 265]=2.437, $p=0.065$), but normality was violated, as indicated by the KS test. However, the large sample size allowed for the use of ANCOVA (Gravetter & Wallnau, 2009). A correlation was conducted to determine which covariates were suitable; however, none of the covariates were above 0.3, which indicated the correlations were weak. Hence an ANOVA was performed and indicated the pre-laboratory activity and the LEI, along with their interaction had no effect on students' specific TRESim anxiety.

General test anxiety. The assumption of homogeneity of variance was satisfied (Levene's Statistic[3, 265]=0.180, p=0.910), but normality was violated, as indicated by the KS test. However, the ANCOVA was justified, even with the violation of the normality assumption, because of the large sample size (Gravetter & Wallnau, 2009). The initial ANCOVA, included only the *performance-avoid goal orientation* because it had a significant correlation with the dependent measure (Pearson Correlation=0.310, p<0.01), which indicated the covariate was significantly different among the groups, F(1, 244)=22.068, p<0.01, partial eta squared=0.083. The two-way ANCOVA revealed that the pre-laboratory activity had a significant effect on students' general test anxiety, F(1, 244)=7.127, p<0.01, partial eta squared=0.028. The LEI and the interaction had no effect. Table 37 shows that students' self-reported general test anxiety was generally higher when they did not receive a pre-laboratory activity than when they did receive the pre-laboratory activity.

Table 37

Mean and Standard Deviation of General Test Anxiety

	n	Mean (SD)
Pre-laboratory	159	3.4167 (1.623)
No Pre-Laboratory	110	4.1509 (1.645)

Motivation to use computers. The assumption of homogeneity of variance was satisfied (Levene's Statistic[3, 265]=2.471, p=0.062), but normality was violated, as indicated by the KS test. The large sample size used in this study allowed for the implementation of the ANCOVA (Gravetter & Wallnau, 2009). A correlation to determine which covariate was suitable for this analysis indicated that all the covariates had weak correlations (i.e., less then 0.3). Hence, an ANOVA was conducted and revealed that neither the pre-laboratory activity, LEI, nor their

interaction led to significant effects on student motivation to use computers. These results are discussed in the next section.

Discussion and Conclusion

Computer simulated science laboratories (CSSL) are popular in classrooms as a supplement to science laboratories (Ma & Nickerson, 2006). They are often used as a teaching and assessment tool to measure students' science knowledge and skills. CSSLs are designed to provide students with a flexible learning environment that allows engagement in scientific inquiry, while at the same time offering an opportunity for dynamic assessment to measure students' acquisition of laboratory knowledge and skills. Despite the many intended benefits CSSLs may bring to the science classroom, the implementation of these tools needs to be further investigated to maximize their potential as a learning and assessment tool.

The investigation reported in this dissertation focused on the implementation of two interventions – pre-laboratory activity and learning error intervention (LEI) – designed to intensify the benefits of CSSLs. The specific CSSL used in this study is called the National Assessment of Educational Progress' (NAEP) Problem-Solving in a Technology Rich Environment science laboratory simulation (TRESim; Bennett et al., 2007). The investigation was guided by three research questions:

(a) What are the effects of a pre-laboratory activity on students' socio-emotional experiences, as well as on their understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(b) What are the effects of an LEI on students' socio-emotional experiences, as well as on their understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

(c) What are the interactions between the pre-laboratory activity and LEI on students' socio-emotional experiences, as well as on their understanding of science knowledge and problem-solving skills as measured by the NAEP TRESim science laboratory?

These questions were designed to investigate whether the two interventions had any effects on students' affective experiences with the TRESim, as well as their performance during the TRESim. The next sections will discuss the results of the study and provide answers to these three research questions based on the results of the study.

Research Question 1: Effects of a Pre-Laboratory Activity

The results of this study indicated that students who were administered a pre-laboratory activity reported lower general test anxiety when compared with their peers who did not receive the activity. Although previous studies have not investigated the specific relationship between pre-laboratory activities and general test anxiety, these results are complementary to results from previous studies in terms of the general benefits of using pre-laboratory activities. Many of the outcome measures used in these previous studies focused on affective indicators, such as self-efficacy and motivation and achievement. For example, research designed to investigate students' affective dispositions towards the use of pre-laboratory activities indicated a positive response towards them (Chittleborough, Mocerino, & Treagust, 2007; Supasorn, Suits, Jones, & Vibuljan, 2008). For example, in Chittleborough and colleagues' study (2007), which investigated students' use of a pre-laboratory activity, the results indicated that 70% of the students reported the activity increased their self-efficacy and motivation towards the laboratory. Students reported that completing the pre-laboratory activity helped them feel more confident about the laboratory, which helped improve their enjoyment of the laboratories (Chittleborough et al., 2007). The finding indicates that students had a positive attitude towards the use of pre-

laboratory activities. These findings are complementary to the results of this current study, which showed that students who received the pre-laboratory activity self-reported significantly lower levels of general test anxiety compared to students who did not receive the activity.

Although general test anxiety was not directly measured in the previous studies cited (e.g., Chittleborough et al., 2007), the concepts of self-efficacy and test anxiety are correlated (Bandalos, Yates, & Thorndike-Christ, 2005; Chu, Guo, & Leighton, 2013; Hodapp & Benson, 1997). That is, students who report greater self-efficacy in relation to an activity tend to report lower general anxiety in relation to the activity. As such, pre-laboratory activities may be viewed as an educational tool that can help increase students' self-efficacy and confidence before a laboratory, which may help to decrease general test anxiety about evaluative aspects of the laboratory. In turn, when students feel less anxiety about their performance on a test or task, it may allow them to increase their enjoyment of the test or task.

Affective dispositions and socio-emotional experiences, such as attitudes towards the use of a pre-laboratory activity, are important to consider in better understanding how students use this tool to enhance their learning. For example, Chaby, Sheriff, Hirrlinger, and Braithwaite (2015) found that exposure to stress during the early stages of the learning process improve students' performance during the later stages of learning. The authors explain this finding in context of the Yerkes-Dodson law which indicates that a students' performance may increase with psychological or mental stimulation up to a certain point when the stimulation becomes too high and the performance decreases. The Yerkes-Dodson law can be used to explain that the exposure to the low-level stresses brought about by a pre-laboratory activity may improve students' performance on the subsequent tasks by providing the stimulation needed to maximize

learning. Furthermore, acute stress has been shown to bolster memory, which is often positively correlated with academic achievement (Smith, Floerke, & Thomas, 2016).

Improving students' learning is a major objective of pre-laboratory activities; thus, it is necessary to discuss the use of the pre-laboratory activity intervention in terms of how it influences achievement. Many studies that have investigated the use of pre-laboratory activities as part of a hands-on laboratory have found significant improvement in students' achievement as measured by written reports and summative tests (Johnstone, Watt, & Zaman, 1998; Supasorn, Suits, Jones, & Vibuljan, 2008). The results of these studies show that pre-laboratory activities were often viewed by teachers as a necessity to help prepare students for the laboratory. The teachers explain that access to the laboratory and equipment is often limited to a pre-set amount of time; as such, students need to enter the laboratory with the required background knowledge so that time spent in the laboratory may be used to deepen understanding of the materials presented. In other words, the pre-laboratory activity acted as an education tool in two ways: it drew out students' prior experiences and served to expose the knowledge and skills they were struggling with ahead of the actual laboratory; thus, the pre-laboratory activity served to provide feedback to students in terms of the required knowledge and skills prior to the laboratory.

The idea of preparing students to arrive at a laboratory with the necessary background knowledge and skills allows the laboratory to be used as intended, and to build on students' existing content knowledge, so that more emphasis can be placed on the process of application and inquiry skills (Reid & Shah, 2007). Having a good understanding of basic content knowledge is often a pre-requisite to application and inquiry skills, in which the content is applied in different ways to solve a variety of problems. Additionally, pre-laboratory activities can serve as an effective tool in providing feedback to students. As students complete the pre-

laboratory activities, they engage in a type of self-assessment where they are prompted to enhance or review areas of weakness if they encounter portions of the activity they find challenging. As indicated by Johnstone et al. (1998) and Supasorn et al. (2005), laboratory instructors can also provide detailed and individual feedback on students' pre-laboratory activities so that learning errors in students' performance and their understanding of the basic content knowledge can be rectified prior to the hands-on laboratory. When pre-laboratories activities are integrated as part of the learning process, some studies indicate these activities can improve students' achievement scores (Chittleborough et al., 2007; Supasorn et al., 2005).

Although the current study did not find a statistically significant, enhancing effect of the pre-laboratory activity on students' TRESim performance and post-intervention question score, the main reasons for this finding may be due to the following limitations of the activity: (a) type of activity, (b) lack of personalized feedback, (c) new reflection format, and (d) explicit explanation of activity as part of learning. First, the pre-laboratory activity essentially involved the first problem of the TRESim and focused on preparing students with the process skills needed to solve the TRESim problems. The nature of this activity may not have resembled or proven familiar to the pre-laboratory activities that students in this study had experienced in the past. Most students may have been more familiar with pre-laboratory activities that focused on reviewing the conceptual theories underlying the problem instead of a practice run at the process of solving the problem. This difference in the pre-laboratory activity may have confused students and caused students to not do much with the activity in preparation for the TRESim.

Second, the pre-laboratory activity feedback provided to students consisted of only a numerical score to ensure consistency among the students who received this intervention. The reason for only providing scores, instead of elaborative and personalized comments, was to

ensure (control for) consistency of the feedback to the students who received the pre-laboratory activity. If elaborative and personalized feedback had been provided to students, then the specific nature of the feedback could have functioned as a confounding variable in the study. This should be included in future studies but controlled. It should be noted that research studies do suggest the benefit of providing timely and personalized feedback on pre-laboratory activities to allow students the opportunity to learn from their mistakes and therefore improve their results during the experiment (Chittleborough et al., 2007; Reid & Shah, 2007; Shute, 2008). Personalized feedback, which highlights each student's learning errors along with ways to tackle those errors, has been deemed useful for improving students' understanding because it targets their specific areas of weaknesses so that they may focus on improvements (Shute, 2008).

Third, the time provided for students to review their pre-laboratory activity may not have been properly used by students to reflect upon their errors to enhance their areas of weaknesses. After the graded pre-laboratory activities were returned to students, it is likely that many students looked at their scores, but did not review their actual performance (i.e., they likely failed to reflect on the reasons behind their performance). This lack of reflection and review would have prevented them from learning based on their learning errors as a source of information to enhance their knowledge and skills. By not reviewing their learning errors, students would have continued to make the same errors during the TRESim assessment as they did on the pre-laboratory activity. Considering the importance of formative feedback to the progression of learning (Shute, 2008), future research needs to focus on administration of pre-laboratory activities and the conditions for feedback, including timing and content of feedback, to enhance student understanding and performance.

Lastly, the pre-laboratory activity was not explicitly introduced as part of the TRESim and post-intervention survey measures. The reason for not introducing the pre-laboratory activity as part of the TRESim and post-intervention survey measures was to minimize the explicit instruction given to students during the first day of data collection. On Day 1, students were introduced to the study and invited to participate. Considering the large amount of information regarding the purpose and value of the study to students, discussion surrounding the pre-laboratory activity was not included.

It is important to note that research results do suggest that students are likely to be receptive to pre-laboratory activities when they are introduced as part of the learning process (e.g., by explicitly explaining that the activity was designed to draw out prior experiences and ideas that prepare students for the actual laboratory; Reid & Shah, 2007; Supasorn et al., 2008). Future research should investigate the conditions for how best to incorporate pre-laboratory activities in the learning process so that they serve their intended function. One condition may be incorporating an explicit discussion of the connection between the activity and the learning goals of the laboratory; variables inherent to the discussion might include content and level of detail in relation to student ability, motivation, and interest.

Research Question 2: Effects of a Learning Error Intervention (LEI)

The results of this study indicated that students who received the LEI scored significantly higher on two components of Problem 3 of the TRESim assessment and on a sub-section of the post-intervention survey measure than students who did not receive the LEI. These results are consistent with the LEAFF model in which an explicit discussion regarding the necessity of learning errors as part of the learning process is expected to improve students' performance (see also Firestein, 2016). Previous studies have found that students exposed to the LEI show stronger

performance on indicators of learning for meaning (e.g., identification of areas of confusion; see Leighton & Bustos Gomez, under review).

Specifically, compared to students who did not receive the LEI, students who received the LEI outperformed their peers on two components of Problem 3: *data organization and conclusion*, and *prediction*. Problem 3 was the most difficult problem administered during the TRESim. The question asked in Problem 3 was "how do the amount of helium and payload mass together affect the altitude of a helium balloon?" (NAEP. 2007, pg. 25). This problem required students to find the relationship between three variables, which involved holding one variable constant while manipulating the second variable to observe the changes in the third variable. Problem 3 allowed students to use several methods to solve the problem, thus, making this problem relatively open-ended. This type of exploration problem would have been well suited to students who felt free to experiment and show innovation following the LEI.

Additionally, the results of the post-intervention survey measure showed that students who received the LEI significantly outperformed their peers who did not receive the intervention on the post-intervention question (see Table 26). An analysis of the four sub-scores (i.e., sub-items) of the post-intervention question indicated that students who received the LEI performed significantly better on the first sub-item, which asked students to list the materials needed for the experiment. Thus, the LEI might have encouraged students to engage in more open and comprehensive thought about the problem. The challenge now is to understand the specific pedagogical mechanisms by which the LEI might encourage this higher-level thinking and performance.

These results are consistent with the LEAFF model. The LEAFF model outlines that students who perceive their learning environment as safe may feel at ease making more errors

during the early training stages of learning because they feel authorized to experiment and state what they do not understand, including taking risks with their thinking and learning. However, as they become more skilled with the content and respond to feedback, the number of errors should decrease, as represented by an increase in summative assessment performance (Leighton, Chu, & Seitz, 2013). Although there were no group differences on Problems 1 and 2 of the TRESim assessment, it is possible that by the time students progressed to Problem 3 and the post-intervention survey measure, the students who received the LEI felt more at ease tackling the most difficult of problems. Alternatively, Problem 3, given its greater demand for higher-level thinking, may have been more sensitive to those students who were willing to engage in more innovative thought as part of their LEI exposure. This result is in line with previous research (see Leighton & Bustos Gomez, under review) that suggests that a simple LEI can help students become more aware of what they do not understand and thus be more open to receiving feedback to improve learning.

However, this present study differed from previous investigations of the LEAFF model in that it did not include a tally of the number of learning errors students made during the TRESim. While prior LEAFF studies have investigated the number of errors identified during the instruction students are receiving in the classroom (Chu & Leighton, 2016; Leighton & Bustos Gomez, under review), this study focused on students' achievement at the end of each TRESim problem and on the post-intervention survey measure. The reason for not focusing on the number of learning errors students made was because the TRESim assessment did not have a good system for identifying and tallying number of errors. Hence, the only type of learning error that was identified by TRESim was the number of incorrect hypotheses or predictions made by students. Incorrect hypotheses or predictions may be a good representation of learning errors as

these hypotheses reflected potentially wrong but educated guesses of what would happen to the balloon after the value of a manipulated variable(s) was selected. Prior to running each trial, students were prompted by the TRESim to make a hypothesis. The TRESim assessment did not limit the number of hypotheses or trials students could perform for each problem. Thus, number of hypotheses and trials is likely to have varied among different students. Students who had an increased number of incorrect hypotheses or predictions might have performed more trials; those incorrect hypotheses could have been coded as learning errors when compared with other students who ran fewer trials. This is worth considering in future analysis of these data.

Additionally, as students progressed through the TRESim assessment, the difficulty of the three problems administered increased as more variables were introduced. For example, Problem 1 of the TRESim required students to investigate the linear relationship between two variables, while Problem 3 required students to investigate the relationships among three variables. Since the number of variables increased with each new TRESim problem, more trials were expected and needed to properly determine the relationship between or among the variables. Again, the difficulty of the problem would have led to an increased number of hypotheses or predictions, some of them incorrect, for more difficult problems. Although problem difficulty is confounded with the expected number of hypotheses, it would be possible and important to examine how students exposed to the different interventions performed in the number of hypothesis proposed.

Although the results of this study are consistent with predictions derived from the LEAFF model, more research is needed to understand the sample and classroom characteristics required for the LEI to lead to improvements in student learning and achievement (e.g., Chu & Leighton, 2016; Leighton & Bustos-Gomez, under review). For example, in Chu and Leighton's (2016)

investigation of the impact of a LEAFF-based intervention on students' affective dispositions, learning errors, and feedback preferences in small, undergraduate computer programming labs, they found that students who received the LEI reported significantly more errors during the learning phase compared to their control peers who did not receive the intervention but their final grades did not show a statistically significant improvement. The Chu and Leighton (2016) study involved a relatively small sample and took place over an entire 13-week term. The present study administered the TRESim over one class period.

In addition, it is important to note the many other variables that distinguish the Chu and Leighton (2016) study compared to the present one – sample characteristics and size were different (e.g., Chu and Leighton included 18 and 10 per treatment group), content matter was different (e.g., Chu and Leighton focused on computer programming skills), and the way in which the LEI was delivered were different (e.g., Chu and Leighton delivered the LEI verbally at the beginning of every laboratory). It is important to note that the LEAFF model does not suggest student summative assessment scores will be different immediately following an LEI but should have an effect over time with increased focus on correcting misunderstandings. For example, a study conducted by Leighton and Bustos Gomez (under review) revealed that students who received the LEI in a single session indicated more positive feelings, higher levels of trust in the instructor, and reported more errors during the learning process than their peers who did not receive the intervention. However, there was no expectation that a single session of the LEI would lead to differences on a summative assessment given the quickness of the intervention.

The results of this current study indicate students who receive the LEI may tend to perform significantly better on assessments that are administered later during a learning process

characterized by exploration and remediation of learning errors. The factor of time may play an important role in improving students' summative assessment scores when using the LEAFF model as a framework. Although Leighton and Bustos Gomez (under review) did not notice an improvement in summative assessment scores after a single LEI session, the current study found improvements in students' performance after repeated exposure to learning environments deemed safe by the students.

Research Question 3: Interaction of Pre-Laboratory Activity and Learning Error Intervention (LEI)

The results of this study indicated that there were no significant interactions between the pre-laboratory activity and LEI on the dependent variables measured. This result is not surprising considering the literature supporting the use of the pre-laboratory activity and LEI did not suggest an interaction effect would be present. Although both interventions were designed to improve students' performance on the TRESim assessment, the interventions included different approaches. The pre-laboratory activity was designed to provide students with the necessary background experiences to succeed in the TRESim, while the LEI encouraged students to explore their thinking about mistakes during the learning process. In a sense, the pre-laboratory activity focused on providing the *content knowledge* needed for the assessment while the LEI focused on the *process of learning*. Future research needs to explore whether specific conditions can bolster the combination of these interventions – for example, by helping students appreciate the value of the pre-laboratory activity as an opportunity to explore learning errors to deepen understanding.

The previous sections discussed the results of the three research questions guiding this study. The research questions were designed to investigate whether two interventions – pre-

laboratory activity and LEI – had an effect on students' socio-emotional experiences and performance during the TRESim assessment. The next sections provide a summary of the full study conducted.

Summary of the Study: Purpose, Method, and Results

Purpose. The main purpose of this research study was to investigate whether two interventions, specifically a pre-laboratory activity and LEI, impacted student socio-emotional experiences and enhanced their learning and performance on a CSSL designed to measure students' science knowledge and skills. These two interventions were hypothesized to improve students' performance during the CSSL for two reasons. First, pre-laboratory activities are tools to cognitively prepare students for hands-on laboratories, but they are seldom used with CSSLs (PhET, 2015). Second, the LEI was designed to encourage students to make training errors during the learning process to promote the development of scientific inquiry skills.

Together, these interventions are closely related to the development of scientific inquiry skills which focus on the different ways of discovering knowledge or solving a problem, similar to how real-life engineers and scientists approach answering questions (National Research Council, 2006). Scientific inquiry focuses on the idea that multiple plausible methods or solutions may be used to arrive at the same conclusion. This skill is considered a fundamental concept in science education. Since scientific inquiry is considered a higher-level application of knowledge, students need a solid foundation of understanding basic concepts. Hence, a pre-laboratory activity was designed to provide students with the opportunity to interact with background knowledge needed for the TRESim, while the LEI focused on the inquiry process by underscoring the necessity of trial-and-error thinking, exploration without concern for errors during training, and encouraging the use of formative feedback to inform later trials.

Method. These two interventions (manipulated variables) – pre-laboratory activity and LEI – were chosen and designed to enhance students' experiences on CSSLs so that the benefits of the simulated assessments could be maximized. In order to investigate whether these interventions enhanced students' performance during and after the CSSL, a 2×2 quasi-experimental design was used to measure Grade 8 science students' performance. In total, 298 students and 10 teachers from four schools were assigned to one of four treatments as shown:

	Pre-Lab Activity	No Pre-Lab Activity
LEI	School A (n=108)	School C (n=69)
No LEI	School B (n=73)	School D (n=48)

All students were administered a series of pre- and post-intervention survey measures and the TRESim assessment. The pre- and post-intervention survey measures consisted of surveys and achievement items. Specifically, the pre-intervention survey measure consisted of items used to determine pre-existing differences among the four groups in terms of goal orientations, motivational learning strategies, use of scientific process, and prior science knowledge and skills. The post-intervention survey measure consisted of a single achievement question designed to assess students' abilities to design the best experiment to solve a problem, as well as survey items designed to measure emotional and cognitive engagement, general and specific anxiety, use of computer technology, and other demographic variables. Students in each of the assigned schools were administered the appropriate combination of the two interventions – pre-laboratory activity and LEI.

The pre-laboratory activity was completed by students in groups A and B only. This activity was designed for students to review basic concepts related to scientific processes needed to solve problems on the TRESim. The pre-laboratory activity included questions (e.g., identifying the manipulated, responding, and controlling variables) that probed students' foundational knowledge and skills needed to solve problems using the scientific method.

The LEI was administered only to students in groups A and C. The LEI consisted of a scripted presentation, which was delivered to students using PowerPoint. The presentation was designed to explicitly inform students that making learning errors or mistakes is an important part of the training phase in learning. Throughout the presentation, students were encouraged to share their own experiences of making mistakes while they were learning a new skill and to describe how those mistakes helped them to better learn the skill. This LEI was based on the LEAFF model (please refer to Leighton et al., 2013 paper for full details of the model).

Results. The results of this study were split into three sections: (a) pre-intervention survey measure, (b) TRESim assessment, and (c) post-intervention survey measure. Results of the pre-intervention survey measure indicated that the groups differed on five subscales, namely performance-avoid goal orientation, learning strategies: critical thinking, frequency of scientific methods used in class, frequency of general computer use, and prior-knowledge questions. These five subscales were then considered as covariates throughout the latter two sections of the analyses to control for pre-existing differences among the treatment groups.

The results of the TRESim assessment indicated that students who received the LEI performed significantly better during the third, and most difficult, TRESim problem compared to their peers who did not receive the LEI. Specifically, students who received the LEI performed significantly better on two components of Problem 3: *data organization and conclusion* and

prediction. Performance on the first and second TRESim problems was not significantly different among the groups; however, these problems were also easier than Problem 3.

Analyses of the post-intervention survey measure revealed that students who were administered the pre-laboratory activity reported significantly lower levels of general test anxiety than those who were not administered the pre-laboratory activity. Furthermore, students who received the LEI performed significantly better on the post-intervention survey measure's single-item question compared to those students who did not. Further analyses of the single-item question – specifically, analyzing the four individual sub-items that made up the question's total score – indicated that students who received the LEI scored significantly higher than control, non-LEI students, on the sub-item that required students to list materials needed for the experiment.

Importance of Study and Implications for Practice

CSSLs have been shown to improve student achievement when used as a supplement to classroom activities (PhET, 2015; Scalise, Timms, Clark, & Moorjani, 2009; Quellmalz, Timms, & Schneider, 2009). This study provided preliminary evidence that a pre-laboratory activity can help reduce general test anxiety and the LEI can improve students' CSSL performance on difficult problems. Knowing the beneficial aspects of these kinds of interventions may help educators better utilize CSSLs in the classroom so that digital learning and associated assessment tools may be maximized. The result of reducing test anxiety by administering a pre-laboratory activity highlights the importance of considering pre-laboratory activity with CSSLs. When students' anxiety about an assessment is reduced, the reliability of their performance on the assessment would be expected to increase as students' performance can be revealed without construct irrelevant variance generated from test anxieties. Moreover, the use of the LEI may

help students develop a more confident and realistic perception of what it means to learn complex material and explore increasingly demanding questions in science (Firestein, 2016). While additional research needs to be conducted, CSSLs may be especially helpful in providing students with a dynamic experience that differs from previous, static educational encounters; thus, making it important for teachers to prepare students by reviewing necessary background knowledge and skills, and ensuring that students realize errors are a part of thinking at higher levels. By helping students understand the higher-level knowledge and skills that CSSLs are designed to measure, educators can prepare students to make the most of these tools, thereby deepening their conceptual understanding of science.

The results of this study may also inform the development of future digital simulations. While designers of these simulated learning environments have a tendency to focus on ensuring programs function properly or developing coherent directions for tools, they should also consider the preparatory work students should complete in order to have the necessary background knowledge and mindset to delve into the dynamic environments that will allow students to profit from these kinds of tools. Preparatory work (e.g., pre-laboratories) and explicit guidance on how errors can inform deeper learning could be easily integrated into a CSSL so that all students will be exposed to the same resources.

Limitations of the Study

There are a number of limitations in the study but two that deserve discussion: first, the sample of convenience and, second, the open-ended data capture of the TRESim assessment. First, there is potential for bias in the data when a sample is not random but rather drawn in convenience. The convenience sampling in this study was composed of students who were enrolled in the same school district that granted ethics approval prior to the start of data

collection. This sample of convenience also resulted in an unbalanced design because group sizes were relatively constrained by the number of grade 8 students enrolled in each school. A larger sample size that included two or three schools for each of the treatment groups could have overcome this limitation. Unfortunately, educational research is often conducted with samples of convenience given the challenges with recruiting schools and attempting to minimize disruptions in classroom time.

Second, although the TRESim assessment was developed by NAEP, the technical support (i.e., development of the log-file data capture mechanism) for using the TRESim was not provided by the NAEP, the original developers, but instead by a team of technology experts at the University of Alberta (i.e., Technologies in Education). Hence, the TRESim program could not be altered significantly by the researcher to have a better method of capturing data. For example, the TRESim program did not capture all the open-ended responses during student performance. The log files of students' actions made during performance on the TRESim only included 266 characters (including spaces) for each of the four open-ended items. This limitation was not known to either the researcher or students during data collection; hence, many students wrote relatively lengthy responses to the open-ended questions but the responses were not fully captured. Only the first portion of many responses was recorded, which led to many answers being abruptly cut off in the log files. To overcome this limitation, the data capturing mechanism of TRESim, or any CSSL, should either (1) not accept any further input and warn the students when the character limit is reached or (2) be programmed to capture more than 266 characters (e.g., 750 characters) for open-ended items. Alternatively, the test administrator, teacher or CSSL instructions needs to alert students that their answers may only contain a maximum number of

characters. One way to ensure students' responses do not exceed the limit is to minimize the response window so that any key strokes exceeding the maximum allowed are not displayed.

Future Studies

The use of CSSLs as learning and assessment tools has potential and may become an integral part of science education. Although many simulated laboratories are currently used as supplements to science lessons, more development and research of these digital learning environments is needed to help maximize their benefits in classrooms (PhET, 2015). Findings from this study contribute towards a few streams of research that are now presented.

Improving the reliability of CSSL score reports is an important area of future research. In general, the reliability of score reports increase with additional supportive evidence (AERA, APA, & NCME, 2014). As an example, a study could be conducted to investigate whether administering a series of construct-related CSSLs during one unit of study to a group of students leads to similar performance results. The reliability of performance scores on each variable could be calculated after each CSSL to observe how the reliability increases as new evidence is made available from each CSSL. The use of multiple CSSLs to improve the reliability of scores (and outlined in score reports) would be beneficial to provide students with increasingly accurate feedback about their performance. For example, if a student's performance on one CSSL indicates that the student does not understand a concept, this result may be due to a variety of reasons (e.g., lack of conceptual understanding, mistake while reading or answering the question, and misunderstanding of the item). However, if that student's performance is similar on five CSSLs, then this presents stronger evidence that the student has not acquired the requisite knowledge and skills in conceptual understanding. Improving the reliability of scoring and reports may also inform future studies by allowing random mistakes to be parsed out from

consistent learning errors. This would allow students to focus on specific areas that they conceptually do not understand instead of expending too much effort on minor mistakes.

A series of CSSLs can also complement learning by enhancing the tracking of student improvement. Administering CSSLs that collect data on achievement over a period of time may allow tracking of whether students have improved their knowledge and skills. The data collected would need to target a specific construct within the context of the CSSL. For example, if five CSSLs were used to track the construct of *scientific inquiry*, the results could be used to indicate students' skills in scientific inquiry over time. Tracking improvement using multiple assessments that focus in on a given construct is congruent with other studies on formative and embedded assessment, such as the SEPUP course (see Wilson & Sloane, 2000), which was presented earlier in the literature review chapter. In particular, formative and embedded assessments highlight the provision and use of feedback from previous assessments to inform students of their areas of weakness, and in turn help them focus on which areas to enhance (Black & Wiliam, 1998). The continued feedback produced from CSSL performance may inform students in terms of whether they have mastered the knowledge and skills associated with a specific construct throughout the period of time. For example, if students did not learn a construct on the first CSSL, they would most likely work on enhancing that area. After a period of time a second CSSL could be administered to measure the same construct, if the feedback from this assessment indicates the student has improved, but still has not mastered the construct, then more work will be needed to ensure the student fully understands the construct. The use of CSSLs as tools that involve a series of formative, and/or embedded assessments to measure a certain construct could allow students to track their learning and improve their understanding for the given construct.

Providing students with relevant feedback (information) about their learning could help students regulate their efforts to learn by focusing on areas of weakness.

A third stream of future research could focus on scientific inquiry by developing CSSLs that allow for more open-ended solutions. By developing CSSLs that allow for more open-ended solutions, it would be possible to investigate whether students can explore different solution methods as opposed to only a few because of restrictions of the program. Future CSSL development could provide students with more open-ended simulations that allow for relatively unlimited digital versions of laboratory equipment. Some simulations, such as the one shown in Figure 10, have been developed to let students use as many resources (e.g., wires, resistors, and batteries) as the screen will fit (PhET, 2015). In this example the light bulb in the direct current circuit will light up as long as the bulb and battery form a closed circuit (i.e., when the wires are connected to each other forming a loop) and this simulation lets students choose from many resource options.

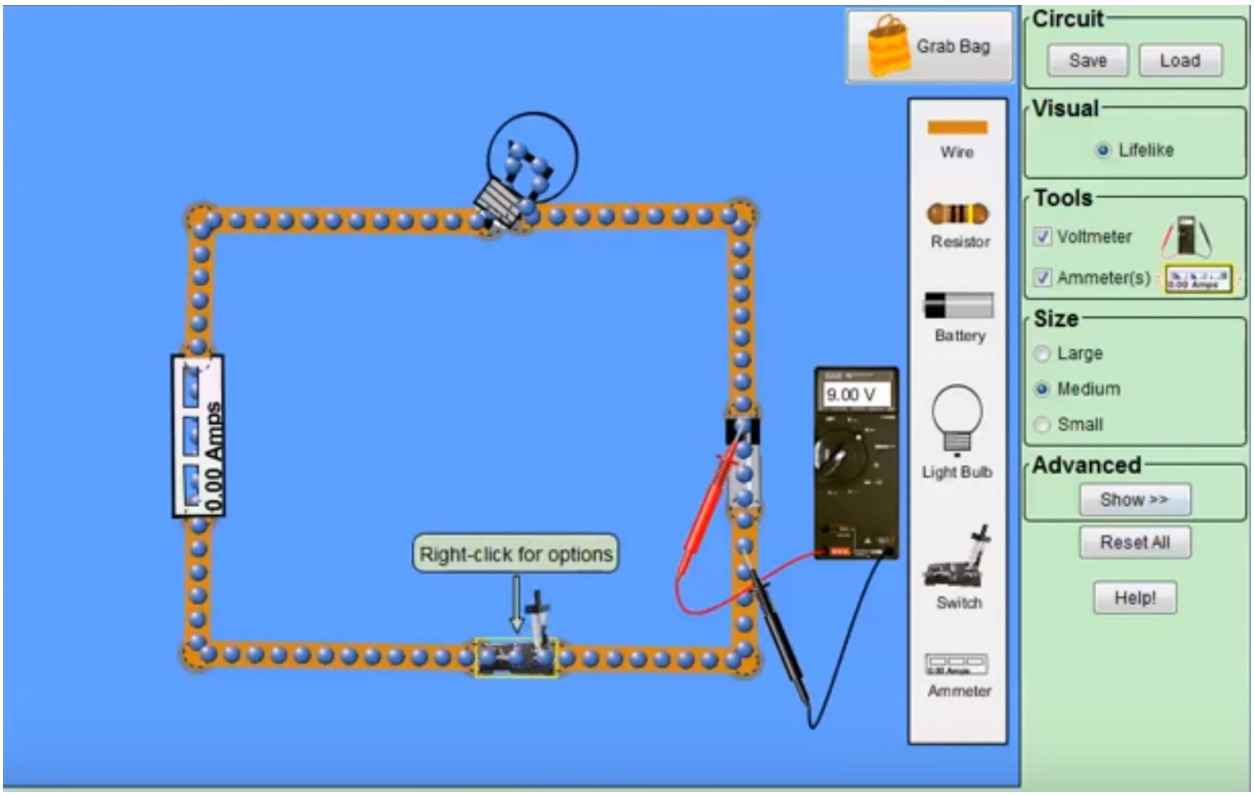


Figure 10. Screenshot of the PhET interactive simulations circuit construction kit. Adapted from "PhET Interactive Simulations: Circuit Construction Kit (Direct Current Only)", by PhET Interactive Simulation, University of Colorado Boulder. Reprinted with permission.

During open-ended simulations, students may use more resources than typically supplied during a physical hands-on laboratory activity. The affordance of open-endedness – having multiple and iterative methods to solve a problem - may be found to create a simulation learning environment that encourages students to "stretch" their scientific inquiry skills (Dwyer & Lopez, 2001). Scientific inquiry is considered fundamental to science education because it focuses on how scientists acquire knowledge and skills in the real-world (National Research Council, 1996, 2014). Thus, an important goal is to develop assessment tools that can consistently measure scientific inquiry in a way that approximates real-world practice and provides students with

feedback to master the needed knowledge and skills. In other words, there needs to be a greater focus on the development of CSSLs that measure scientific inquiry and allow for open-ended solutions so that students can explore different methods to approach a problem.

Overall, CSSLs may help contribute towards further assessment research by providing educators with a new and dynamic format to assess science knowledge and skills. The digital format of these assessments allows students to work with resources that would be too costly or dangerous to use in a real-life experiment. Additionally, this format provides a laboratory environment that allows students to repeat their experiments so that the results from previous trials may inform later iterations. This format does hold potential in terms of being a flexible, dynamic learning and assessment tool for measuring process knowledge and skills. It is, therefore, important to continue investigating the best ways to implement and administer CSSL assessments so that their benefits may be maximized.

References

Ackroyd, J. E., Anderson, M., Berg, C., Martin, B. E., McGuire, B. L. S., Sosnowski, C., … Wolfe, E. (2007). *Physics*. Toronto, ON: Pearson Education Canada.

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67*(4), 1012-1028. doi: 10.1111/j.1741-3737.2005.00191.x

Alberta Education. (1996). Science (elementary). Edmonton, Alberta: Alberta Education. Retrieved from http://www.education.alberta.ca/media/654825/elemsci.pdf

Alberta Education. (2014a). *Physics 20-30*. Alberta: Alberta Education. Retrieved from http://www.learnalberta.ca/ProgramOfStudy.aspx?lang=en&ProgramId=187927#

Alberta Education. (2014b). Science grades 7-8-9. Edmonton, Alberta: Alberta Education. Retrieved from https://education.alberta.ca/media/3069389/pos_science_7_9.pdf

Alberta's Commission on Learning. (2003). *Every child learns, every child succeeds: Report and recommendations*. Retrieved from https://archive.education.alberta.ca/media/413413/commissionreport.pdf

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Babenko, O., & Rogers, W. T., (2014). Comparison and properties of correlational and agreement methods for determining whether or not to report subtest scores. *International Journal of Learning, Teaching and Educational Research, 4*(1), 61-74. Retrieved from http://ijlter.org/index.php/ijlter/article/view/50/pdf

Baker, E. L., Niemi, D., & Chung, G. K. W. K. (2008). Simulations and the transfer of problem-solving knowledge and skills. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil

(Eds.), *Assessment of problem solving using simulations* (pp. 1-17). New York: Taylor & Francis Group.

Bandalos, D. L., Yates, K., & Thorndike-Christ, T. (2005). Effects of math self-concept, perceived self-efficacy, and attributions for failure and success on test anxiety. *Journal of Educational Psychology 87*(4), 611-623. doi: 10.1037/0022-0663.87.4.611

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play. *Educational Researcher 39*(7), 525–536. doi: 10.3102/0013189X10386593

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world* (CRESST Report 778). Retrieved from http://files.eric.ed.gov/fulltext/ED520431.pdf

Bennett, R.E., Persky, H., Weiss, A.R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project* (NCES 2007–466). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London, United Kingdom: School of Education, King's College.

Bybee, R. W. (2010). What is STEM education? *Science, 329*(5995), 996. doi: 10.1126/science.1194998

Chaby, L. E., Sheriff, M. J., Hirrlinger, A. M., & Braithwaite, V. A. (2016). Can we understand how development stress enhanced performance under future threat with the Yerkes-Dodson law? *Communicative and Integrative Biology, 8*(3), e1029689. doi: 10.1080/19420889.2015.1029689

Chalmers, A. F. (1999). *What is this thing called science?* Indianapolis, IN: Hackett Publishing Company.

Cheronis, N. D. (1962). The philosophy of laboratory instruction. *Journal of Chemical Education,* 39(2), 102-106. doi: 10.1021/ed039p102

Chittleborough, G. D., Mocerino, M., & Treagust, D. F. (2007). Achieving greater feedback and flexibility using online pre-laboratory exercises with non-major chemistry students. *Journal of Chemical Education 84*(5), 884-888. doi: 10.1021/ed084p884

Chu, M-W. (2010, August). *Exploring science curriculum emphases in relation to the Alberta physics program-of-study.* Unpublished Master's Thesis: University of Alberta.

Chu, M-W., & Leighton, J. P. (2016). Using errors to enhance learning and feedback in computer programming. In S. Tettegah & M. P. McCreery (Eds.), *Emotions, technology, and learning* (pp.89-117). London Wall, London: Elsevier Incorporated.

Chu, M-W., Guo, Q., & Leighton, J. P. (2013). Students' interpersonal trust and attitudes towards standardized tests: Exploring affective variables related to student assessment. *Assessment in Education: Principles, Policy & Practice, 21*(2), 167-192. doi: 10.1080/0969594X.2013.844094

Cisco, Intel, and Microsoft (2008). Transforming education: Assessing and teaching the skills needed in the 21st century. Retrieved from http://listar.hi.is/pipermail/feki/attachments/20081013/f8c88285/callforaction-0001.pdf

City of Calgary (2015). *Community Profiles*. Retrieved from http://www.calgary.ca/CSPS/CNS/Pages/Research-and-strategy/Community-profiles/Community-Profiles.aspx

Coller, B. D., & Scott, M. J. (2009). Effectiveness of using a video game to teach a course in mechanical engineering. *Computers and Education 53*(3), 900–912. doi:10.1016/j.compedu.2009.05.012.

Council of Canadian Academies. (2015). *Some assembly required: STEM skills and Canada's economic productivity: The expert panel on STEM skills for the future*. Ottawa (ON): The Expert Panel on STEM Skills for the Future, Council of Canadian Academies. Retrieved from http://scienceadvice.ca/uploads/ENG/AssessmentsPublicationsNewsReleases/STEM/STEMFullReportEn.pdf

Council of Ministers of Education Canada [CMEC]. (2016). *Measuring up: Canadian results of the OECD PISA study: The performance of Canada's youth in science, reading, and mathematics.* Toronto, ON: Author. Retrieved from http://cmec.ca/Publications/Lists/Publications/Attachments/365/Book_PISA2015_EN_Dec5.pdf

De Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 179-201. doi: 10.3102/00346543068002179

Domin, D. S. (1999). A review of laboratory instruction styles. *Journal of Chemical Education, 76*(4), 543-547. doi: 10.1021/ed076p543

Doran, R. L., Boorman, J., Chan, F., & Hejaily, N. (1993). Alternative assessment of high school laboratory skills. *Journal of Research in Science Teaching, 30*(9), 1121–1131. doi: 10.1002/tea.3660300909

Duncan, T. G. & McKeachie, W. J. (2005) The making of the motivated strategies for learning questionnaire. *Educational Psychologist, 40*(2), 117-128. doi: 10.1207/s15326985ep4002_6

Dwyer, W. M., & Lopez, V. E. (2001). *Simulations in the learning cycle: A case study involving "exploring the Nardoo*". Paper presented at the National Educational Computing Conference, "Building on the Future", Chicago, IL. Retrieved from http://files.eric.ed.gov/fulltext/ED462932.pdf

Empirical Games (2013). *Physics playground: A game to teach qualitative physics*. Retrieved from http://www.empiricalgames.org/games/

Firestein, S. (2016) *Failure: Why science is so successful*. New York, NY: Oxford University Press.

Fredricks, J. A., Blumenfeld, P., Friedel, J., & Paris, A. (2005). School engagement. In K. A. Moore & L. Lippman (Eds.), *What do children need to flourish? Conceptualizing and measuring indicators of positive development* (pp. 305-321). New York, NY: Springer Science and Business Media.

Gobert, J., O'Dwyer, L., Horwitz, P., Buckley, B., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' epistemologies of models and conceptual learning in three science domains: biology, physics, & chemistry. *International Journal of Science Education, 33*(5), 653-684. doi: 10.1080/09500691003720671

Gravetter, F. J., & Wallnau, L. B. (2009). *Statistics for the behavioral sciences*. Belmont, CA: Wadsworth.

Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress and Coping, 10*(3), 219-244. doi: 10.1145/1132960.1132961

Hodson, D. (1998) *Teaching and learning science: Towards a personalized approach.* Maidenhead, Berkshire: Open University Press.

Hodson, D. (2003). Time for action: Science education for an alternative future. *International Journal of Science Education, 25*(6), 645-670. doi: 10.1080/09500690305021

Hofstein, A., & Lunetta, V. N. (2003). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, 88(1), 28-54. doi: 10.1002/sce.10106

Institute of Education Sciences (2006). *National assessment of educational progress: Problem solving in technology-rich environments*. Retrieved from http://nces.ed.gov/nationsreportcard/studies/tba/tre/

Ioannidou, A., Repenning, A., Webb, D., Keyser, D., Luhn, L., & Daetwyler, C. (2010). Mr. Vetro: A collective simulation for teaching health science. *International Journal of Computer-Supported Collaborative Learning, 5*(2), 141-166. doi: 10.1007/s11412-010-9082-8

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnstone, A. H., Watt, A., & Zaman, T. U. (1998). The students' attitude and cognition change to a physics laboratory. *Physics Education, (33)*1, 22-29. doi: 10.1088/0031-9120/33/1/016

Leighton, J. P., & Bustos Gomez, M. C. (Under review). *A pedagogical alliance for trust, wellbeing and the identification of errors for learning and formative assessment.* Paper submitted for publication.

Leighton, J. P., Chu, M-W., & Seitz, P. (2013). Cognitive diagnostic assessment and the learning errors and formative feedback (LEAFF) model. In R. Lissitz (Ed.), *Informing the practice*

of teaching using formative and interim assessment: A systems approach (pp. 183-207). Information Age Publishing.

Lytton, H., & Pyryt, M. (1998). Predictors of achievement in basic skills: A Canadian effective schools study. Canadian Society for Study in Education, 23(3), 281-301. Retrieved from http://www.jstor.org/stable/1585940

Ma, J., & Nickerson, J. V. (2006). Hands-on, simulated, and remote laboratories: A comparative literature review. *ACM Computing Surveys, 38*(3), Article 7. doi:10.1145/1132960.1132961

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., … Urdan, T. (2000). *Manual for the patterns of adaptive learning scales*. Ann Arbor, MI: University of Michigan Press.

Mislevy, R. J., Almond, R. G., & Lukas, J. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). Retrieved from Educational Testing Service website: https://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., … John, M. (2014). Psychometric considerations in game-based assessment. GlassLab: Institute of play. Retrieved from http://www.instituteofplay.org/wp-content/uploads/2014/02/GlassLab_GBA1_WhitePaperFull.pdf Mizuko, I. (2009). *Engineering play: A cultural history of children's software.* Cambridge, MA: The MIT Press.

National Assessment of Educational Progress [NAEP]. (2007). *NAEP technology-rich environment simulation scenario [Computer software].* Retrieved from https://nces.ed.gov/nationsreportcard/studies/tba/tre/sim-description.aspx

National Research Council. (1996). *National science education standards*. National Committee for Science Education Standards and Assessment. National Committee on Science Education Standards and Assessment, Board on Science Education, Division of Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press. Retrieved from http://www.nap.edu/catalog/4962/national-science-education-standards

National Research Council. (2006). *America's lab report: Investigations in high school science*. Committee on High School Science Laboratories: Role and Vision, In S. R. Singer, M. L. Hilton, and H. A. Schweingruber, (Eds.). Board on Science Education, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog/11311.html

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=13165

National Research Council. (2014). *Developing assessments for the next generation science standards*. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J. W. Pellegrino, M. R. Wilson, J. A. Koenig, and A. S. Beatty (Eds.), Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=18409

National Science Teachers Association. (2004, October). *National science teachers association NSTA position statement: Scientific inquiry.* Retrieved from http://www.nsta.org/docs/PositionStatement_ScientificInquiry.pdf

Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by states.* Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=18290

O'Connor, K. (2010). *A repair kit for grading: 15 fixes for broken grades (2nd ed.)*. Boston, MA: Pearson.

Office of Assessment Services, Northern Illinois University (n.d.). *Assessment term glossary*. Retrieved from http://www.niu.edu/assessment/Resources/Assessment_Glossary.htm

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103(2),* 241-262.

Organization for Economic Co-operation and Development [OECD]. (2016). PISA 2015 results (volume 1): Excellence and equity in education. Retrieved from http://www.oecd.org/pisa/pisa-2015-results-volume-i-9789264266490-en.htm

Nichols, P.D. (2010). What is a learning progression? Test, Measurement & Research Services Bulletin. Issue 12. Pearson Education. Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/Bulletin_12.pdf?WT.mc_id=TMRS_Bulletin_12_What_is_a_learning_progression

Perkins, K. K., Loeblein, P. J., & Dessau, K. L. (2010) Sims for science: Powerful tools to support inquire-based teaching. *Science Teacher, 77*(7), 46-51. Retrieved from http://static.nsta.org/files/tst1010_46.pdf

PhET. (2015). *PhET interactive simulations: Research*. Retrieved from https://phet.colorado.edu/en/research

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation, 7*(4), 353-383. doi: 10.1076/edre.7.4.353.8937

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1991). Reliability and predictive validity of the motivational strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement, 53,* 801-813. doi: 10.1177/0013164493053003024

Popham, W. J. (2008). *Transformative Assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, Alberta: Joint Advisory Committee. Retrieved from http://cdn.aac.ab.ca/wp-content/uploads/2015/10/eng_principles.pdf

Quellmalz, E. S., Timms, M. J., & Buckley, B. C. (2009). Using science simulations to support powerful formative assessments of complex science learning. Retrieved from http://simscientists.org/downloads/Quellmalz_Formative_Assessment.pdf

Reid, N., & Shah, I. (2007). The role of laboratory work in university chemistry. *Chemistry Education Research and Practice, 8*(2), 172-185. Retrieved from http://www.rsc.org/images/Reid%20paper%20final_tcm18-85040.pdf

Rosen, K. R. (2008). The history of medical simulations. *Journal of Critical Care, 23*, 157-166. doi: 10.1016/j.jcrc.2007.12.004

Rowe, E., Asbell-Clarke, J., & Baker, R. S. (2015). Serious games analytics to measure implicit science learning. In C.S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 343-360). Springer International Publishing. doi: 10.1007/978-3-319-05834-4_15

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment, 8*(4). Retrieved from http://edgaps.org/gaps/wp-content/uploads/rupp2010.pdf

Rutherford, F.J., & Ahlgren, A. (1990). *Science for all Americans.* New York: Oxford University Press. Retrieved from http://www.project2061.org/publications/sfaa/online/sfaatoc.htm

Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers and Education, 58*(1), 136-153. Doi: 10.1016/j.compedu.2011.07.017

Sahin, S. (2006). Computer simulations in science education: Implications for distance education. *Turkish Online Journal of Distance Education, 7*(4): 132-146. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.8977&rep=rep1&type=pdf

Sawyer, R. K. (2006). *Explaining creativity: The science of human innovation*. New York: Oxford University Press.

Scalese, R. J., Obeso, V. T., & Issenberg, B. (2008). Simulation technology for skills training and competency assessment in medical education. *Journal of General Internal Medicine, 23*(1), 46-49. doi: 10.1007/s11606-007-0283-4

Scalise, K. (2009, June). *Computer-based assessment: "Intermediate constraint" questions and tasks for technology platforms*. Retrieved from http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html

Scalise, K., Timms, M., Moorjani, A., Clark, L., Holtermann, K., & Irvin, P. S. (2011), Student learning in science simulations: Design features that promote learning gains. *Journal of Research in Science Teaching, 48*(9), 1050–1078. Doi: 10.1002/tea.20437

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.; pp. 307-353). Westport, CT: Praeger Publishers.

Shute, V. J. (1993). A comparison of learning environments: All that glitters... In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 47-75). Hillsdale, NJ: Erlbaum.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. doi: 10.3102/0034654307313795

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers. Retrieved from http://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf

Shute, V. J., & Becker, B. J. (2010). Prelude: Issues and assessment for the 21st century. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs* (pp. 1-11). New York, NY: Springer-Verlag. Retrieved from http://myweb.fsu.edu/vshute/pdf/prelude.pdf

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: Massachusetts Institute of Technology Press. Retrieved from http://myweb.fsu.edu/vshute/pdf/white.pdf

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. *International*

Journal of Artificial Intelligence and Education, 18(4), 289-316. Retrieved from http://myweb.fsu.edu/vshute/pdf/shute%202008_a.pdf

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education, 80*, 58-67. doi: 10.1016/j.compedu.2014.08.013

Smith, A. M., Floerke, V. A., Thomas, A. K. (2016). Retrieval proactive protects memory against acute stress. *Science, 356*(6315), 1046-1048. doi: 10.1126/science.aah5067

Squire, K., & Patterson, N., (2010). *Games and simulations in informal science education* (Working Paper No. 2010-14). Retrieved from http://files.eric.ed.gov/fulltext/ED514369.pdf

Statistics Canada. (2008). Average scores and confidence intervals for provinces and countries: Combined science. Retrieved from http://www.statcan.gc.ca/pub/81-590-x/2007001/charts/5002606-eng.htm

Statistics Canada. (2010). *Measuring up: Canadian results from the OECD PISA study*: *The performance of Canada's youth in reading, mathematics and science PISA 2009 first results for Canadians aged 15*. Ottawa, ON: Ministry of Industry. Retrieved from http://www.statcan.gc.ca/pub/81-590-x/81-590-x2010001-eng.htm

Stieff, M., & Wilensky, U. (2003). Connected chemistry—Incorporating interactive simulations into the chemistry classroom. *Journal of Science Education and Technology, 12*(3), 285-302. doi: 10.1023/A:1025085023936

Strauss, R., & Kinzie, M. B. (1994). Student achievement and attitudes in a pilot study comparing an interactive videodisc simulation to conventional dissection. *American Biology Teacher, 56*(7), 398-402. doi: 10.2307/4449869

Strogatz, S. (2007) *The end of insight.* In J. Brockman (Ed.), *What is your dangerous idea? Today's leading thinkers on the unthinkable* (pp. 130-131). New York: Harper Perennial.

Supasorn, S., Suits, J. P., Jones, L. L., & Vibuljan, S. (2008). Impact of a pre-laboratory computer simulation of organic extraction on comprehension and attitudes of undergraduate chemistry students. *Chemistry Education Research and Practice, 9*(2), 169-181. doi: 10.1039/b806234j

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson/Allyn & Bacon.

Tamir, P., Nussinovitz, R., & Friedler, Y. (1982). The design and use of practical tests assessment inventory. *Journal of Biological Education, 16*(1), 42–50. doi: 10.1080/00219266.1982.9654417

Tsupros, N., Kohler, R., & Hallinen, J. (2009). *STEM education in southwestern Pennsylvania: Report of a project to identify the missing components,* Intermediate Unit 1 Center for STEM Education and Leonard Gelfand Center for Service Learning and Outreach at Carnegie Mellon University, Pennsylvania. Retrieved from http://www.cmu.edu/gelfand/documents/stem-survey-report-cmu-iu1.pdf

U.S. Congress, Office of Technology Assessment. (1992, February). Testing in American schools: Asking the right questions. (OTA-SET-519). Washington, DC: U.S. Government Printing Office. Retrieved from http://www.fas.org/ota/reports/9236.pdf

Waternaux, C.M. (1976). Asymptotic distribution of the sample roots for a nonnormal population. *Biometrika, 63(3),* 639-645.

Waternaux, C. M. (1984). Principal components in the nonnormal case: The test of equality of Q roots. *Journal of Multivariate Analysis, 14*, 323-335.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.

Wilkinson, J. W., & Ward, M. (1997). The purpose and perceived effectiveness of laboratory work in secondary schools. *Australian Science Teachers' Journal, 43*(2) 49–55.

Wilson, M., & Sloane, K. (2000) From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208. doi: 10.1207/S15324818AME1302_4

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. New York: Routledge Academic.

Zeidner M. (1998). *Test anxiety: The state of the art*. New York, NY: Plenum

Appendix A

Student Information Letter and Consent Form

Project Title: Computer Simulated Science Laboratory Assessment
Principal Investigator: Man-Wai Chu & Dr. Jacqueline Leighton

What is a research study?

- A research study is a way to find out new information about something. Children do not need to be in a research study if they don't want to.

Why are you being asked to be part of this research study?

- You are being asked to take part in this research study because we are trying to learn more about how students learn science lab skills using computer simulated science labs. We are asking you to be in the study because your science class has been chosen to help with this study. About 240 children will be in this study.

If you join the study what will happen to you?

- We want to tell you about some things that will happen to you if you are in this study.
- You will be in the study for one science class period.
- We will ask you to work through a simulated science lab on a computer, short assignment/quiz, and survey

Will any part of the study hurt? NO

Will the study help you? The results of this study will help you better understand your level of knowledge and skills in a science lab.

Will the study help others?

- Understanding students' use of computer simulated labs and learning outcomes will help educators maximize these classroom tools more effectively and provide students with activities and assessments that are good measures of students' knowledge and skills.

Do your parents know about this study?

- This study was explained to your parents and they said that we could ask you if you want to be in it.

Who will see the information collected about you?

- The information collected about you during this study will be kept safely locked up. Nobody will know it except the people doing the research.
- The study information about you will be given to your teachers. The researchers will not tell your friends or anyone else.

What do you get for being in the study?

- You will gain insightful information about your lab knowledge and skills so that you may improve areas of weaknesses.

Do you have to be in the study?

- You do not have to be in the study. No one will be upset if you don't want to do this study. If you don't want to be in this study, you just have to tell us. It's up to you.

What if you have any questions?

- You can ask any questions that you may have about the study. If you have a question later that you didn't think of now, either you can call or have your parents call 780-996-5216.

What choices do you have if you say no to this study?

- You will complete an activity pre-approved by your teacher while your classmates work on the computer simulated science lab.

Other information about the study.

- If you decide to be in the study, please write your name below.
- You can change your mind and stop being part of it at any time. All you have to do is tell the person in charge. It's okay. The researchers and your parents won't be upset.
- You will be given a copy of this paper to keep.

☐ Yes, I will be in this research study. ☐ No, I don't want to do this.

_____	_____	_____
Child's Name	Signature of the Child	Date
_____	_____	_____
Person Obtaining Assent	Signature	Date

_____ _____

Student Code (made up) Computer Simulated Laboratory Code

Appendix B

Parent Information Letter and Consent Form

Project Title: Computer Simulated Science Laboratory Assessment

Your child is invited to take part in a research study that looks at how students learn science lab skills using computer simulated science labs. Knowing how students learn science lab skills will help us to better make simulated science labs that can be used to tell teachers what a student knows and can do, and also where he or she may need some help.

Teachers have helped us identify the possible ways students may learn science lab skills. But, we are interested in the level of science lab skills learned through computer simulated science labs. To find the level of science lab skills learned, we are asking students to work through a computer simulated science lab. The computer will record their actions throughout the simulated science lab and analyze those actions for an understanding of science lab skills.

We are asking you to give permission for your child to participate in our study. The results of this research will be shared with assessment specialists to help design better educational tests. The results will also be shared with other educational researchers through papers or professional conferences. All individual student information will be kept confidential and only group results will be shared.

Methods:

The simulated science lab has been developed for Grade 8 science students in the topics of Flight and Mix & Flow of Matter. Students will work through a simulated science lab on a computer to answer three questions in one class period. A computer will record students' actions, such as which buttons were clicked, on the computer while they work through the simulated science lab.

Understanding of participation and consent:

I understand that:

- ✓ my child will be asked to complete a simulated science lab.
- ✓ my child's computer actions, such as which buttons were clicked, will be recorded.
- ✓ I can withdraw my child's participation in the study at any time until one month after the simulated science lab has been completed.
- ✓ there are no negative consequences for not participating in the study.

Two copies of this form have been provided. Please indicate in the boxes below if you give permission for your child to participate in this study and then sign your name. Please return this copy to your child's science teacher. The other copy should be kept for your own records.

Consent:

Please indicate whether or not your child can participate in this study.

____ I give permission for my child, _____, to participate in this
(Please print child's name)
study, and that my child may withdraw from the study at any time without penalty.

____ I do not give permission for my child, _____, to participate in
(Please print child's name)
this study.

_____ _____

Signature (Parent/Guardian) Date

If you have any questions or concerns please do not hesitate to contact Dr. Jacqueline Leighton (Chair – Department of Educational Psychology) or the researcher, Man-Wai Chu:

Man-Wai Chu
E-mail: manwai@ualberta.ca
Phone: 780-996-5216

Dr. Jacqueline Leighton
E-mail: Jacqueline.Leighton@ualberta.ca
Phone: 780-420-1167

The plan for this study has been reviewed for its adherence to ethical guidelines and approved by the Faculties of Education, Extension and Augustana Research Ethics Board (EEA REB) at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Chair of the EEA REB at 780-492-3751.

Appendix C

Teacher Information Letter and Consent Form

Project Title: Computer Simulated Science Laboratory Assessment

You are invited to take part in a research study that looks at how students learn science lab skills using computer simulated science labs. Knowing how students learn science lab skills will help us to better make simulated science labs that can be used to tell teachers what a student knows and can do, and also where he or she may need some help.

Teachers have helped us identify the possible ways students may learn science lab skills. But, we are interested in the level of science lab skills learned through computer simulated science labs. To find the level of science lab skills learned, we are asking students to work through a computer simulated science lab. The computer will record their actions throughout the simulated science lab and analyze those actions for an understanding of science lab skills.

We are asking you to give permission for your class to participate in our study. Parental and student consent will also be obtained before the start of the computer simulated science labs. The results of this research will be shared with assessment specialists to help design better educational tests. The results will also be shared with other educational researchers through papers or professional conferences. All individual student information will be kept confidential and only group results will be shared.

Methods:

The simulated science lab has been developed for Grade 8 science students in the topics of Flight and Mix & Flow of Matter. Students will work through a series of assessments including a simulated science lab on a computer to answer three questions over two class periods. A computer will record students' actions, such as which buttons were clicked, on the computer while they work through the simulated science lab. A survey will also be administered to capture students' opinions regarding the simulated science lab.

Understanding of participation and consent:

I understand that:

- ✓ my science class will be asked to complete a simulated science lab and fill out a survey.
- ✓ I will need to provide the researcher with 2 class periods to administer the simulated science lab and the possibility of assigning additional assignments that have been prepared by the researcher.
- ✓ my students' computer actions, such as which buttons were clicked, will be recorded.
- ✓ there are no negative consequences for not participating in the study.

Two copies of this form have been provided. Please indicate in the boxes below if you give permission for you and your class to participate in this study and then sign your name. The other copy should be kept for your own records.

Consent:

Please indicate whether or not you and your class can participate in this study.

____ I give permission for myself and my class, _____, to
(Please print teacher's name)
participate in this study, and that my students may withdraw from the study at any time without penalty.

____ I do not give permission for myself and my class, _____, to
(Please print teacher's name)
participate in this study, and that my students may withdraw from the study at any time without penalty.

_____ _____

Signature (Teacher) Date

If you have any questions or concerns please do not hesitate to contact Dr. Jacqueline Leighton (Chair – Department of Educational Psychology) or the researcher, Man-Wai Chu:

Man-Wai Chu
E-mail: manwai@ualberta.ca
Phone: 780-996-5216

Dr. Jacqueline Leighton
E-mail: Jacqueline.Leighton@ualberta.ca
Phone: 780-420-1167

The plan for this study has been reviewed for its adherence to ethical guidelines and approved by the Faculties of Education, Extension and Augustana Research Ethics Board (EEA REB) at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Chair of the EEA REB at 780-492-3751.

Appendix D

Script for Obtaining Verbal Consent

Hi Mrs./Mr. _____, this is Mrs./Ms./Mr. _____. I am _____'s science teacher.

I am calling because your child is invited to take part in a research study that looks at how students learn science lab skills using computer simulated science labs. Knowing how students learn science lab skills will help us to better make simulated science labs that can be used to tell teachers what a student knows and can do, and also where he or she may need some help.

Teachers have helped us identify the possible ways students may learn science lab skills. But, we are interested in the level of science lab skills learned through computer simulated science labs. To find the level of science lab skills learned, we are asking students to work through a computer simulated science lab. The computer will record their actions throughout the simulated science lab and analyze those actions for an understanding of science lab skills.

We are asking you to give permission for your child to participate in the study. The results of this research will be shared with assessment specialists to help design better educational tests. The results will also be shared with other educational researchers through papers or professional conferences. All individual student information will be kept confidential and only group results will be shared.

The simulated science lab used in this study has been developed for Grade 8 science students in the topics of Flight and Mix & Flow of Matter. Students will work through a simulated science lab on a computer to answer three questions in one class period. A computer will record students' actions, such as which buttons were clicked, on the computer while they

work through the simulated science lab. A survey will also be administered to capture students' opinions regarding the simulated science lab.

I hope that you understand and will grant consent to allowing the researcher to:

- *Ask your child to complete a simulated science lab and fill out a survey.*
- *Record your child's computer actions, such as which buttons were clicked.*

The participation is voluntary, and there are no negative consequences for not participating in the study. You can withdraw your child's participation in the study at any time until one month after the simulated science lab has been completed and, again, there are no negative consequences for not participating in the study. Do you have any question? If you have any additional question at a later time please feel free to call me at the school or the researcher Ms. Man-Wai Chu (780-996-5216).

Do you provide verbal consent for your child to participate in this study?

Thank you for your time.

Appendix E

Pre-Laboratory Activity

You have been hired as a science intern at the local weather station. The weather station uses a weather balloon to determine the weather and have assigned you the problem of finding different aspects (mass hanging from balloon and volume of helium in balloon) that can affect a balloon's altitude (how high a balloon can raise) when flying. Since this is a science internship, the station is hoping you can use scientific lab principles to solve the problem. Use your knowledge of scientific lab processes to work on the station's problem.

1. What is your hypothesis of the relationship between the amount of mass hanging from a balloon and the altitude of the balloon?

_____

_____

_____

2. The weather station is giving you $500 to buy lab equipment to test your hypothesis. Design an experiment which will test your hypothesis. Include the tools you will use to record and organize your data.

a) List the materials needed for your experiment: _____

_____

_____

b) Steps of your experimental design: _____

_____

_____

_____

c) What tools will you use to record your data:_____

_____

d) How will you organize your data: _____

_____

3. Indicate the different types of variables of your experiment.

Manipulated variable: _____

Responding variable: _____

Controlled variable(s): _____

_____

Appendix F

Learning Error Intervention

Slide 1

Computer Simulated Laboratory Assessment

Man-Wai Chu, PhD Candidate

Dr. Jacqueline Leighton, PhD

Slide 2

What is our rationale for this study?

Well, as you might have experienced in past classes, learning is a rewarding experience but it also can be risky.

Learning takes us from a state of NOT KNOWING to a state of COMING TO KNOW, and this complex process involves several elements such as making mistakes.

Making mistakes is part of learning. Actually, psychologists tell us that –in most cases- making mistakes help us learn better.

Slide 3

Why is that?

Well, mistakes help our brain clearly separate what is correct and incorrect. In the process of learning, being able to identify mistakes, where they can happen and talking about them can help us learn better.

You may recall an experience when you were learning something and made a mistake (or more than one) and this helped you learn that knowledge or skill

really well; for example, when you were learning to tie your shoes or play a new video game.

Have you ever experienced this?

Slide 4

As I said, this class is about scientific laboratory skills and learning these skills involves making mistakes. Why is that?

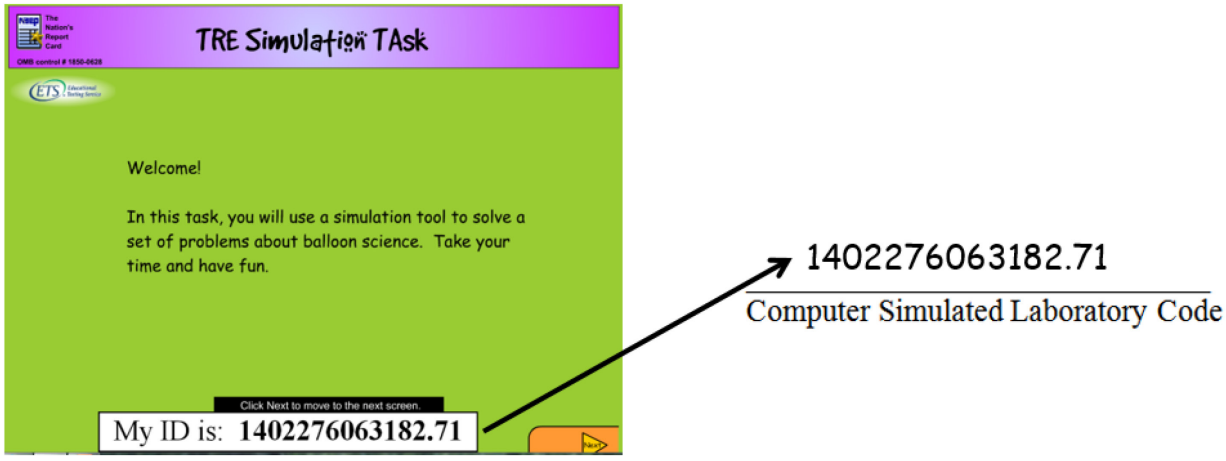
Well, in learning scientific laboratory skills there are many steps and concepts which makes it is very easy to get confused and make a mistake. So, it is very important to recognize the presence or potential to make mistakes. As you work through this simulated laboratory, please feel free to make mistakes so that you can move from a state of not knowing to a state of knowing – learn. During the simulation you may experiment and make mistakes, but before you hit the final 'next' button ensure that the answers you have there are what you hope to submit

Slide 5

Please start the simulation by going to the following site:

http://tresim.educ.ualberta.ca/

Please write the Computer Simulated Laboratory Code on your consent form.



Appendix G

Pre-Intervention Survey Measure

Pre-Intervention Survey Measure Items

Thank you very much for participating in this study on Computer Simulated Science Laboratory Assessment. Please ensure you have completed the consent form.
Please write your student code below:

_____

Survey 1: Patterns of Adaptive Learning Scale

Using the scale below and **thinking about your classes in general**, please rate the following items. Please answer all items, even if you are not sure. Please select only a single rating for each item.

	Not at all true 1	**2**	**Somewhat true 3**	**4**	**Very true 5**
1. It's important to me that I learn a lot of new concepts this year.					
2. One of my goals in class is to learn as much as I can.					
3. One of my goals is to master a lot of new skills this year.					
4. It's important to me that I thoroughly understand my class work.					
5. It's important to me that I improve my skills this year.					
6. It's important to me that other students in my class think I am good at my class work.					
7. One of my goals is to show others that I'm good at my class work.					
8. One of my goals is to show others that class work is easy for me.					
9. One of my goals is to look smart in comparison to the other students in my class.					
10. It's important to me that I look smart compared to others in my class.					
11. It's important to me that I don't look stupid in class.					
12. One of my goals is to keep others from thinking I'm not smart in class.					
13. It's important to me that my teacher doesn't think that I know less than others in class.					

14. One of my goals in class is to avoid looking like I have trouble doing the work.					

Survey 2: Motivated Strategies for Learning Questionnaire

Using the scale below and **thinking about your classes in general**, please rate the following items. Please answer all items, even if you are not sure. Please select only a single rating for each item.

	Not at all true of me 1	**2**	**3**	**4**	**5**	**6**	**Very true of me 7**
1. I prefer class material that really challenges me so I can learn new things.							
2. I'm certain I can understand the most difficult material presented in the textbook for the class.							
3. The most satisfying thing for me in class is trying to understand the content as thoroughly as possible.							
4. When I have the opportunity in class, I choose topics of class projects that I can learn from even if they don't guarantee a good grade.							
5. Getting a good grade in the class is the most satisfying thing for me right now.							
6. The most important thing for me right now is improving my overall school average, so my main concern in the class is getting a good grade.							
7. If I can, I want to get better grades in the class than most of the other students.							
8. I want to do well in the class because it is important to show my ability to my family, friends, or others.							
9. I often find myself questioning things I hear or read in class to decide if I find them convincing.							
10. When a theory, interpretation, or conclusion is presented in class or in the textbook, I try to decide if there is good supporting evidence.							

11. I treat the class material as a starting point and try to develop my own ideas about it.							
12. I try to play around with ideas of my own related to what I am learning in class.							
13. Whenever I read or hear an assertion or conclusion in class, I think about possible alternatives.							

Survey 3: NAEP TRESim Background Questionnaire

Using the scale below and **thinking about your science classes in particular**, please rate the following items. Please answer all items, even if you are not sure. Please select only a single rating for each item.

How often did you do the following:	**Never** **1**	**Sometimes, but less than once a month** **2**	**Once a month or more** **3**
1. Design your own science experiment or investigation			
2. Carry out the science experiment or investigation you designed			
3. Write up results of the experiment or investigation you designed			
4. Talk to the class about the results of your experiment or investigation			
5. Collect data using lab equipment that interfaces with computers (for example, probes)			
6. Download data and related information from the Internet			
7. Analyze data using the computer			
8. Use the Internet to exchange information with other students or scientists about science experiments or investigations			
9. Use computer simulations to perform experiments or explore science topics			

Prior-Knowledge Questions

Please circle the correct answer.

1. Which of the following is the best example of the concept of mass?

A. length of a piece of material
B. amount of material in an object
C. amount of space that a liquid takes up
D. energy it takes a person to carry an object

2. Which statement best describes what happens to a specific amount of gas when it is moved from a larger to a smaller closed container?

A. mass of the gas decreases

B. density of the gas increases
C. volume of the gas increases
D. temperature of the gas decreases

3. A rubber gas balloon can hold 10 cubic feet of helium. Ellen puts 5 cubic feet of helium inside the balloon, so its starting volume is 5 cubic feet. The balloon rises and expands. When the balloon stops rising, its final volume is 10 cubic feet. Why did the balloon volume change from start to finish? As the balloon rises:

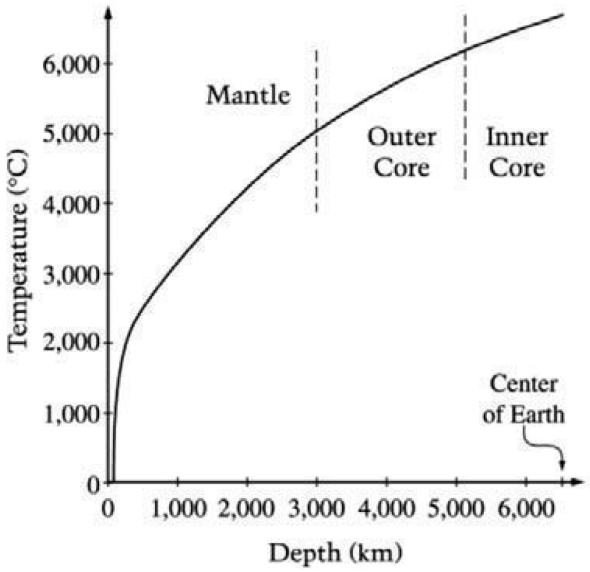
A. decreasing air pressure allows the amount of helium gas inside the balloon to increase
B. increasing air pressure makes the helium gas inside the balloon denser and therefore heavier
C. increasing air pressure makes the helium gas inside the balloon less dense so it expands
D. decreasing air pressure allows the helium inside the balloon to expand and push out the sides of the balloon

4. Brad thinks that water will evaporate at different rates depending on the temperature of a room. If he wants to do an experiment to test his idea, what would be the best experimental set up? Put equal amounts of water at the same temperature in bowls of:

A. equal size, each in a different room with each room having the same temperature but different humidity
B. equal size, each in a different room with each room having a different temperature but the same humidity
C. different sizes, each in a different room with each room having the same temperature and the same humidity
D. different sizes, each in a different room with each room having a different temperature and a different humidity

The graph below shows the change in temperature inside the Earth as the depth below the surface increases.

Graph 1: Change in Temperature with Increasing Depth Below Earth's Surface



5. Which of the following is true of the temperature inside the Earth? It increases:

A. with depth at a constant rate
B. rapidly with depth near the surface, then remains constant
C. slowly with depth near the surface, then it increases more rapidly in the inner layers
D. rapidly with depth near the surface, then it increases more slowly in the inner layers

6. Which statement best describes what makes a gas balloon rise into the air?

A. temperature of the air increases as the balloon rises into the air
B. gas inside the balloon decreases in volume as the balloon rises into the air
C. mass of the balloon material is greater than the mass of the gas inside the balloon
D. density of the air surrounding the balloon is greater than the density of the gas inside the balloon

A scientist questioned the ability of fish raised in a hatchery (farm) to survive in the wild. She believed the fish raised in hatcheries had lost their fear of predators. To test her idea, she placed 15 hatchery salmon and 15 wild salmon of the same age into two separate but identical tanks. She then placed a clear piece of plastic into each tank. In each tank, she put the salmon on one side of the plastic and a large predatory fish, the cod, on the other side of the plastic. She then recorded the amount of time it took the salmon in each tank to move to the back of the tank away from the cod. She found that the hatchery fish were much slower in moving away than the wild fish. This led her to believe that the hatchery fish have less fear of predators than do wild fish.

7. What is the control variable in the experiment?

A. wild salmon
B. hatchery salmon
C. time it took the wild salmon to move away from the cod
D. time it took the hatchery salmon to move away from the cod

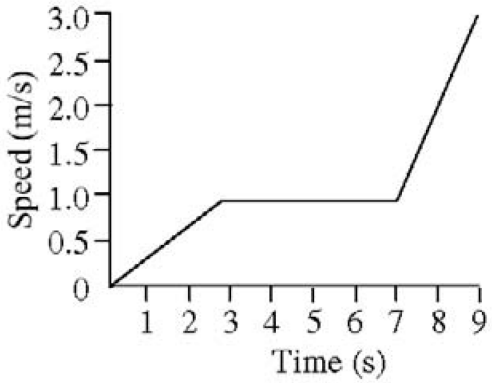
8. What is the hypothesis in the experiment?

A. hatchery fish have lost their fear of predators
B. wild fish have less fear of predators than hatchery fish
C. hatchery fish will move rapidly away from predators placed in their tanks
D. wild fish will survive attacks from predators more often than hatchery fish

9. What is the conclusion of the experiment?

A. wild fish swim more rapidly than do hatchery fish
B. hatchery fish have less fear of predators than do wild fish
C. hatchery fish will be able to survive in a wild environment
D. wild fish take more time to move away from predators than do hatchery fish

The graph below contains information about the movement of a bicycle.



10. At which time is the bicycle's speed constant? At:

A. 1 second
B. 2 seconds
C. 4 seconds
D. 8 seconds

Appendix H

Post-Intervention Survey Measure

Thank you very much for participating in this study on Computer Simulated Science Laboratory Assessment. Please ensure you have completed the consent form.
Please write your student code below:

_____

Post-Intervention Question

You have been hired as a science intern at the local weather station. Your supervisors at the local weather station were very impressed with your work on solving the three problems presented in the simulation. They now present you with the last problem for your internship 'How do the amount of helium, payload mass, and ***temperature*** together affect the altitude of a helium balloon?' However, your internship is almost over and you will not have time to test your experimental design so your supervisors have asked you to write up the best procedure to solve this problem so that your colleagues at the station may follow your procedure to run your experiment.

a) List the materials needed for your experiment: _____
_____
_____

b) Steps of your experimental design: _____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

c) What tools will you use to record your data:_____
_____

d) How will you organize your data: _____

_____

Post-Intervention Survey Measure Items

Survey 1: School Engagement Scale – Behavioral, Emotional, and Cognitive Engagement
Using the scale below and **thinking about your experiences of this simulated science lab**, please rate the following items. Please answer all items, even if you are not sure. Please select only a single rating for each item.

	Never 1	**On Occasion 2**	**Some of the time 3**	**Most of the time 4**	**All of the time 5**
1. I feel happy when using the simulated science lab					
2. I feel bored when using the simulated science lab					
3. I feel excited when using the simulated science lab					
4. I like the simulated science lab					
5. I am interested in using the simulated science lab					
6. The simulated science lab is fun					
7. When I read the instructions and post-lab conclusions, I ask myself questions to make sure I understand what it is about					
8. I plan to study the contents of the simulated science lab even when I won't have a test in the subject					
9. I plan to watch TV shows about the topics covered in the simulated science lab					
10. I checked my work on the simulated science lab for mistakes before clicking 'next'					
11. I plan to read extra books to learn more about the topics covered in the simulated science lab					

Survey 2: Revised Motivated Strategies for Learning Questionnaire

Using the scale below and **thinking about tests in general**, please rate the following items. Please answer all items, even if you are not sure. Please select only a single rating for each item.

	Not at all true of me 1	**2**	**3**	**4**	**5**	**6**	**Very true of me 7**
1. When I take a test I think about how poorly I am doing compared with other students.							
2. When I take a test I think about items on other parts of the test I can't answer.							
3. When I take tests I think of the consequences of failing.							
4. I have an uneasy, upset feeling when I take an exam.							
5. I feel my heart beating fast when I take an exam.							

Survey 3: Motivated Strategies for Learning Questionnaire

Using the scale below and **thinking about your experiences of this simulated science lab**, please rate the following items. Please answer all items, even if you are not sure. Please select only a single rating for each item.

	Not at all true of me 1	**2**	**3**	**4**	**5**	**6**	**Very true of me 7**
1. When I did the simulated science lab I thought about how poorly I was doing compared with other students.							
2. When I did the simulated science lab I thought about problems on other parts of the simulated science lab I couldn’t answer.							
3. When I did the simulated science lab I thought of the consequences of failing.							
4. I had an uneasy, upset feeling when I did the simulated science lab.							
5. I felt my heart beating fast when I did the simulated science lab.							

Survey 4: NAEP TRESim Background Questions

To what extent do you do the following on a computer? Include things you do in school and things you do outside of school.

	Not at all 1	**Small extent 2**	**Moderate extent 3**	**Large extent 4**
1. Play computer games				
2. Write using a word processing program				
3. Make drawings or art projects on the computer				
4. Make tables, charts, and graphs on the computer				
5. Find information on the Internet for a project or report for school				
6. Use e-mail and social networking site/apps to communicate with others				
7. Talk in chat groups or with other people who are logged on at the same time				

	Never or hardly ever 1	**Once every few weeks 2**	**About once a week 3**	**Two or three times a week 4**	**Everyday 5**
8. How often do you use a computer at school?					
9. How often do you use your own mobile device/tablet at school?					
10. How often do you use a computer outside of school?					

Please indicate the extent to which you AGREE or DISAGREE with the following statements.

	I never use a computer 1	**Strongly disagree 2**	**Disagree 3**	**Agree 4**	**Strongly agree 5**
11. I am more motivated to get started doing my schoolwork when I use a computer					
12. I have more fun learning when I use a computer					
13. I get more done when I use a computer for schoolwork					

14. Who taught you the most about how to use a computer?
    A. I learned the most on my own.
    B. I learned the most from my friends.
    C. I learned the most from my teachers.
    D. I learned the most from my family.
    E. I don't really know how to use a computer.

15. Is there a computer at home that you use?
    A. Yes
    B. No

16. Please indicate your gender:

☐ Male ☐ Female ☐ I prefer not to respond

17. Please indicate your birth date: _____ (month)/_____ (day)/_____(year)

18. Please indicate one or more of the following groups to which you self-identify in terms of ethnicity:

☐ Caucasian
☐ Chinese
☐ South Asian (e.g., East Indian, Pakistani, Sri Lankan etc.)
☐ African American
☐ Filipino
☐ Latin American
☐ Southeast Asian (e.g., Cambodian, Indonesian, Laotian, Vietnamese, etc)
☐ Arab
☐ West Asian (e.g., Afghan, Iranian, etc.)
☐ Japanese
☐ Korean
☐ Other: _____

Appendix I

Dependent Variables and TRESim Observables Measured during Study

Dependent variables measured before, during, and after TRESim laboratory

Type of Dependent Variables	Measures of Dependent Variables
Achievement	Prior-knowledge questions score (Appendix G) TRE computer simulated laboratory observables Post-intervention question score (Appendix H)
Engagement and Motivation	Pre- and post-intervention survey measure items (Appendices G and H respectively) which provided information regarding: mastery goal orientation performance-approach goal orientation performance-avoid goal orientation emotional engagement cognitive engagement intrinsic motivation extrinsic motivation text anxiety critical thinking learning strategies demographics

Rubric of TRESim observables measured during Problem 1

Observable Variables	Type of data collected
Degree of computer help	Frequency count: number of times this button was pushed
Degree of science help	Frequency count: number of times this button was pushed
Degree of use of glossary	Frequency count: number of times this button was pushed
Number of correct predictions made	Frequency count of the number of correct predictions made: # of times the 'It will rise into the air high above the ground' or D was pushed (this is the only correct hypothesis)
Number of predictions made	Frequency count of the number of predictions made.
Data organized with table or graph	Both graph and table were used (coded as 3); only table used (coded as 2); only graph used (coded as 1)

	Rubric description of data collected				
	3	2	1	0	-9
Choice of best experiment to solve problems	running all experiments systematically (e.g., increasing payload mass or amount of helium)	running experiments but not systematically (i.e., needs three experiments or more to determine it was not systematic)	running only one experiment that is not sufficient to determine whether it was done systematically or not OR only running two experiments	does not run any experiments	
Number, range, and distribution of experiments (running a set of experiments sufficient in number, range, and distribution to reveal the linear relationship between altitude and mass) *if the experiment does not span at least 40lbs, then drop them down a point level	running more than three experiments with the first and last being at least 40lbs apart (e.g., 10lbs, 30lbs, 60lbs, and 90lbs)	running three experiments with the first and last being at least 40lbs apart (e.g., 10lbs, 40lbs, and 80lbs)	running two experiments with the first and last being at least 40lbs apart (e.g., 10lbs and 80lbs)	running one experiment only	
Graph is useful to problem	x-axis is weight of payload and y-axis is altitude of balloon	x-axis is weight of payload and y-axis is time to final altitude	either the payload mass or the altitude is on the correct axis but the wrong variable is on the opposite axis (e.g., x-axis is weight of payload and time to final altitude is y-axis)	any other combination of graphs	did not create a graph
Table is useful to problem	table includes only altitude and payload mass	table includes altitude, payload	table includes only one of the required variables (i.e.,	any other combination of the table	did not create a table

		mass, and any other variables	altitude or payload mass) and other variables		
Accuracy of response to multiple-choice question			B or the 2nd choice is chosen	any other answer	
Degree to which conclusions are correct and complete	correct and complete ("best") responses to the constructed-response question with specific references to experiments (e.g., "As the payload mass increases, the balloon's altitude decreases. For example, when I put 90 lb. of payload on the balloon, it only went to 10,000 feet. But when I put 50 lb. of payload mass on the balloon, it went to 22,326, and when I put 10 lb., it went to 36,211 feet.")	correct but incomplete ("partial") responses that express the linear relationship between mass and altitude to the concluding question (e.g., "As the payload mass increases, the balloon's altitude decreases") with no specific references to experiments	wholly inaccurate response to the concluding question	did not produce scorable response for this observable	Did not complete item

Rubric of TRESim observables measured during Problem 2

Observable Variables	**Type of data collected**
Degree of computer help	Frequency count: number of times this button was pushed
Degree of science help	Frequency count: number of times this button was pushed
Degree of use of glossary	Frequency count: number of times this button was pushed
Number of correct predictions made	Frequency count of the number of correct predictions made: # of times the 'It will sit on the ground' is indicated for the volumes 2400 cu. ft. and below 'It will rise into the air high above the ground' for the volumes 2500 cu. ft. and above.

Number of predictions made	Frequency count of the number of predictions made.				
Data organized with table or graph	Both graph and table were used (coded as 3); only table used (coded as 2); only graph used (coded as 1)				
	Rubric description of data collected				
	3	2	1	0	-9
Choice of best experiment to solve problems	running all experiments systematically (e.g., increasing payload mass or amount of helium)	running experiments but not systematically (i.e., needs three experiments or more to determine it was not systematic)	running only one experiment that is not sufficient to determine whether it was done systematically or not OR only running two experiments	does not run any experiments	
Number, range, and distribution of experiments (running a set of experiments sufficient in number, range, and distribution to confirm that the relationship between altitude and amount of helium takes the form of a step function) *if the experiment does not span at least 1000cu. ft., then drop them down a point level	running more than four experiments to confirm a step function relationship which means they MUST have 2400cu. ft. and 2500cu. ft. in their series of experiments (e.g., 910cu. ft., 1700cu. ft., 2400cu. ft., 2500cu. ft., 2616cu. ft., and 3083 cu. ft.)	running four or more experiments to confirm a step relationship which means running two values ≤2400cu. ft. and two values ≥2500cu. ft. (e.g., 910cu. ft., 2275cu. ft., 2616cu. ft., and 3083 cu. ft.)	running three experiments that indicate a hyperbolic relationship which means at least one value ≤2400cu. ft. and two values ≥2500cu. ft. or vice versa (e.g., 910cu. ft., 2275cu. ft., and 3083 cu. ft.)	running one or two experiments that indicate a linear relationship (e.g., 910cu. ft. and 3083 cu. ft.)	
Graph is useful to problem	x-axis is amount of helium and y-axis is altitude of balloon	x-axis is amount of helium and y-axis is time to final altitude	either the amount of helium or the altitude is on the correct axis but the wrong variable is on the opposite axis (e.g., x-axis is amount of	any other combination of graphs	did not create a graph

			helium and time to final altitude is y-axis)		
Table is useful to problem	table includes only altitude and amount of helium	table includes altitude, amount of helium, and any other variables	table includes only one of the required variables (i.e., altitude or amount of helium) and other variables	any other combination of the table	did not create a table
Accuracy of response to multiple-choice question			E or the 5th choice is chosen	any other answer	
Degree to which conclusions are correct and complete	correct and complete ("best") responses to the constructed-response question that explain how the relationship between amount of helium and balloon altitude for a payload mass of 100 lb. takes the form of a step function (e.g., "Once the balloon has enough helium to rise into the air, the balloon will rise to a maximum height and go no higher matter how much helium is added.")	correct but incomplete ("good") responses referring either to the top or the bottom of the step function to the concluding question (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added.")	partially correct responses that express a linear relationship between altitude and amount of helium to the concluding question (e.g., "More helium inside the balloon will make the balloon go higher.")	wholly inaccurate responses to the concluding question OR did not produce scorable response for this observable	Did not complete item

Rubric of TRESim observables measured during Problem 3

Observable Variables	**Type of data collected**
Degree of computer help	Frequency count: number of times this button was pushed
Degree of science help	Frequency count: number of times this button was pushed

Degree of use of glossary	Frequency count: number of times this button was pushed
Number of correct predictions made	Frequency count of the number of correct predictions made: (different combinations have different correct hypothesis) 'It will sit on the ground' is only correct for the following combinations: 20lbs and ≤1400cu. ft. 30lbs and ≤1400cu. ft. 40lbs and ≤1400cu. ft. 50lbs and ≤1400cu. ft. 60lbs and ≤1700cu. ft. 70lbs and ≤1700cu. ft. 80lbs and ≤1875cu. ft. 90lbs and ≤2025cu. ft. 100lbs and ≤2025cu. ft. 110lbs and ≤2275cu. ft. 'It will rise into the air high above the ground' is only correct for the following combinations: 10lbs and ≥975cu. ft. 20lbs and ≥1500cu.ft. 30lbs and ≥1500cu.ft. 40lbs and ≥1500cu.ft. 50lbs and ≥1875cu. ft. 60lbs and ≥1875cu. ft. 70lbs and ≥2025cu. ft. 80lbs and ≥2275cu. ft. 90lbs and ≥2275cu. ft. 100lbs and ≥2500cu. ft. 110lbs and ≥2616cu. ft. 'It will bob lightly up and down on the ground' is only correct for the following combinations: 10lbs and 910cu.ft. 50lbs and 1500/1700cu. ft. 70lbs and 1875cu. ft. 80lbs and 2025cu. ft. 100lbs and 2275/2400cu. ft. 110lb and 2400cu. ft./2500cu.ft.

Number of predictions made	Frequency count of the number of predictions made.				
Data organized with table or graph	Both graph and table were used (coded as 3); only table used (coded as 2); only graph used (coded as 1)				
	Rubric description of data collected				
	3	2	1	0	-9
Choice of best experiment to solve problems	running all experiments systematically (e.g., increasing payload mass or amount of helium)	running experiments but not systematically (i.e., needs three experiments or more to determine it was not systematic)	running only one experiment that is not sufficient to determine whether it was done systematically or not OR only running two experiments	does not run any experiments	
Number, range, and distribution of experiments (running experiments for at least two values of mass and, for at least one of those values, conducting a set of experiments with amounts of helium sufficient in number and in range to confirm that the relationship between altitude and volume takes the form of a step function) *if the experiment does not span at least 1000cu. ft., then drop down a point level	running at least two masses, in addition to the 100lbs already done previously, and for at least two of the mass values conducting a set of experiments with a range (more than four experiments; the first and last being at least 1000cu. ft. apart) of values of helium to show a step function (e.g., 910cu. ft., 1700cu. ft., 2400cu. ft., 2500cu. ft., 2616cu. ft., and 3083 cu. ft. for 10lbs and 90lbs) *threshold for step must be included in range	running only one mass, in addition to the 100lbs already done previously, and conducting a set of experiments with a range (four experiments) the first and last being at least 1000cu. ft. apart) of values of helium to show a step function (e.g., 910cu. ft., 2275cu. ft., 2500cu. ft., and 3083 cu. ft.)*two values on each side of the step function threshold should be present	Running one or no masses and conducting a set of experiments with a range in helium to indicate a linear relationship (e.g., 910cu. ft. and 3083 cu. ft.)	Running one or no masses and conducting a set of experiments with a range in helium to indicate a linear relationship (e.g., 910cu. ft. and 3083 cu. ft.)	

Graph is useful to problem *There is no '0' in this category because no other choices are possible	x-axis is amount of helium and y-axis is altitude of balloon OR x-axis is payload mass and y-axis is altitude of balloon	x-axis is amount of helium and y-axis is time to final altitude OR x-axis is payload mass and y-axis is time to final altitude	x-axis is amount of helium and y-axis is balloon volume OR x-axis is payload mass and y-axis is balloon volume		did not create a graph
Table is useful to problem	table includes only altitude, payload mass, and amount of helium	table includes altitude, payload mass, amount of helium, and any other variables	table includes only one of the required variables (i.e., altitude, payload mass, or amount of helium) and other variables	any other combination of the table	did not create a table
Controlling variables	three or more mass or volume was controlled (i.e., same mass was ran for four volumes or one volume was ran for three masses)	two mass or volume was controlled (i.e., same mass was ran for four volumes or one volume was ran for three masses)	one mass or volume was controlled (i.e., same mass was ran for four volumes or one volume was ran for three masses)	running an insufficient number of experiments for controlled experimentation to be evaluated	Did not produce scorable response for this observation
Accuracy of response to multiple-choice question			E or the 5th choice is chosen	any other answer	
Degree to which conclusions are correct and complete	correct and complete ("best") responses to the constructed-response question that concludes how the relationship between amount of helium and balloon altitude for more than one payload mass takes the form of a series of step functions (e.g., "Once the balloon has enough helium to rise into the air, the	correct but incomplete ("good") responses to the constructed-response question that concludes either the top or the bottom of the step function (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added, and the maximum altitude	partially correct responses that can be derived from Simulation problems 1 or 2 to the concluding question (e.g., "Below a certain amount of helium the balloon cannot get off the ground.")	wholly inaccurate responses to the concluding question OR did not produce scorable response for this observable	Did not complete item

	balloon will rise to a maximum height and go no higher no matter how much helium is added.")	the balloon can reach decreases as payload mass increases.")

Conclusion questions were scored as binary items:

Question 1: correct solution is B (second answer)

Question 2: correct solution is D (fourth answer)

Question 3: correct solution is C (third answer)