Fall 11-22-2019

Assessing the Efficacy of Virtual Experiments in the General Chemistry Laboratory

Cory Jerome Hensen
Portland State University

Assessing the Efficacy of Virtual Experiments in the General Chemistry Laboratory

by

Cory Jerome Hensen

A dissertation submitted in partial fulfillment of the
requirements of the degree of

Doctor of Philosophy
in
Chemistry

Dissertation Committee:
Jack Barbera Jr., Chair
Erin E. Shortlidge
Dean Atkinson
Joel Steele

Portland State University
2019

ABSTRACT

As more students enroll in chemistry courses, institutions are faced with increasing costs and limited laboratory space to keep up with the demand. One solution some institutions have turned to is the incorporation of virtual experiments into the curriculum, as this can lower costs and increase the availability of laboratory space. Some institutions have offered sections that complete all of their experiments in a virtual environment, others have offered sections that alternate between a traditional hands-on experiment and a virtual experiment, and some institutions have replaced only select experiments throughout the curriculum with a virtual experiment. To begin to be able to assess the affective outcomes in laboratory settings that include virtual experiments, six existing affective scales were modified for use in the laboratory setting. Sufficient evidence of the reliability and validity of the data from the existing scales was found. The functioning scales were then used to assess the affective outcomes of a Beer's Law experiment, a calorimetry experiment, and a titration experiment in both a hands-on and virtual learning environment. To assess the cognitive outcomes in these experiments, rubrics based on common learning objectives were used to determine if students in both learning environments were able to meet instructors' learning objectives for the experiment. The affective and cognitive outcomes were compared for each experiment to determine whether there was a difference between learning environments and also across the three experiments. The findings of this work are presented throughout this dissertation.

DEDICATION

To my grandfather, Dr. Ronald Jerome Hensen, and my great-grandmother, Marie K. Denny, for teaching the importance of education

ACKNOWLEDGMENTS

First and foremost, I want to thank Dr. Jack Barbera. Jack invited me to follow him to Portland from the University of Northern Colorado where he started a new chemistry education research program. His continued support throughout my time at Portland State University made the world of difference as we both navigated the building of a program. No matter how busy Jack was, he always made sure there was the time in his day to answer whatever questions I had. I would also like to thank Dr. Erin Shortlidge for her guidance on the qualitative portions of this research, Dr. Joel Steele for his guidance on the quantitative portions of this research, and Dr. Dean Atkinson for his guidance with the laboratory portions of this research. I would also like to thank Dr. Eric Sheagley for his continued support and guidance throughout this research. It would not have been possible to implement any of the experiments without him. Additionally, I would like to thank Kirk Fisher, graduate teaching assistants, and the entire chemistry stockroom staff for their help in this research.

Next, I would like to thank the entire Barbera and Shortlidge research groups for their guidance and support throughout my time here. They have spent countless hours listening to my presentations, looking over my writing, and being a soundboard for my ideas, without them this would not have been possible. Additionally, there are two undergraduate researchers that were a tremendous help with this research, Medina Glenn and Gosia Glinowiecka-Cox. The entire groups' support has meant the world to me.

Outside of the academic support I have received, words cannot describe how much my family's support throughout this process has meant to me. The frequent visits and phone calls made this process easier knowing I had a team of support behind me.

Even when times were tough, my family was just a phone call away. We may not always live near each other but we always are there for each other. I love you all and thank you for your continued support in everything I do.

Last but certainly not least, I would like to thank Kackie Sitton. I know a long-distance relationship was not easy for either of us but your encouragement to pursue my Ph.D. no matter what has meant the world to me. You are there for me every day and are my biggest fan and supporter. I cannot wait to spend the rest of our lives together and not have to be in a long-distance relationship anymore. Thank you for everything you have done these past five years to make this possible. I love you!

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: Introduction

The human race has long been fascinated with the transformations chemicals can undergo. Whether this is the creation of the alloy bronze during the Bronze Age around 3500 BC, the study of alchemy in 300 AD, or the current research taking place in laboratories all around the world. While humans have come a long way in their understanding of chemistry since the days of alchemy and the Bronze Age, the way in which knowledge is passed down from an expert to a novice today is surprisingly similar, in many regards, to the past, but distinct in a few important areas. In the past, an expert alchemist would teach an apprentice the craft of alchemy and knowledge of chemical reactions with a combination of hands-on work and verbal explanations. Similarly, we teach students about chemical reactions by having them set up experiments under the supervision of an expert in a teaching laboratory. The first laboratories in chemistry were modeled off this apprenticeship model, where an aspiring chemist would learn from an expert chemist (Lindberg, 2015) . However, as education became more formal with universities starting to form, so did the way we taught chemistry. Chemistry quickly became a class that had a formal lecture component where experts could teach aspiring chemists about the theory of chemistry (Lucas, 1994). After this shift toward formal lecture, a young professor by the name Justus von Liebig, noted that the way in which chemistry was being taught was starting to drift from the roots of how it had been previously taught. Liebig asked his department to include a laboratory component to the lecture where students could not only learn the theory but also experience the reactions first hand. This gave birth to the teaching laboratory as we know it today (Sommer, 1931).

The laboratory was first used to teach the students basic facts about the beginning few elements on the periodic table, such as hydrogen, carbon, and nitrogen, and various reactions to either produce or use that element (Pickering, 1993). This approach was similar to an apprenticeship where the expert is teaching the novice how to carry out basic procedures used in the trade. After several years of each expert teaching their students their own way, the first laboratory manual was created in 1876 by Charles Eliot that unified how to teach these experiments and techniques across institutions (Pickering, 1993). Over the past 200 years, the laboratory has seen a shift from element first teaching to a focus on the conceptual underpinnings behind experiments, rather than trying to teach specific skills. As a result of this shift, there is no longer a unified laboratory manual with a number of different types of curricula having been implemented ranging from inquiry-based experiments (Novak, 1963) to experiments conducted in virtual reality (Georgiou et al., 2007) to many others. Each of these curricula tries to teach chemistry experiments in an innovative way that is a departure from how Liebig first envisioned the experiments in a teaching laboratory.

As different people have developed different curricula that depart from the original laboratory manual, they have tried to replicate what it means to be a chemist at the given time and incorporate it into the laboratory. For example, during Liebig's time, a chemist was someone who carried out specific procedures to create and use the elements. In the 1960s, the push was to reform the classroom to once again think like a chemist. However, chemists in the 1960s were tackling new problems and not generating the same chemicals from the 1800s. Instead, they were focused on inquiring about why principles of chemistry

worked the way they did. This gave rise to a large push toward inquiry-based learning (Novak, 1963; Rutherford, 1964; Schwab, 1958). After the 1960s, there have been many different curricula implemented that try to tackle the problem of having the laboratory closer mimic what a chemist does for a living. Curricula that have been tested in the laboratory include problem-based experiments (Nielsen et al., 2014), cooperative learning experiments (Cooper, 2005), and many more. While different laboratory curricula have been developed, there has not been any widespread adoption of new laboratory curricula, with some institutions using similar procedures to those from the first laboratory manual published in the 1800s. One such experimental procedure is the production of hydrogen gas by mixing an alkali earth metal with a strong acid. One reason that inquiry-based curricula and many other curricula that have been developed since the 1960s do not frequently have widespread adoption is that laboratory coordinators are not given enough tools to implement consistently (Hofstein, 2004; Hofstein & Lunetta, 1982). However, there has been more consistent implementation in recent years with more and more laboratory coordinators and faculty attending professional development workshops each year such as those at the Biennial Conference on Chemical Education (Stegall et al., 2016).

An additional factor that has limited widespread adoption of a unified curriculum is that there no longer a single job description of a chemist. With more career paths opening for people who have an understanding of chemistry, it makes it harder to design a one-size-fits-all laboratory experience for the students. In fact, the American Chemical Society (ACS) lists over 50 different career paths in chemistry on their website. (ACS) Some students may want an experience that is close to the traditional chemist path, whereas a

different student may be interested in a field that has recently emerged, such as working with polymer bioconjugates or immunoengineering. This leads to a problem that the field of education as a whole is facing and especially poses a challenge to the chemistry laboratory (Hagay & Baram-Tsabari, 2015). Laboratory coordinators have many different curricula they can choose from when designing the laboratory and no one approach may be best for all their students. Previous research has shown that some students do best in a guided-inquiry curriculum (Pavelich & Abraham, 1979) while other research suggests students do best in a problem-based curriculum (Kelly & Finlayson, 2007). It can be challenging for the laboratory coordinator to make sense of the different studies on the different curricula and decide which type of curricula best aligns with the interests of the students and their career paths.

Not only have the career possibilities for a chemist drastically increased, so have the tools chemists use in their career. Technology has been expanding at a record pace and has been integrated into many careers to either replace or assist humans (Washbon, 2012). For instance, there are auto-titrating apparatuses that replace a majority of titrations once done by hand. Yet in most undergraduate laboratories, students are still being taught to titrate by hand, as universities do not have the funding for or access to this technology for their curriculum. Technology has not only changed the job descriptions of chemists but, but it has also drastically changed the educational experience for students. Students now have the option to take classes in a variety of formats that would not be possible before. For the past decade, enrollment in online classes has continued to increase and does not show any signs of stopping (Nguyen, 2015). To accommodate this change in the way of

learning, universities, like Arizona State University, have created completely online degree programs in many subjects including in chemistry. However, even these online programs struggle with how to include technology in the laboratory. Many of them, as well as the field as a whole, are hesitant to get rid of the chemistry laboratory because they firmly value the role of the laboratory even if it does have flaws (Reid & Shah, 2007). Thus, one way to support an online program, but still have students gain laboratory skills, is by offering a summer program where the students from all over can come and spend a brief period of time on campus learning laboratory skills. During this brief period, students can get familiar with the techniques and equipment but may not be able to gain much deeper knowledge as all equipment and techniques are covered in a short time.

With the changing landscape in careers students taking general chemistry are seeking as well as and the incorporation of technology in these professions, laboratory coordinators may be unsure of which experiments are best to include in their curriculum. One factor in these decisions is the administration at their respective institution. Recently, laboratory coordinators and university administrations are facing increasing pressure to find solutions to increase enrollment while operating under a fixed amount of laboratory space. To alleviate these pressures, universities have added weekend sections, distance learning sections, and virtual sections (Tüysüz, 2010). While these solutions may alleviate the pressure on the administration, these decisions typically have relatively few peer-reviewed research studies to use as evidence for their decision. Thus, they are left with implementing what they view as ‘best for their situation’ without knowing what the best practices are. For example, to support distance learning, at-home experiment kits have been

sold by companies to universities and students for dozens of years. However, during that same time frame, there has been limited research conducted on the use of at-home kits (Kennepohl, 2007; Lyall & Patti, 2010; Reeves & Kimbrough, 2004). The same goes for virtual laboratories. As technology has advanced, companies have been able to develop virtual experiments that replicate some or all of the traditional laboratory experience, such as the LearnSmart Labs by McGraw-Hill Education (MHEducation, 2014). However, research on their effectiveness typically lags behind their implementation. It takes several years to conduct a thorough research study on the virtual laboratories, meanwhile, the technology may have already changed by the time the research is published.

Statement of Problem

While at-home kits and virtual laboratories are both possible solutions to alleviate some of the current pressures laboratory coordinators face, there is one major difference between these solutions. That difference is that at-home kits are still limited to experiments that can safely and easily be done in someone's home, regardless of the type of kit used. This limitation does not exist in the virtual environment. In the virtual environment, students can carry out near-unlimited possible experiments regardless of safety concerns or ability to get equipment. While this possibility is attractive, there has been limited adoption of virtual experiments to date.

Before there can be a clear plan of adoption, there first needs to be a clear understanding of the virtual experiments themselves as a baseline. There are many different companies producing virtual experiments that cover a range of topics, however, there is currently limited understanding on which topics are best to complete virtually, if any.

While a general chemistry laboratory curriculum has a wide variety of experiment types, some of the experiments focus more on tactile skill-building whereas other experiments focus more on conceptual understanding. It currently remains unseen if these different types of experiments give students the same benefits when the experiment is completed virtually. For instance, there are many different versions of titration experiments and the benefits the students get out of the laboratory from each different version is unknown. This leaves many remaining questions about the best adoption strategy for virtual experiments.

To detect any possible benefits or drawbacks from the adoption of a virtual experiment, it is important to measure relevant aspects of the learning process to determine the efficacy of the virtual experiments. These aspects may exist across one or more of the domains of learning. Bloom and colleagues proposed that students have three distinct domains of learning: the cognitive (Bloom et al., 1956), the affective (Krathwohl et al., 1964), and the psychomotor (Simpson, 1971). The cognitive domain focuses on the process of learning, the affective domain focuses on the 'feeling' aspects, and the psychomotor domain focuses on the 'doing' aspects. While all three are important for learning, the large majority of the research done on the laboratory, and the field of chemistry education overall, focuses heavily on the cognitive benefits and drawbacks (National Research Council, 2012). Universities are primarily concerned if students are learning the material well enough to move on to the next class and graduate, therefore it is natural for them to focus primarily on the process of learning. However, the affective and psychomotor domains are both equally important and can influence the cognitive domain. For example,

if someone is in a stressful learning environment then they may not learn as well (LePine et al., 2004).

For measuring cognitive domain outcomes, most laboratory studies use scores on laboratory assessment as a measure of how well the students learned during the experiment. For instance, measuring post-lab quiz scores (Hawkins & Phelps, 2013). However, these questions are typically self-authored without known psychometric properties. Alternative ways of measuring the laboratory cognitive domain can be used that have more evidence of validity and reliability. One way of doing this is by measuring students based on how well they meet faculty members' learning goals, as the students' ability to meet a learning goal is not inflated. While most studies on the laboratory measure the cognitive domain, in one form or another, very few measure the affective or psychomotor domain.

There is a body of literature that suggests there are only minor or no differences in the cognitive domain but there is much less support in the literature for what affective differences may exist between the two environments. Several laboratory studies that have included the affective or psychomotor domain indicate that students who show the same performance outcomes (cognitive domain) may be experiencing drastically different emotions during the laboratory (Galloway & Bretz, 2015b; Woodfield et al., 2004). That is, just because a student is doing well in the course does not necessarily mean that they experience positive feelings about and during the class. Thus, it is important to evaluate to what degree a learning environment both helps promote the learning of a student and supports the best experience for that student.

Purpose of Study

This study aims to add to the growing literature on what the best use of virtual experiments are and for who they work best for. Many different curricula have been designed and implemented over the past century but fail to gain widespread adoption because of the lack of clarity on how to best implement them (Hofstein & Lunetta, 1982, 2004). Virtual experiments are quickly falling into this same fate. More and more lab coordinators are starting to integrate virtual experiments to alleviate the increasing pressures that they are facing from the administration across STEM disciplines. For example, many universities include virtual experiments as part of their curricula, such as Motlow State Community College, Portland State University, Oregon State University, Arizona State University, and many more. However, the research on virtual experiments has been slow to catch up to the rate professors are incorporating these experiments into their classrooms. For instance, although many universities are using virtual experiments, there have been only a handful of recently published studies on their benefits or drawbacks at the general chemistry level (Chan & Chan, 2001; Hawkins & Phelps, 2013). Thus, there need to be additional studies conducted on the efficacy of virtual experiments to determine what their best use is so that laboratory coordinators can make more informed decisions and those decisions can be more consistent across institutions.

With a lack of studies conducted on virtual experiments as it pertains to the affective, a large part of this study will focus on assessing the impacts of virtual experiments on affective domain outcomes. Previous studies have measured aspects of the affective domain in both the classroom and the traditional laboratory (Galloway & Bretz,

2015b; Kim, 2005). However, there has been much less evidence put forth about the affective domain in a virtual environment. For example, are students less stressed when not working with chemicals or does working with technology cause them to be more stressed? Differences like these are remaining questions about when and for whom to use virtual experiments. Another remaining question this study will aim to address is: What type of experiment may be more beneficial to offer in a virtual environment? If laboratory coordinators are going to pick and choose a handful of virtual experiments to offer, it is imperative that they know which experiments work best in the virtual environment and which should be left for the traditional hands-on environment.

It is also possible that there is no one-size-fits-all adoption possible for virtual experiments and instead the impacts are seen differentially based on the characteristics of students. For instance, it is possible that students who enjoy coming to the laboratory may be more hesitant to perform an experiment online whereas students who do not enjoy coming to the laboratory may prefer it. It could also be that students that need the laboratory skills in the future, such as chemistry majors, need to build the skills whereas students who will not use the laboratory skills in the future may not need to perform the experiment in the traditional hands-on way. Therefore, it is important to characterize if there are any sub-populations that show differential benefits. With this knowledge it is possible to further assess why the differences may exist and how to design future curricula to be better for all sub-groups.

This study adds to the body of literature on virtual experiments, allowing laboratory coordinators to make more informed decisions that are based on evidence. This study will

address the gaps that currently exist in the literature by looking at how virtual experiments compare with traditional hands-on experiments across multiple terms, covering a range of content, and by including measurements in the affective domain when considering what makes a learning environment effective.

Research Questions

1. To what degree can previously developed scales be adapted for use in the laboratory environment?
2. What evidence of validity and reliability supports data collected with the adapted scales?
3. How do various aspects of the affective domain compare when students complete a general chemistry experiment virtually as opposed to a traditional hands-on experiment?
4. To what degree can experiments in a virtual environment meet the same learning objectives as similar traditional hands-on experiments?
5. To what degree do student characteristics explain differential outcomes in a general chemistry laboratory course?

Significance of Study

There is a need to establish a baseline for the efficacy of virtual laboratories before they can be further examined and have wide-spread adoption. Currently, the virtual experiment studies are focused on single institutions covering a single experiment and usually only measure what differences are present in the cognitive domain. To add to this literature and help establish a baseline, various aspects of the affective domain were measured to provide evidence on how aspects of students' affective domain change when the environment changes.

In addition to measuring these aspects of the affective domain, this study compared multiple experiments that cover a wide range of content and learning goals. In the general chemistry laboratory, some experiments emphasize conceptual understanding more than skill-building and vice versa (Hofstein, 2004). Thus, it is important to understand if there

were any differences in a virtual versus hands-on comparison that were due to the type of experiment itself. This information can provide evidence for laboratory coordinators to arrange their schedule in such a way that experiments are offered in the environment that students perform the best in with regards to both experiment performance and the specific components of affective state.

This dissertation will provide a body of evidence to use when decisions need to be made about what types of experiments to include in curricula and what possible impact those have on students. By including multiple experiments, it is possible to present evidence on what type of experiments may provide better outcomes in one environment versus another. By measuring aspects of the affective domain, it is possible to present evidence on how this study compares to previous studies. Additionally, it is possible to determine if certain students have increased affective aspects in a specific environment.

Limitations

There are many things to consider whenever active research takes place in a classroom and this study is no different. One of the biggest limitations was that the laboratories were already in place, with students paying for a quality experience. Thus, the virtual experiments needed to seamlessly fit into the curriculum without an extra burden on the laboratory coordinator or students. This limited the research design and type of studies that were carried out in several different ways. The most significant way is that the implementation of the virtual experiments needed a scaled implementation to establish a proof-of-concept before implementing across all laboratory sections at all participating institutions. Therefore, the sampling pool of institutions included only those willing to

scale the implementation across three years. However, future work will be able to build off this work and be implemented in many different types of institutions to work towards consistent implementation.

Relying on technology was another significant limitation. In fact, if the technology breaks down, the students miss out on the laboratory altogether. It is important that there are backup plans in place for the students to still be able to complete a laboratory if the technology breaks during their experiment. This troubleshooting is no different than troubleshooting when a pH probe doesn't work in a traditional laboratory. However, teaching assistants and other staff may be less comfortable doing this troubleshooting than normal laboratory equipment troubleshooting. If the problem cannot be solved quickly it may involve talking to the technical support at the company that owns the virtual experiments or canceling the experiment altogether.

In addition to these specific limitations with the virtual experiment itself, there were also limitations with how the data was collected. These limitations are as follows:

1. Convenience sampling was used, which limited the faculty members interviewed and students participating to the same geographic region in the United States.
2. The faculty members and students that did participate, had to provide informed consent and thus may not be representative of general chemistry instructors or students overall.
3. With only a few faculty members interviewed, it is possible that the stated learning goals are not representative of an average faculty member's goals for their course.
4. The virtual experiments were implemented only in general chemistry, therefore, results are not intended to be generalized to other courses or subject areas.

Given these limitations, all conclusions drawn from the data may have limited generalizability to other populations.

Definitions of Terms

Virtual experiment: For this research, a virtual experiment replicates an entire procedure that could be found in a traditional hands-on environment. They go above and beyond simulations by replicating all procedural steps.

LearnSmart Lab: An online suite of laboratory experiments offered by McGraw-Hill Education

CHAPTER II: Review of literature

The laboratory has long been a part of the curriculum across many disciplines, especially disciplines with a focus in science, technology, engineering, and/or mathematics (STEM). In chemistry, the practice of including chemical experiments as a way to teach novices about chemistry concepts can trace its roots back to at least 1820 with Justus von Liebig. Liebig was influential in shifting the teaching of chemistry to be taught from a laboratory perspective instead of a theoretical perspective (Ashley et al., 2011; Farrokhi & Mahmoudi-Hamidabad, 2012; Kohn, 1951; Lindberg, 2015; Pickering, 1993; Sommer, 1931). He convinced his colleagues and department to allocate a space where they could have a laboratory to teach students about chemistry through the act of doing an experiment with their hands rather than listening to a lecture. He adamantly pushed for this because of his firm belief in allowing students to learn by performing experiments with their own hands in a laboratory. This type of laboratory, a place to learn by performing experiments with the students' own hands, is commonly referred to as the teaching laboratory.

Laboratory experiments quickly expanded beyond Liebig's university and into many other universities, with each professor creating their own unique demonstrations and experiments. After decades of professors creating their own demonstrations and experiments, in 1876 Charles Eliot authored what is considered to be the first laboratory manual (Pickering, 1993). This manual is a compilation of hundreds of laboratory experiments, organized by element, that helped unify the laboratory and provided a snapshot of the current state of the chemistry teaching laboratory at the time of the writing (Eliot & Storer, 1876). By organizing the manual by element, it provided the current

knowledge of the properties of the element and reactions that produced or used that element.

Changes in the Laboratory Curricula and Environment

In the 150 years that have passed since the manual was published, laboratory manuals have become an integral part of the chemistry laboratory with many different types of laboratory manuals published to date (Beran, 2010; Cooper, 2005; Pavelich & Abraham, 1979). Although all of these manuals approach the laboratory in different ways, laboratory manuals at some universities still contain experiments near identical to experiments from the first laboratory manual authored in 1867. For example, the generation of hydrogen gas, experiment nineteen from Eliot's laboratory manual, is an experiment carried out in a similar way in today's teaching laboratory at some universities . In this experiment, zinc is reacted with an acid to produce hydrogen gas, which is then collected over water. In today's laboratory, this experiment is often the reaction of magnesium with an acid to produce hydrogen gas, which is then collected over water. While procedurally similar to the original experiment, the focus of today's teaching laboratory is conceptual knowledge related to the experiment whereas the original purpose was simply a way to prepare hydrogen gas. Although many universities are still using some of the fundamental concepts and procedures from the original laboratory manual, others have made more significant advances and departures from the original manual.

Changes in the Curricula

One significant departure from the original manual came when educators started to push for curriculum rooted in inquiry-based learning during the 1960s (Novak, 1963;

Rutherford, 1964; Schwab, 1958). While inquiry-based learning has many working definitions, one common definition breaks inquiry-based learning into five key core components (NRC, 2000).

These components are that:

- Learners are engaged by scientifically oriented questions
- Learners give priority to evidence, which allows them to develop and evaluate explanations that address scientifically oriented questions
- Learners formulate explanations from evidence to address scientifically oriented questions
- Learners evaluate their explanations in light of alternative explanations, particularly those reflecting scientific understanding
- Learners communicate and justify their proposed explanations

The rise of inquiry-based learning represented a drastic shift in curriculum from the primary focus of teaching facts to teaching how to think like a scientist and ask scientific questions (Schwab, 1958). While this new curriculum gained traction in the 1960s, the positive outcomes of early studies were not reproducible when tried in different environments and thus the curriculum typically was not implemented well (Williams & Hmelo, 1998) and did not immediately catch on (Tamir & Lunetta, 1981). However, the National Research Council renewed a push towards inquiry learning with their 2000 report (NRC, 2000). While not universally adopted, many institutions are using inquiry-based learning in their laboratory today (King et al., 2018; Rusek et al., 2018; Wheeler et al., 2017) including the development of the Advanced Placement (AP) guided-inquiry laboratory manual for AP high-school classes (CollegeBoard, 2013).

Inquiry-based learning was the first major departure from the original way of teaching the chemistry laboratory but it has not been the only departure. Educators and researchers have tried implementing numerous new curricula since the 1960s. These

curricula include cooperative experiments (Cooper, 1994), problem-based learning (PBL) (Kelly & Finlayson, 2007), course-based undergraduate research experiences (CUREs) (Linn et al., 2015), discovery-based experiments (Ricci & Ditzler, 1991), science writing heuristic (Burke et al., 2006), and many more. While each of these curricula have reported positive gains in student outcomes, no single curriculum has had widespread adoption. Instead, only some institutions have adopted new curricula with many institutions still using experiments that have their roots in the first chemistry manual. However, what was clear is that lab coordinators were starting to seek alternative solutions to the traditional confirmatory laboratory.

Seeking Alternative Solutions

The laboratory curriculum was not the only place that lab coordinators sought alternative solutions to the traditional way of doing experiments. Lab coordinators have historically had few alternatives to offer students who were not able to complete the laboratory in the physical space, such as at-distance learners (Holmberg & Bakshi, 1982). Additionally, there have been recent challenges with the laboratory that have left lab coordinators seeking other alternatives for all their students, not just the at-distance students. The first of these challenges is that there are more students enrolling in STEM majors than ever before (2017). This has left universities and lab coordinators to seek out creative options to ensure they have enough sections to meet the demand while still working with a finite amount of laboratory space and staff. One way that lab coordinators have tried to combat this challenge is by reexamining what the purpose of the laboratory actually is. For instance, universities are asking if all majors need to take the laboratory or

are there some students that the laboratory is not beneficial for. There is little consensus on what the purpose of the laboratory is, although most chemistry faculty members agree that the laboratory is a necessary part of the curriculum (Hofstein & Lunetta, 1982, 2004).

At-home Experiments

Not only is there variability in the curriculum two students attending separate institutions may get, but there is also variability in the type of environment they may experience. With a recent increase in students electing to enroll in non-traditional programs (Miller & Lu, 2003), such as online or distance learning programs, educators have faced a challenge to ensure the students completing the laboratory away from the physical campus have the same outcomes as those students completing the laboratory on the physical campus (Casanova et al., 2006).

One way that educators have combated this issue is to design experiments that could be safely done at the student's home and parallel the experiments done on campus. A popular way of designing these experiments is for an instructor to put together a list of what chemicals and equipment the student would need. From there, there are three typical solutions to how the students can get the chemicals and equipment. These three solutions are buying a commercially available kit, buying an instructor-made kit, or putting together a kitchen kit (Jeschofnig, 2004). A commercially available kit is one that a third-party company has built and contains all the materials that are needed for an experiment. An instructor-made kit has the added advantaged that the instructor had tailored made the kit to fit their experiments and they are not limited to only the experiments that the third-party

company supports. A kitchen kit is one that uses commonly found materials around the kitchen, which limits the number of supplies that need to be bought and sent to the student.

Regardless of the type of kit used, at-home experiments provide students an environment that tries to mimic the experience they would get on-campus in the comfort of their own home. Researchers have found that students using the at-home kits and students completing the experiments on-campus perform equally well in the course (Böhmová & Šulcová, 2007; Kennepohl, 2007). While they perform equally well, the kits are restricted on what materials they can contain since the materials need to be shipped and the experiment is conducted without proper laboratory safety equipment at the student's home. Therefore, the curriculum used for these experiments is most often the traditional confirmatory experiments rather than any of the new types of curricula.

Computer-based Experiments

To overcome the limitations imposed by students completing the experiments in their own home without proper safety equipment, educators have turned to alternatives. One attractive possibility is to use computer-based experiments to conduct experiments that have safety concerns present. This was a natural possibility since computers have long had a role in STEM laboratories. The earliest way that computers were adopted into the teaching laboratory was as data analysis and collection devices (Feisel & Rosa, 2005). For example, in 1972 Hewlett-Packard released the HP-35 that was capable of replacing the slide rule used in data analysis (Whitney et al., 1972). In today's laboratory, computers are still primarily used in data collection and analysis, including the use of the popular Vernier software, Logger Pro (Hechter, 2013). Logger Pro is a software that interfaces with probes

and sensors to assist with data collection from these devices. After the data is collected from the sensor, Logger Pro can also help aid in the data analysis. Since computers are already an integral part in the experiment process at most institutions, the question remained if the entire experiment could be simulated on a computer for the students, practically eliminating any safety concerns and allowing for any experiment to take place and indeed as computing power has rapidly increased since the 1980s, it became possible not only to collect data from sensors but also to predict what the data should be before the experiment takes place (Groth et al., 1981). Therefore, if computers can predict what the data should be from an experiment and help analyze the data, a computer can replicate most parts of a traditional experiment. This possibility gave rise to a common alternative approach to the traditional laboratory beside the at-home kits, the virtual laboratory (Dalgarno et al., 2009).

Virtual and Remote Experiments in STEM

A virtual laboratory experiment is one where the students interact with an artificial computer-based environment to conduct the experiment themselves (Martínez-Jiménez et al., 2003). For instance, the students conduct an experiment by interacting with simulated objects of the equipment and chemicals on their web browser. This allows the students to conduct the procedural steps themselves but does not give real experimental data (Tatli & Ayas, 2013). There are many different types of virtual environments that have been implemented in STEM laboratories that all try to replicate aspects of the traditional environment to different degrees. A simulation is not concerned with replicating the procedure of an experiment but instead simulates conceptual material that students cannot

gain as easily from the experiment (Clark & Chamberlain, 2014). In chemistry, simulations frequently illustrate interactions on the particulate level that cannot be visually observed in a hands-on experiment. A virtual experiment, however, replicates the entire procedure and data collection mimicking the major components of the hands-on experiment. A virtual world replicates as many as possible aspects of the chemistry experiment, including the interaction with others, which is accomplished by using avatars (Winkelmann et al., 2017).

While these types of virtual experiments are a popular alternative for students that cannot come in to complete the traditional laboratory, virtual experiments have been criticized for not generating real data with real experimental errors (Ma & Nickerson, 2006). Therefore, institutions that prefer real experimental data also have the option of using a remote laboratory. A remote laboratory experiment is one where the student watches a video, typically a live stream, of the experiment conducted at a remote site that has the proper safety equipment allowing them to conduct the experiment. (Senese & Bender, 2000). For example, a student can log in to the website where they instruct a robot to carry out the experiment for them. This allows the student to get real data and see how the experiment should look (Kennepohl et al., 2004). However, the students can only watch the experiment and not experience it with their own hands. Both of these environments allow the flexibility to build experiments for students based on any number of curricula since there are significantly fewer limitations on what types of experiments can be carried out in these environments.

Remote and virtual experiments in engineering

Within STEM fields, engineering has widely adopted virtual experiments in their curriculum whereas the natural sciences have been much slower to adopt them as summarized by a reviewed of the literature conducted by Ma and Nickerson (2006). For their literature search they used the Association for Computing Machinery (ACM), the Institute of Electrical and Electronics Engineers (IEEE), and ScienceDirect databases, along with various criteria to narrow their search to include only articles relevant to remote and virtual experiments. Of the 60 articles that they reviewed, 39 come from engineering with only 13 coming from the natural sciences. Of the 13 studies conducted in natural sciences, only 4 were looking at simulated laboratories in chemistry. The difference between the number of studies conducted in engineering and natural sciences could possibly be contributed to the fact that since engineering is an applied science that their labs may be seen as a "place to practice the application of scientific concepts" (Ma & Nickerson, 2006) whereas science professors "may see laboratories as a way of confirming beliefs". Since the publication of their review in 2006, virtual and remote laboratories have been studied more frequently in the natural sciences.

Virtual experiments in science

There has been a long-standing belief in many fields of the natural sciences that students get great benefits from experiencing phenomena with their own hands. This is the primary reason that these disciplines include a laboratory component to their classes that complements the lecture component. Therefore, the natural sciences were slower than engineering to begin to examine how virtual experiments could replace or enhance the

traditional experiments. However, over the course of the past decade, virtual experiments have become more sophisticated and have started to gain traction in the natural sciences including the disciplines of physics, biology, geology, and chemistry.

In physics, traditionally students explore concepts by setting up apparatuses and then experimenting with them. However, recent studies in physics have investigated whether students would understand the same concepts these apparatuses teach without ever physically using the apparatus (Moser et al., 2017; Zacharia & De Jong, 2014; Zacharia & Michael, 2016). For example, in a study by Puntambekar et al. (2012) students were broken into two experimental groups where the first group learned how to use a pulley system with physical manipulates the first week and a virtual pulley system the second week and the second group did the opposite. Both groups of students were then given a conceptual test one week after the experiment and there were no statistically significant differences found between the groups. This result is similar to many previous studies across STEM disciplines that find no difference between virtual and physical experiments (Brinson, 2015). In fact, a similar study in physics also found the same null result. In this study by Zacharia and Constaninou (2008) students were broken into two experimental groups and completed an experiment over the concept of heat and temperature using either physical or virtual manipulatives, respectively. Regardless of the concept the students were tested on, both studies showed no statistically significant difference on their conceptual knowledge when using either virtual or physical manipulatives. Other studies exploring virtual experiments in physics also showed no statistically significant difference between the two types of experiments (Hawkins & Phelps, 2013; Ma & Nickerson, 2006). While most

studies find no statistically significant difference, there are examples where the students completing the virtual experiment outperform students completing the experiment in the traditional hands-on environment. For example, in a study by Finkelstein et al., students completed a direct current circuit experiment in either a simulated environment or the traditional hands-on environment (Finkelstein et al., 2005). It was found that the students who completed the experiment in the simulated environment outperformed their counterparts not only on a conceptual survey but also in their ability to build a real circuit in person. This means that even though the students in the simulated group learned the technique in a virtual environment, they still outperformed the students who learned the technique in a hands-on environment when both groups were tested in a hands-on environment. Since there is a body of literature that suggests that there is either no difference between the two environments on the measured variables or a slight advantage to completing experiments in a virtual environment, researchers have suggested virtual experiments can be used to conduct experiments in physics that are not accessible or safe in a normal teaching laboratory. For instance, nuclear experiments are capable of being conducted virtually but not in a traditional teaching laboratory.

Biology is another natural science discipline that has been able to use virtual experiments to explore concepts that are not as feasible in the traditional laboratory. For example, visualizing fundamental processes such as the growth and death of cells (Slepchenko et al., 2003). Biology virtual experiments are also able to provide students practice with experimental techniques that they learn in the traditional laboratory. One example of virtual experiments built for biology is the iLaBS developed by Dr. Raineri and

published by McGraw-Hill. Raineri (2008). These virtual experiments allow students to replicate costly traditional experiments, such as sequencing experiments or cloning experiments, in a virtual environment to reduce the cost while allowing students multiple opportunities to run these procedures.

While biologists have been keen on using virtual experiments to expose students to experiments or techniques that could not traditionally be as easily performed, such as those at the cellular level, there has also been research conducted on the effectiveness of using a virtual experiment to replace a traditional experiment. One common laboratory experiment in general biology is an experiment covering the technique called Polymerase Chain Reaction (PCR). To see how students would respond to learning this technique virtually, Henderson-Begg et al. (2009) developed an online experiment that replicated the technique. The students were split into two groups with one group completing the experiment in the traditional environment and the other in the virtual environment. After completing the experiment in either environment, the students were assessed on their gains in content knowledge. In addition, the students also completed an assessment to evaluate how well they learned the technique of PCR. To do this, both sets of students completed the technique in a traditional setting regardless of the environment they first practiced the technique in. Like many of the studies conducted across STEM disciplines, the students did not have a statistically significant difference in their learning gains (Ma & Nickerson, 2006). However, similar to outcomes found from virtual physics experiments (Finkelstein et al., 2005), it was reported that the students who completed the experiment virtually needed less guidance from the demonstrator when performing the in-person practical.

Another study in biology also highlighted the potential benefits of virtual experiments beyond just content knowledge. Lazarowitz et al. (2002) found that students who otherwise would not have been successful in the laboratory were able to use the virtual environment to perform better than expected. The students who completed a microbiology experiment virtually and had low reasoning abilities were able to still successfully learn the concepts whereas the students who completed the experiment traditionally were not able to.

Virtual experiments in chemistry

In addition to the studies conducted in other natural science disciplines, there have also been studies conducted on virtual experiments in chemistry. Similar to other STEM disciplines, studies conducted in chemistry typically find no difference between virtual and traditional hands-on experiments when measuring content knowledge (Ma & Nickerson, 2006). Therefore, several studies have gone beyond content knowledge and examined what additional benefits may exist (Woodfield et al., 2005; Woodfield et al., 2004). In these studies, Dr. Woodfield tested virtual experiments in both the organic and inorganic curricula at BYU. Unlike previous studies that have a control and comparison group, all students enrolled in the corresponding sections completed the virtual experiment. In the organic class, the students completed sophisticated organic synthesis simulations and qualitative analysis (Woodfield et al., 2005). In the inorganic class, the students completed a sophisticated inorganic qualitative analysis (Woodfield et al., 2004). They then were able to use historical data to see if there were any benefits to switching several experiments to be virtual. The exam scores in both the organic and inorganic classes both showed no change or slight improvement, which matches previous findings across STEM disciplines

(Ma & Nickerson, 2006). In this case, the researchers concluded that the slightly higher exam scores, in some cases, may be contributed to the students' deeper understanding of content when they can establish trends for all the reactions possible and not just the one assigned to them for that lab.

These studies go beyond comparing just grade data to examine what other benefits virtual experiments may provide. The first benefit was that many students reported preferring the virtual environment (Woodfield et al., 2005; Woodfield et al., 2004). When the virtual experiment was implemented in the organic laboratory, students had a better transfer of knowledge at the end of the experiment when they could explore all the different reactions that were available to them in the virtual experiment that could never be explored in a traditional lab due to time limitations. That is, students were able to use what happened in the reaction they saw and answer questions about a fictional reaction. Thus, students were learning higher-order cognitive skills, specifically how to apply their acquired knowledge in new contexts, and in turn, performing better on the exams than previous terms where these virtual labs were not used. Instead of focusing on the "how" that a traditional lab usually focuses on, students were shifting toward the "why" and asking more conceptual questions. This led to an increase in exam performance, as most of the exam questions were conceptual "why" questions.

Aspects of the affective domain were also measured in both the inorganic and organic implementation to determine what benefits may exist in the affective domain and for whom virtual experiments are benefiting. To do this, Hermann Brain Dominance Instrument (HBDI) scores were measured and there was a statistically significant

difference for people with high left-brain scores being more satisfied with the virtual laboratory than those with low left-brain scores. A high left-brain score indicates people who are "verbal and structured in their thinking, efficient, time-oriented, linear, and precise" (Herrmann, 1995). The design of the virtual environment allowed the experiment to be linear in nature with a clear path for completion and thus it tended towards a left-brain preference (Woodfield et al., 2004). After implanting virtual experiments in both organic and inorganic laboratories, Woodfield and colleagues suggest that the virtual labs were best used as a supplement to the classroom and laboratory to increase the flexibility and exploration of an experiment.

In addition to the organic and inorganic levels, virtual experiments have also been studied at the general chemistry level (Hawkins & Phelps, 2013; Irby et al., 2017; Winkelmann et al., 2017). For example, Hawkins and Phelps examined how well students can learn laboratory skills from replacing traditional experiments with virtual experiments. To measure the differences that may exist between these two environments the researchers administered both a pre- and post-test. During the lab practical, all students assembled an electrochemistry cell regardless of the laboratory environment in which they initially completed the experiment. Interestingly, the authors found that regardless of the environment both groups performed equally well at constructing the cell. This finding aligned with previous studies showing that even when students learn a skill in a virtual environment they are able to do that skill in person given the right equipment (Banerjee et al., 2007; Cobb et al., 2009; Finkelstein et al., 2005). Additionally, students performed equivalently on the posttest after having either laboratory environment. Thus, for the

general chemistry students at their institution, the virtual experiment allowed the students to gain the same knowledge and skills that they would have otherwise learned in the traditional environment.

There have also been studies done at the high school level in chemistry to compare virtual experiments to traditional hands-on experiments (Davenport et al., 2012; Hou & Lin, 2017; Pyatt & Sims, 2012; Tüysüz, 2010). The studies at the high school level offer insight into how students respond to virtual laboratories and the different ways of implementing them. For example, Pyatt and Sims (2012) used a crossover design that allowed them to monitor if any changes they observed were due to the experiment being virtual or hands-on or if they were related to the content that was covered in the experiments. The students in this study were split into two groups with the first group performing a virtual experiment the first week and a traditional hands-on experiment the second; the order was reversed for the second group. The students' answered content questions about the experiment and their scores were compared across groups. It was found that the students performed equally well in both environments as is similar across STEM disciplines. Additionally, the authors go beyond the typical comparison and also examined what aspects of the affective domain may be different between the two environments. One of the affective findings mirrored earlier studies that found students prefer virtual experiments (Woodfield et al., 2005; Woodfield et al., 2004). Although this study was done at the high school level and has not been replicated at the college level, it is in line with the college studies that have found students perform better or the same in a virtual environment and tend to prefer the virtual environment.

Cognitive Domain Differences

When implementing new educational changes, such as the inclusion of virtual experiments within a STEM discipline, researchers often focus on performance gains and other measures that capture the cognitive domain. The cognitive domain encompasses the aspects that relate to students' mental skills or knowledge. Therefore, researchers can determine if an intervention had any cognitive impact on the students using some sort of cognitive assessment. To measure this in a laboratory setting, researchers have used scores on assessments given in the laboratory such as lab practical scores (Hawkins & Phelps, 2013), lab report scores (Choi et al., 2013) and pre/post-lab quizzes (Winkelmann et al., 2017). However, these measures frequently produce skewed distributions with a majority of students performing above 80%, since most laboratory curricula are designed in such a manner that most students pass the course. Some studies instead use other measures of the cognitive domain that are less sensitive to the skew, such as interviews with the students (Woodfield et al., 2005).

One way to address this challenge is to use cognitive measures that are not based on graded assessments. For example, it is possible to measure how students perform on learning goals instead of the points associated with meeting the learning goal. However, it can be harder to quantify how well students are meeting learning goals, which is why it is easiest to measure graded assessments. Bruck, Towns, and Bretz (2010) researched what learning goals various faculty members had for the laboratory to determine what consistent learning goals existed. Following up on the initial project, they found seven different goals for that laboratory that were informed by 312 faculty members (Bruck & Towns, 2013).

These seven goals were: “research experience”, “group working and broader communication skills”, “error analysis, data collection and analysis”, “connection between lab and lecture”, “transferable skills (lab-specific)”, “transferable skills (not lab-specific)”, and “laboratory writing”. These goals establish common ground for professors and lab coordinators to evaluate if the students are meeting the expectations for the chemistry laboratory course. These studies have helped establish common learning goals professors have for the course as a whole but not every experiment may address these goals. For example, while some experiments in the course may give students research-like experience, some experiments may not meet that goal. Therefore, there is a difference in ‘course’ learning goals versus ‘experiment-specific’ goals. While the previous studies help establish a consensus on what professors desire their students to get out of the course as a whole, there has yet to be established learning goals for specific experiments, due to the wide variation in experiments conducted over the same topic in the general chemistry laboratory. This, in turn, means that students at different universities may be doing an experiment on the same topic but leave with very different cognitive outcomes. Thus, to be able to compare specific experiments across different universities it would be beneficial to measure learning goals for specific experiments rather than broad course learning goals.

Affective Domain Differences

Potential differences between a virtual and a traditional hands-on laboratory may exist in the affective domain across many different variables, in addition to any differences on cognitive assessments. Previous classroom intervention research in STEM disciplines have found many different variables, or constructs, that can be potentially different

between the treatment and control groups. While a variable in the cognitive domain is typically a measurable result such as a lab report score, variables in the affective domain are typically not directly measurable variables. Instead of directly measuring the variable, indirect measurements are taken that allow for the direct measurement to be inferred. When a variable is indirectly measured like this, it is referred to as a construct.

There have been many affective constructs that have been studied in classroom intervention research. For example, self-concept (Kim, 2005), interests (Jones et al., 2000), attitudes (Astin, 1977), self-esteem (Crocker et al., 1994), ego development (Adams & Fitch, 1983), feelings (Simpson, 1978), locus of control (Abouserie, 1994), personality (Felder et al., 2002), and anxiety (McCarthy & Widanski, 2009). Teaching assistants and laboratory professors typically have a good feeling for the state of many of these constructs for their students. For example, it is easy to notice which students are more interested in the laboratory material than others or which students seem highly anxious handling chemicals. While the professor may be aware of these constructs while teaching, the National Research Council found that in less than half of the research studies they examined the researchers measured any aspects of the affective domain. (National Research Council, 2012) Therefore, while there has been literature that suggests there are either minor or no differences in the cognitive domain when a virtual laboratory environment is compared to the traditional environment, there is much less support for what affective differences may exist between environments.

One difference that may exist between individual students is their level of motivation. Motivation has been frequently researched because motivation and its sub-

components have been found to be one of the biggest predictors of success in the chemistry classroom (Zusho et al., 2003) as well other classroom settings (Bandura, 1997; Dweck & Leggett, 1988; Pintrich, 1999; Pintrich & Garcia, 1991; Pintrich & Schrauben, 1992; Schunk, 1991). Another difference that may exist that is closely related to the students' levels of motivation is their initial interest. To measure initial interest in chemistry, Ferrell, Phillips, and Barbera (2015) adapted an interest scale developed by Harackiewicz et al. to determine how initial interest in chemistry predicted final course performance in chemistry (Ferrell et al., 2016). It was found that chemistry majors reported higher self-efficacy and interest than non-science majors, which correlated to their academic performance. Thus, initial interest may help explain any differences found in the cognitive domain.

Unsafe 1 2 3 4 5 6 7 Safe

Figure 2.1: Example of semantic differential

A widely used instruments in chemistry to measure aspects of the affective domain and how it correlates to the cognitive domain is the Attitude toward the Subject of Chemistry Inventory (ASCI) (Bauer, 2008). This instrument measures five aspects of the affective domain; intellectual accessibility, interest and utility, fear, emotional satisfaction, and anxiety. It was developed from a framework for semantic differential questions based on the works by Osgood (Osgood et al., 1975; Snider & Osgood, 1969). A semantic differential is a question where the respondent has to choose how to respond on a scale of polar opposite words. An example of a semantic differential is shown in Figure 2.1.

After coming up with twenty adjective pairs that would work for a semantic differential, the instrument was administered and 379 usable data points were collected.

With these data points, Bauer ran an exploratory factor analysis (EFA). An EFA gives a mathematical model that examines how the items are related to each other. It was first formalized by Spearman in 1904 but has traces of many other sources (Mulaik, 1987). By running an EFA, Bauer was able to see that of the twenty items initially authored, there were five factors (or constructs) that his items were addressing. After reviewing these items and knowing where they originated from, Bauer named the first factor "Interest and Utility', the second factor "Anxiety", the third factor "Intellectual Accessibility", the fourth factor "Fear", and lastly the fifth factor "Emotional Satisfaction".

While five factors were resolved when the EFA results were analyzed, Bauer provided little theoretical evidence for these five factors. Thus, the instrument was further revised by Xu and Lewis to provide more evidence for the construct validity of the instrument. In other words, to ensure that the items were derived from theory and not just random item groupings (Xu & Lewis, 2011). Xu and Lewis revised the ASCI with a goal of making a psychometrically sound instrument that aligned with the general theory of the intended affective domain aspects. To accomplish this, each of the five original constructs was run as a single factor confirmatory factor analysis (CFA) to measure how well the items fit onto a single construct. A CFA is an analysis technique used to measure how well items identify the overall latent construct they are hypothesized to belong to. For example, how well Bauer's anxiety items fit the construct of anxiety. After analyzing the CFA results, they were left with 8 items that resolved into two components. This left the revised four-item "Intellectual Accessibility" construct and the revised four-item "Emotional Satisfaction" construct as part of the theoretically-driven ASCI V2.

The ASCI has been used widely to look at different classroom interventions. For example, it has been used to identify at-risk students in general chemistry (Chan & Bauer, 2014), evaluating a large-enrollment flipped classroom (Mooring et al., 2016), to examine students' attitudes in a modified POGIL classroom (Vishnumolakala et al., 2017) and many other applications. Although it has been widely used, there are no instances of the ASCI being used to examine interventions in the laboratory. Despite a lack of use of the ASCI or ASCIv2 in the laboratory, there have been several other instruments used to measure aspects of the affective domain in the laboratory. One example at the college level is from Galloway and Bretz who researched "meaningful learning" in the laboratory, which is a term relating back to Novak's theory of meaningful learning that states that the affective domain, the cognitive domain, and the psychomotor domain all need to interact in order for meaningful learning to occur (Novak, 2002). To capture this, Galloway and Bretz developed the Meaningful Learning in the Laboratory Instrument (MLLI) (Galloway & Bretz, 2015a). Overall, the results from the MLLI were that students did not have their expectations met after completing the laboratory.

Another example of an affective instrument for the laboratory is the Virtual and Physical Experimental Questionnaire (VPEQ) (Pyatt & Sims, 2012). This example is from the K-12 literature where Pyatt and Simms researched how specific aspects of the affective domain, as measured by the VPEQ, and performance changed when students conducted a virtual experiment instead of a traditional hands-on experiment. The VPEQ was created by compiling existing items, as well as writing new items, and was used to measure aspects of the affective domain in both a virtual and a physical laboratory. The VPEQ uses the

‘Usefulness of computers’ and ‘Anxiety towards computers’ from the Attitudes towards Computer and Computer Courses instrument (Woodrow, 1994) and the ‘Open-endedness of lab’ scale from the Science Laboratory Environment Inventory (SLEI) (Fraser et al., 1993). In addition to these published scales, Pyatt and Simms added an ‘Equipment usability’ scale and a ‘Usefulness of lab’ scale to capture expected differences between the virtual and hands-on experiments. While this instrument measures aspects of the affective domain when doing a virtual experiment compared to a hands-on experiment, this instrument has only been used in the initial study by Pyatt and Simms, which was conducted at the K-12 level.

Another affective instrument that has been used and developed for the laboratory measures anxiety. Unlike the ASCI that measured a general anxiety construct, Bowen developed the Chemistry Laboratory Anxiety Instrument (CLAI) to measure laboratory-specific anxiety, such as students’ anxiety with working with chemicals, using equipment and procedures, collecting data, working with other students, and having adequate time to complete an experiment (Bowen, 1999). Students have shown differing levels of anxiety when in the laboratory and thus the students’ anxiety level may be informative of the differences they perceive between a virtual and a hands-on environment (Alkan & Koçak, 2015; Ercan, 2014; Ural, 2016).

While there are many expected differences in the students’ affective domain between a hands-on experiment and a virtual experiment, not everything can be measured in one study due to power limitations and test fatigue. Therefore, students should be asked the least amount of questions that allow the constructs to be measured accurately to limit

test fatigue. With a limitation on the number of questions that can be asked, it is important that the questions that are asked are relevant to the research questions being addressed and that the questions have support from existing literature to how well they work. Additionally, large sample sizes are needed to accurately measure how multiple constructs interact when doing a virtual experiment compared to a hands-on experiment.

Measurement

In both the cognitive and affective domains, it is important that researchers are able to accurately and reliably measure what they are intending to measure. For example, if one wishes to measure a student's anxiety levels using an assessment instrument, it is important that the researcher has evidence that the instrument accurately measures anxiety levels and that it works equally well across a range of different students. The researcher, therefore, needs to provide evidence for the accuracy and reliability of the instruments they use to measure aspects within both the cognitive and affective domains. Within psychological and educational research, the accuracy of the measure is termed 'validity'. Validity is defined as a property that assesses how well the measure assesses what the measure intends to (Hammersley, 1987) and there are many different ways of providing evidence of validity such as consequential validity (Sambell et al., 1997), external validity (Calder et al., 1982), structural validity (Muenjohn & Armstrong, 2008), and more. Reliability is defined as the reproducibility of the measure, and like validity, there are many different ways of providing evidence to support the reliability of a measure such as test-retest reliability (Weir, 2005), interrater reliability (James et al., 1984), internal consistency (Henson, 2001), and more.

Validity

Researchers provide validity evidence for their measure to ensure that it is accurately measuring what it is intending to measure. The intended measure is also referred to as a construct and therefore this type of validity is referred to as construct validity. Construct validity has been theorized to be multi-faceted and includes six underlying aspects (Cronbach & Meehl, 1955; Messick, 1989). These six aspects are content validity, substantive validity, structural validity, external validity, consequential validity, and generalizability.

Content validity examines the extent to which the measure is relevant and represents the domain of interest. For example, researchers have provided evidence that the questions on the Chemical Concepts Inventory (CCI) do not fully span the domain of interest and lack this aspect of validity (Schwartz & Barbera, 2014). Substantive validity is similar to content validity but instead examines how well individual items represent the domain of interest (Loevinger, 1957). Structural validity examines the relationship between the items and the dimensionality of the measure. A Confirmatory Factor Analysis (CFA) generates item loadings based on the data observed. A higher item loading indicates that the factor has a stronger influence on that specific item. If all items have an acceptable loading and the overall model has acceptable fit, there is evidence for the structural validity. Evidence for convergent and discriminant validity is provided by relating the measure of interest to other measures with either a positive or a negative expected correlation. (Campbell & Fiske, 1959) A measure that is expected to correlate positively with the measure of interest provides evidence for convergent validity whereas a measure that is expected to have no correlation with the measure of interest provides evidence for

divergent validity. For example, a researcher demonstrated that the American Chemical Society (ACS) first-semester exam is positively correlated with the students' second-semester performance, which provides evidence for convergent validity of the first semester exams (Lewis, 2014). In this study, the authors also provide evidence for consequential validity. Consequential validity provides evidence for the consequences of the measures score interpretation (Messick, 1989). For example, evidence of consequential validity is provided for the ACS exam by examining if the score from a given ACS exam can accurately indicate future chemistry performance (Lewis, 2014). Lastly, generalizability evidence is closely related to external validity. Generalizability supports that a measure maintains its accuracy within other settings. One example of providing evidence for generalizability is by conducting national cross-section surveys that capture a wide audience to ensure the measure of interest is working equally well across different participants (Galloway & Bretz, 2015b).

Reliability

In addition to validity evidence, researchers should also provide reliability evidence for their instruments. Reliability is an indication of how consistent a measure is and is typically tested through one of three methods; coefficient alpha, test-retest, and interrater reliability (Komperda et al., 2018). Of these three types, researchers most frequently use coefficient alpha to report evidence for reliability. Coefficient alpha is a measure of how well the items that identify a construct are related to each other (Cronbach, 1951). In other words, it measures if all the anxiety items on an instrument are consistent and measuring anxiety or if one or more of the items are dissimilar to the rest. While the items may be

consistently measuring the same construct, it is possible that that is only true for the population that the items were originally tested on and the items may function differently with a different population. Test-retest reliability evidence is provided to help alleviate this concern. Test-retest reliability measures how consistent the instrument is when it is given multiple times (Guttman, 1945). When the items are consistent within the construct and the instrument as a whole is able to produce consistent findings there is strong evidence that the instrument is reliable. Sometimes, however, the variables of interest are measured through qualitative methods rather than quantitative. In this case, reliability evidence can be provided through interrater-reliability. Interrater-reliability is measured based on how well two independent coders reach the same conclusion (McHugh, 2012). For example, whether both coders think the interview transcript means the same thing (Miller et al., 2003). Through these three different ways, researchers are able to provide evidence for how consistent the variables of interest are measured. Evidence for reliability can then be paired with evidence for validity to demonstrate that not only are the variables of interest measuring what the researcher was hoping to measure but also that the variables are measured consistently.

Conclusion

Virtual experiments have been widely used across STEM disciplines and are starting to gain traction in chemistry. As they gain traction, it is imperative that any differences between a virtual and traditional hands-on environment be accurately measured. These differences may exist in either the affective or cognitive domain and as such, there is a need to measure relevant aspects of both domains rather than solely focusing

on one domain. To measure the affective domain, there need to be instruments that have provided evidence for the reliability and validity of their use in the laboratory. Only with instruments that are properly functioning can it be determined what potential positives, and negatives, exist when virtual experiments are implemented in the general chemistry curriculum. Future work should provide validity and reliability evidence for the intended measures to strengthen the comparison of the two environments and provide insight into what potential differences may exist for students completing laboratory experiments in either environment.

CHAPTER III: Methodology

As more students enroll in online and other non-traditional chemistry degree programs (Seaman et al., 2018), it is imperative that the efficacy of their laboratory options are well understood. One of the biggest challenges that these types of programs face when offering a chemistry degree is how to offer the laboratory component of chemistry classes in an online environment (Brewer et al., 2013; Dalgarno et al., 2009; Georgiou et al., 2007). While various types of virtual environments and experiments have been implemented in the chemistry laboratory sequence, there is no clear consensus on all the benefits and drawbacks of virtual experiments (Hawkins & Phelps, 2013; Winkelmann et al., 2014; Woodfield et al., 2005; Woodfield et al., 2004). A majority of the studies are single-institution studies that have students complete one virtual experiment and only compare performance, as indicated by a grade on that experiment, between students completing the experiment in a hands-on environment with those completing the experiment in a virtual environment. This focus leaves many questions unanswered about the efficacy of virtual experiments and what other benefits and drawbacks exist outside of performance measures.

The first of these questions is whether students meet the instructor's learning goals differently when completing an experiment in a virtual or traditional hands-on environment. With most studies focusing on performance as measured by grades, they do not capture the potential difference in learning goals. The second of these questions is whether students differ on affective aspects when completing an experiment in a virtual or hands-on environment. Very few studies consider the affective impact of virtual environments on students (Tüysüz, 2010) with the majority only focusing on the cognitive

differences (Hawkins & Phelps, 2013). Another question is if all students experience the same gains in the affective and cognitive domain or if certain types of students, e.g. chemistry majors, experience differential gains as compared to their peers. These questions are further broken down into five specific questions that guided this research.

Research Questions

1. To what degree can previously developed scales be adapted for use in the laboratory environment?
2. What evidence of validity and reliability supports data collected with the adapted scales?
3. How do various aspects of the affective domain compare when students complete a general chemistry experiment virtually as opposed to a traditional hands-on experiment?
4. To what degree can experiments in a virtual environment meet the same learning objectives as similar traditional hands-on experiments?
5. To what degree do student characteristics explain differential outcomes in a general chemistry laboratory course?

Human Subject Research

All parts of this dissertation that involved human subjects received institutional review board (IRB) approval.

Phases of Data Collection and Analysis

The data was collected and analyzed over several phases with the results of each phase informing the next phase as outlined in Figure 3.1 below. The first phase of the project was to decide which concepts the virtual experiments should cover. Once the experiments were chosen, the second phase consisted of interviewing faculty members at nearby institutions about what learning goals they have for the chosen experiments. These learning goals were then used to measure student's learning in the later phases of the data collection. There were two phases of piloting the selected experiments before the full implementation phase took place. The first phase of piloting took place one-on-one with

students coming in and completing the virtual experiment with the researcher to get a better understanding of how the designed procedures were working. Additionally, the students were asked to complete affective items after completing the experiment to pilot test how the items were working. The experiments were then beta-tested by having individual sections of 24 students complete one of the selected virtual experiments and completing the affective items. These students were observed while they completed the laboratory to further refine the procedures. These observations, as well as initial data analysis from the data collected during beta-testing, informed the full implementation. For the full implementation, the selected virtual experiments were implemented in half of all possible sections being offered in any given term. The other sections served as a reference group to compare the virtual experiment results to. The results from the full implementation can then inform what possible benefits and drawbacks exist for virtual experiments across a range of selected experiments.

Phase One: Experiment Selection

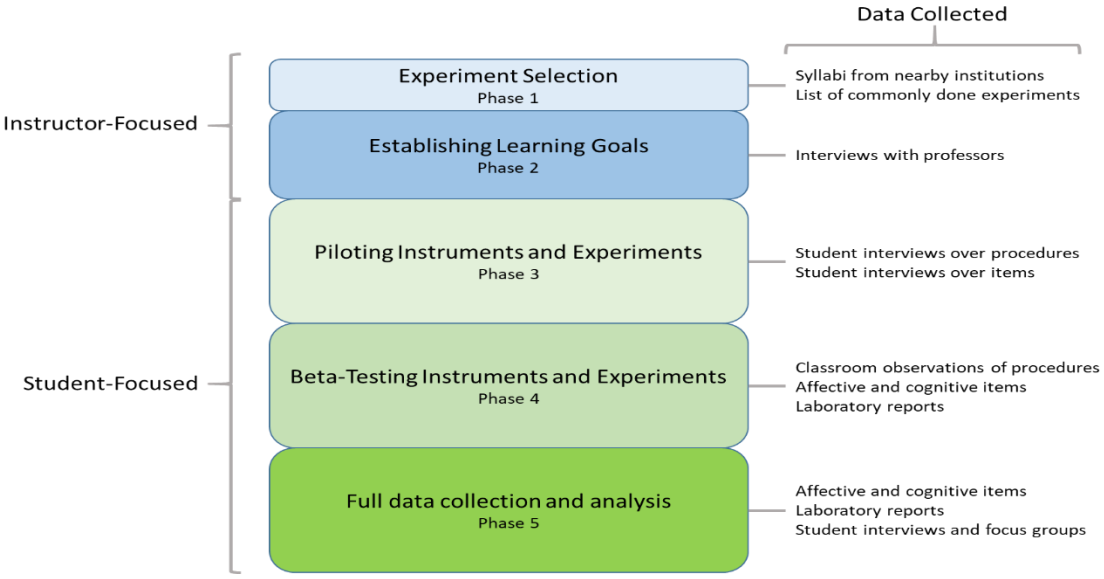


Figure 3.1: Outline of the five phases of data collection

Learning goals can span across the three domains of learning. The affective domain is broadly defined as the domain that emphasizes a feeling tone, an emotion, or a degree of acceptance or rejection (Krathwohl et al., 1964). The cognitive domain is broadly defined as the domain that emphasizes mental skills and content knowledge (Bloom et al., 1956). The psychomotor domain is broadly defined as the domain that emphasizes manual or physical skills (Simpson, 1971). Professors may have experiment-specific goals that span all three domains or goals that only cover one domain, depending on what the specific experiment emphasizes or what their preference is. For example, some experiments have laboratory techniques that students are focusing on whereas other experiments have minimal techniques and the primary focus is conceptual understanding. To adequately cover different possible types of experiments, a range of experiments were chosen.

The experiments were selected based on two criteria. The first criterion was that the experiment needed to be a commonly completed experiment in a typical general chemistry laboratory sequence. This criterion allowed professors from multiple institutions to be interviewed since, even though they may have slightly different procedures, they all share a basic understanding of how that type of experiment works. Data was obtained from Dr. Deb Exton (2016) at the University of Oregon to determine what the most common experiments performed were. As part of the development of the ACS laboratory exam, Dr. Exton and Dr. Reeves (2014) analyzed 37 laboratory manuals from a variety of institutions and found that the top seven experiments conducted in the general chemistry series were:

1. Volumetric Analysis (titrations)
2. Stoichiometry
3. Kinetics
4. Spectrophotometry (Beer's Law)
5. Properties of Acids and Bases
6. Calorimetry
7. Gas Laws

The second criterion was that the type of experiment needed to be done during back-to-back weeks. This criterion allowed for the possibility of a crossover design for the implementation of virtual experiments in the curriculum. A crossover design is a 2 by 2 design that allows differences at the individual level and temporal order to be assessed (Grant, 1948). In this design, half the students completed an experiment virtually one week and then a hands-on experiment the second week and the second half of the students did the opposite. This design is summarized in Table 3.1.

Table 3.1: Crossover design

	VIRTUAL ENVIRONMENT	**HANDS-ON ENVIRONMENT**
FIRST EXPERIMENT	First half of students (Group A)	Second half of students (Group B)
SECOND EXPERIMENT	Second half of students (Group B)	First half of students (Group A)

To identify potential experiments meeting the previously defined criteria, laboratory syllabi from a Pacific Northwest higher research university (R2, University 1), a Pacific Northwest highest research activity (R1, University 2), and nearby community colleges were analyzed (2015). These institutions were chosen using convenience sampling, which is that they were chosen because of the relationships already established during the interview phase of this research (Farrokhi & Mahmoudi-Hamidabad, 2012). At University 1, the syllabus for the first term of the general chemistry laboratory sequence revealed that there were two experiments conducted over the interaction of light and matter

(Beer's Law). In the second term, there were two experiments conducted over calorimetry and two experiments conducted over gas laws. In the third term, there were two experiments conducted over titrations. The same instructor has been teaching the course for the last decade and thus the experiments do not change much from year to year. At University 2, Beer's Law and titrations were the only two topics that had two experiments completed back-to-back weeks. One of the local community colleges (Community College 1) completed back-to-back experiments over gas laws and calorimetry and the other community college (Community College 2) completed back-to-back experiments over calorimetry and Beer's Law, as summarized in Table 3.2. With the primary focus being on the primary institution (University 1), all four experiments were selected to interview faculty members over.

Table 3.2: Outline of experiments completed at the selected institutions

	BEER'S LAW	CALORIMETRY	GAS LAWS	TITRATIONS
UNIVERSITY 1	x	x	x	x
UNIVERSITY 2	x			x
COMMUNITY COLLEGE 1		x	x	
COMMUNITY COLLEGE 2	x	x		

Phase Two: Establishing Learning Goals

One question that remains as a gap in the literature on the efficacy of virtual experiments is if students completing a hands-on experiment differ from students completing a virtual experiment on their ability to meet learning goals set out by the faculty in charge of the laboratory course. A learning goal, also sometimes referred to as a mastery goal, is broadly defined as the ability to demonstrate one's competence (Dweck, 1986). More generally, a learning goal is an outcome that the professor wants the students to master after completing an experiment or completing the course. While there has been

previous work done on course-based learning goals for the laboratory (Bretz et al., 2013; Bruck et al., 2010), little work has been on done on experiment-specific learning goals. An example of a course-based learning goal for the general chemistry laboratory is that 'students will be able to learn laboratory techniques' whereas an example of an experiment-based learning goal is that 'students will be able to successfully perform an acid-base titration using an end-point indicator'. In this example, the experiment-specific goals allow the professor to narrow the exact techniques they want the students to master rather than the broad goal of general laboratory techniques. Without experiment-specific goals, a student completing an experiment in a virtual environment and a different student completing the experiment in a hands-on environment may both meet the overall course learning goals equally but may not meet the goals for that specific experiment equally. Thus, there is a need to focus on exactly what professors want students to master at the experiment level before it is possible to compare students across institutions or types of laboratory environment. Faculty members in charge of the laboratory were interviewed to capture what specific learning goals they have.

Participants

The participant pool for the learning goal interviews was created using purposive expert sampling. Purposive expert sampling is when the participants are selected for being knowledgeable in their field and is one of the most common sampling techniques when qualitative research starts without a lot of background literature support (Etikan et al., 2016). Any faculty member that has previously taught the laboratory could be considered an expert on what they want their students to gain from specific experiments. However,

due to time and accessibility, faculty members at the home institution, a Pacific Northwest higher researcher university (R2), and at nearby community colleges were included in the participant pool for expert interviews (Carnegie Commission, 2015). These institutions all teach with different curricula that cover the same content of experiments. Therefore, the experiment-specific learning goals will be compared to measure how generalizable across curricula they are.

Data Collection

Once the experiments were selected, expert interviews were conducted with the faculty members in charge of the general chemistry laboratory sequence at their institution to understand their specific learning goals. Expert interviews are a qualitative technique that is commonly used in chemistry education research (Bretz et al., 2013; Bruck et al., 2010; Mutambuki & Fynewever, 2012) that involves interviewing an individual in the field to gain insight into the topic in which they are the expert (Bogner & Menz, 2009). Since there was specific information that the interviews were targeting, a semi-structured approach was used. A semi-structured interview approach is when the interviewer has specific questions they are interested in knowing about and use natural follow-ups to gain more information based on the participant's response (Wengraf, 2001) and has been frequently utilized to address similar questions within similar research. (Abell & Bretz, 2018; Canpolat et al., 2006; Nakhleh, 1994). The faculty members participating in this research were interviewed about the selected experiments that occur throughout the year of the general chemistry sequence.

Interview Protocol

The semi-structured interviews took place by asking the professors guiding questions with natural follow-up questions based on their responses. Examples of guided questions that professors were asked are noted below, for a full interview protocol see Appendix B.

1. Describe what students do in this procedure [referring to each specific experiment].
2. What learning goals, or things you want your students to get out of this lab, do you have?
3. If a student was sick and missed this experiment, what would they miss?

Since professors have differing levels of comfort with the term learning goals, the third guiding question was used as an alternative way of elucidating their learning goals. This question allowed professors not as familiar with learning goals to think about the important parts of the individual aspect that students show evidence of mastering after completing the experiment. Based on their answers to the three guiding questions, natural follow-up questions were asked to clarify and elaborate on their answers, such as asking the professor to elaborate on why they had those specific learning goals for the experiment.

Data analysis

The audio-recorded faculty interviews were transcribed verbatim to assist in the data analysis process. The transcripts were then open coded to determine the experiment-specific goals each professor had expressed for each experiment they were interviewed over. Open coding is when the researcher analyzes the transcripts without letting any preconceived idea influence the coding and instead looks for thematic elements to appear naturally (Strauss, 1987). This open coding process was done by the primary researcher to generate a list of experiment-specific learning goals that each professor had. The lists of

each professor's experiment-specific goals were further analyzed to determine what commonality between professors existed. This generated a second list of the shared experiment-specific learning goals between all professors. Generating these lists allowed for student laboratory reports to be analyzed on whether the student meets their professor's experiment-specific learning goals.

Phase Three: Piloting Instruments and Experiments

After establishing experiment-specific learning goals, the virtual experiments could then start to be piloted to measure what differences may exist in both the cognitive and affective domain. Before the affective domain could be measured in pilot studies, the constructs of interest first had to be determined and appropriate scales selected.

Selecting the Virtual Experiment Platform

There are many different platforms for virtual experiments at the general chemistry level such as the Second Life® virtual world (LindenLab, 2003), Late Nite Labs (Macmillan Learning), LearnSmart Labs (McGraw-Hill Education), PhET (University of Colorado), the Chemistry Collective (Carnegie Mellon University), and many more. The platform that was ultimately used was narrowed down from the possible platforms based on how much of the environment was replicated in the virtual environment. Some platforms, such as the PhET simulations, do not replicate all aspects of the laboratory but instead focus more on conceptual understanding. Other platforms, such as Second Life, focus on an immersive experience that replicates as many aspects of the physical environment as possible. Then there were also platforms that fell in between those two extremes. For example, LearnSmart Labs incorporate many aspects of the environment but

leave out other aspects to simplify the interface for the students. Ultimately, this middle platform was selected as the experiments needed to mirror the physical environment but also be simple and clear. McGraw-Hill Education generously provided access to the LearnSmart Labs for this research.

Designing the Virtual Procedures

LearnSmart labs have many built-in features available for users to use including an electronic notebook and a premade procedure for the laboratory. However, these procedures were not sufficient to use for this research as they were not as closely aligned with the traditional hands-on laboratory as possible. The traditional hands-on laboratory procedure for each of the selected experiments was used as the starting place for the virtual experiment. From there, the procedures were modified to adapt the experiment to the virtual environment and what chemicals were present. For example, the traditional hands-on procedure called for students mixing phosphoric acid and sodium hydroxide in a calorimetry experiment whereas the LearnSmart Lab only had hydrochloric acid available for the calorimetry experiment. The formatting and as many steps as possible were consistent between the procedures in the two environments to minimize any potential differences that could cause one environment to have a different experience than the other.

The initial procedures that were developed for the virtual experiments were tested and further refined to maximize how aligned the two procedures were, as well as for clarity, in Phases 3 and 4 of this dissertation before the final versions were implemented in Phase 5.

Measuring the Affective Domain

The affective domain is more broadly defined than the cognitive domain and covers many different aspects including constructs such as anxiety, motivation, interest, etc. A construct is a postulated attribute of a person that cannot be directly measured (Cronbach & Meehl, 1955). Therefore, indirect measures are used as a proxy of the underlying construct. For example, rather than hooking up a probe to someone's brain to measure anxiety it is instead measured by asking the participant items that are predicted, or shown in previous work, to relate to their level of anxiety. However, researchers have to be selective with how many constructs they try to measure at the same time because people, especially students, can experience testing fatigue. Testing fatigue is when the person is given a set of questions that is too long and results in answers that are not as valid (Kendall, 1964). For example, rather than thinking about each question thoughtfully, students may just repeat the same answer for many questions in a row. Therefore, researchers must limit the number of affective constructs they measure in any individual study to ones they are most interested in studying.

Selecting affective constructs and items

The constructs for this study were selected to address three areas that may contribute to benefits and drawbacks between the two environments. The area that was addressed was if pre-existing attributes exist that could cause students to perform differently in the two laboratory environments. As each student enters the laboratory with a different background and having different attributes, it is possible that there are constructs that help to explain their performance in the laboratory. One of these constructs that can

explain the students' performance is interest in chemistry and the laboratory overall. If a student is highly interested in chemistry, and in coming to the laboratory, then they are more likely to perform better than someone who has little to no interest in these things. Interest has been shown to be a significant predictor of performance in many disciplines (Harackiewicz et al., 2008; Rotgans & Schmidt, 2011) and specifically chemistry (Ferrell et al., 2016). Therefore, interest was chosen as a construct that could help explain why students perform differently in the laboratory.

Specifically, initial interest was chosen because this is the interest that students enter the classroom with. One way that initial interest has been measured in classrooms is through a set of items that Harackiewicz et al. (2008) developed. These items were later adapted for use in chemistry (Ferrell & Barbera, 2015) and shown to have similar psychometric properties when applied in a new context. Therefore, the initial interest set of items, as adapted by Ferrell and Barbera, were administered to the students at the beginning of their participation in the study to capture their initial interest. For a list of the items used, see Appendix A.

The next area that was addressed examined potential benefits and drawbacks that exist because of the experiment that took place. To examine this, it was important to understand what the goals of the laboratory were in the first place. One stated goal (Hofstein, 2017) of the laboratory is to increase students' attitude toward science. An existing instrument, the Attitudes toward the Subject of Chemistry Inventory (ASCI) (Bauer, 2008; Brandriet et al., 2011; Chan & Bauer, 2014), has scales that are well aligned with this goal. It has the scales of "Emotional Satisfaction", which measures how satisfied

the students are emotionally, and "Intellectual Accessibility", which measures how challenged the students felt. The ASCI was revised in 2011 to have better psychometric properties and be more aligned with a theoretical framework (Xu & Lewis, 2011).

In addition to determining if students find virtual experiments less challenging and are less satisfied performing them, previous literature has documented that anxiety is more prevalent in chemistry laboratories than other environments (Bowen, 1999). Therefore, it is possible that the anxiety levels between students completing the experiment in a virtual environment and a traditional hands-on environment are drastically different since they have different sources of anxiety present. An instrument, the chemistry laboratory anxiety instrument (CLAI), was developed previously to measure the various component of students' anxiety levels in the laboratory (Bowen, 1999). Additionally, the original ASCI also had a general anxiety scale. Therefore, the items from both the CLAI and ASCI were analyzed for similarities and new semantic differential items were developed based on the existing scales to measure students' general anxiety in both the virtual and traditional hands-on experiment. See Appendix A for a copy of these scales.

The third area that was focused on was examining the differences that exist because of the environment the laboratory was performed in and not necessarily the experiment itself. There are specific aspects of the affective domain that are context-specific. Therefore, it was of interest to ask questions that captured specific affective aspects in each environment to be able to directly compare aspects of students' affective domain between the virtual and hands-on environment. For example, if not requiring students to wear proper personal protective equipment (PPE) in the virtual environment changed their perception

of the experiment. A previous cross-over design study used existing items as well as developed their own set of paired questions that captured some of the expected affective differences between the two environments. Specifically, the questions covered 'Equipment Usability', the 'Usefulness of Lab', the 'Usefulness of Computers', 'Anxiety towards Computers' and the 'Open-Endedness of Lab' as these are possible differences between the environments that the researchers hypothesized. These five scales make up the Virtual and Physical Experimentation Questionnaire (VPEQ) (Pyatt & Sims, 2012). This questionnaire was developed and initially administered in a similar cross-over design experiment to capture the specific differences between the two environments making it a very good fit for this study (Pyatt & Sims, 2012). However, the original instrument had thirty-six items, to limit testing fatigue, only items that directly or indirectly compared the two environments were used. If an item indirectly compared the two environments, a new item was paired with the question to make the comparison explicit. After narrowing down the original thirty-six items, there were eighteen items covering nine direct comparisons between a virtual environment and a traditional hands-on environment. The new items from the 5 scales will be referred to as the VPEQ items. See Appendix A for a copy of the items used.

These chosen scales and items allowed for targeted aspects of the affective domain to be captured without the students being overwhelmed and fatigued by the number of questions they were asked. Overall, students were asked questions about their 'initial interest', 'anxiety', 'emotional satisfaction', 'intellectual accessibility', and the VPEQ items covering specific affective differences between the environments.

Initial wording changes

Minor wording modifications were made prior to initial piloting to some scales to better fit the application of the project. The first major change was to the emotional satisfaction and intellectual accessibility scales. The original versions of these scales used chemistry overall or the classroom environment as a frame of reference. Since this research took place in the laboratory, the frame of reference was changed to better fit the laboratory environment. Therefore, the frame of reference used for the equipment usability and intellectual accessibility scales was changed from "Chemistry is…" to "This experiment was…".

Participants

Students were recruited through purposive sampling to participate in piloting of the virtual laboratories and the scales and items chosen for the project. The students were recruited out of the general chemistry laboratory course at University 1, Community College 1 and Community College 2. These universities were chosen because of the possibility of implementing a large-scale study in the future at these institutions.

Think-aloud Interviews

Participants were asked if they were willing to participate in an interview. This interview followed a think-aloud protocol, where the students are asked to explain their thought process to the interviewer throughout a process. For each interview, the student met the interviewer and first completed a virtual experiment. During the virtual experiment, students were asked to think aloud about their initial impression and decision making when working with the experiment. Then following the experiment, the students were given all

of the chosen items and then after completing the items were asked to explain why they selected the responses that they did. This interview process allowed for the opportunity to see how the students were interpreting all the items and test the virtual environments on a small scale. Further item modifications were made after seeing how students were responding to the items.

Further wording changes

After conducting the think-aloud interviews, it became evident that further wording modifications were needed. The VPEQ items were initially administered on a frequency-type response scale ranging from 'almost never' to 'very often'. However, in the interviews, it was clear that the students were struggling to place their responses on a frequency scale of how many times during the virtual experiment the statement happened. Therefore, the responses were piloted a second time using a Likert-type response scale from 'strongly disagree' to 'strongly agree'. In this form, students provided more consistent answers and had an easier time responding to the Likert-type scale. Additionally, all students struggled to answer a pair of items from the 'usefulness of lab' scale. The item read "The _____ experiment provided me opportunities to pursue my own experimental interests". However, students did not understand what it meant by experimental interests and were giving inconsistent responses on their interpretation of this question. Therefore, it was not included moving forward when the scales were distributed in the pilot study data collection.

Phase 4: Beta-testing instruments and experiments

Participants:

Participants for this phase were chosen based on purposive sampling. Before implementing virtual experiments across all sections offered at University 1, the selected virtual experiments were beta-tested in select sections. Each of these sections completed only one of the selected virtual experiments. Students enroll randomly into various sections based on their preferred day and time and therefore did not know that their section would be doing a virtual experiment. However, the section was chosen based on the TA that was teaching the class as the more senior TAs had the experience of teaching similar labs in previous terms.

Affective data collection

The scales and virtual experiments were piloted after being refined through think-aloud interviews. This pilot study data collection consisted of one graduate teaching assistant (TA), per selected experiment, at University 1 implementing the crossover design (Table 3.1 from above) in their two sections. The laboratory course was offered twice a year meaning that for the pilot data collection each selected experiment was implemented twice as a crossover design. Students were administered the initial interest, emotional satisfaction, intellectual accessibility, and anxiety scales using paper and pencil immediately after completing the first experiment in the crossover design. The next week, students completed the second experiment in the crossover design using the environment that they did not use the previous week. Then, after completing the second experiment, students were administered the emotional satisfaction, intellectual accessibility, and

anxiety scales using paper and pencil again and were also given the VPEQ items. The initial interest scale was only administered at the beginning of the data collection to capture the incoming level of interest students had but was not administered a second time because their initial interest is not expected to change over the course of a week. The VPEQ items were administered only after finishing both experiments since these scales have paired questions that directly compare the two environments. The emotional accessibility, intellectual accessibility, and anxiety scales measure general affective aspects and therefore were administered after the experiment completed each week. This allowed these scales to be used to compare overall changes in affective aspects within a person across the two environments. The scale administration within the crossover design is summarized in Table 3.3.

Table 3.3: Scale administration throughout crossover design

	VIRTUAL ENVIRONMENT	**HANDS-ON ENVIRONMENT**	**SCALES ADMINISTERED**
FIRST EXPERIMENT	First half of students	Second half of students	Initial Interest, Anxiety, Emotional Satisfaction, Intellectual Accessibility
SECOND EXPERIMENT	Second half of students	First half of students	Anxiety, Emotional Satisfaction, Intellection Accessibility, VPEQ Items

Cognitive data collection

Students completed two types of laboratory reports based on which selected experiment was completed. For half of the experiments at University 1, the students completed brief worksheets with their laboratory partner that were due before leaving for the day, whereas for the other half of experiments the students completed formal laboratory reports on their own that were due one week after the experiment ends. The students turned

in their report to a learning management system regardless of which type of report, a worksheet or a formal report, they completed. The reports pertaining to the selected experiments for the students who consented to be a part of the study were then downloaded and stored for further analysis. These reports were then analyzed using a rubric generated from the list of experiment-specific learning goals to determine how well students met the desired goals.

Affective data analysis

After collecting responses to the eight scales that the students answered at the various time points outlined in Table 3.3, the data were analyzed to determine if they were functioning appropriately. One way of determining if the scales were measuring the intended constructs is by performing a confirmatory factor analysis (CFA) and examining how well they fit hypothesized models. A CFA has the ability to determine how well indicator items are informing a latent construct (Kim & Mueller, 1978). A CFA is chosen over an EFA when the scales are previously developed because there is an established model that the items are hypothesized to fit. Indicator items are the questions that we ask students to indirectly measure the construct. For example, if we are interested in the students' anxiety, we ask them indirect questions to measure their anxiety levels. These questions can then be put into a confirmatory factor analysis model to determine how well they identify the latent construct of anxiety. An example of a confirmation factor model is shown in Figure 3.2.

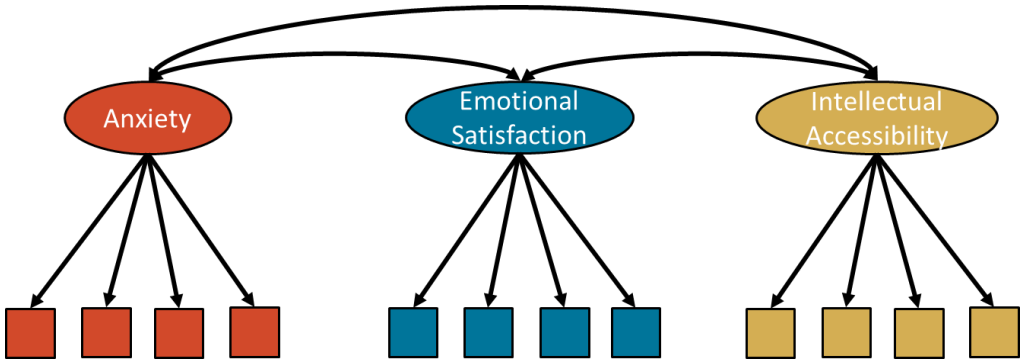


Figure 3.2: CFA model of ASCI scales

In this example, the three constructs measured by the ASCI and ASCIv2 are shown in the ovals. An oval in a CFA model represents a latent construct as it cannot be directly measured. Instead, the construct is measured with indicator items as shown by the squares under the latent construct. The arrow between the construct and item represents how well that item identifies the construct. Since the three scales have been shown in previous research to be correlated with each other, there is the need to also add correlation arrows between the constructs, shown as the double-headed arrows. Together this represents the hypothesized model for how anxiety, emotional satisfaction, and intellectual accessibility are related.

To test this model and run a CFA, It is generally suggested that there is a need for at least 200 responses (Marsh et al., 1988). Above 200 responses there starts to be enough variation in the responses that enough information about the fit of the model can be calculated. Therefore, after administering the scales in 16 sections of 24 students per section there were enough students in the sample to run one-factor CFA models for each of the scales that had at least four items on the scale. A one-factor CFA model is similar in structure to the model shown in Figure 3.2 except only a single construct is represented.

For example, the one-factor CFA model for anxiety consists only of the items for anxiety and how well they measure the latent construct of anxiety. These CFA models were used to determine how well the indicator items are informing the latent constructs they represent and to determine if any changes to the scales are needed. The CFA output contains information in the fit statistics, various statistics that indicate how well the data fit the hypothesized model, on how well the indicator items are informing about the same latent construct. There are many different fit statistics that can be calculated from the output of a CFA model but the most commonly reported are the comparative fit index (CFI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA). However, the RMSEA is known to be less informative for models that have low degrees of freedom such as the one-factor CFA models for each scale (Kenny et al., 2015). These scales have low degrees of freedom due to having only either four or five indicator items per scale. Thus, the CFI and the SRMR were used to determine how well the indicator items were informing the same latent construct.

Previous simulation work has been done to determine what general numeric values for the CFI and SRMR are deemed to provide acceptable fit (Hu & Bentler, 1999). Although they only ran specific models in their simulation study, the values provided by Hu & Bentler have been tested in a wide range of other models and held up equally well (Miyake et al., 2000). Therefore, if the CFI is above 0.95 and the SRMR is below 0.08 the data fit the proposed model, indicating that the items are adequately fitting together to inform the same latent construct. Further investigation about the model was warranted if the fit indices did not indicate adequate fit. Modification indices are one way of examining

why there is not adequate model fit. A modification index indicates how well the model fit would be improved if another path in the model was added (Steiger, 1990). For example, it may suggest that two of the indicator items have correlated errors and it is necessary to add that path to the model to achieve adequate fit. However, the researcher must also use a theory-driven approach to see if the modification makes sense in the model rather than just modify the model until achieving adequate fit (Schreiber et al., 2006). If there is a theoretical reason to add the path then the researcher can modify the original model and determine if the new model has adequate fit.

This process of running CFA models and examining the fit indices was performed on all the individual scales that had at least four items before administering the scales in a large data collection. The CFAs were run and analyzed using the Laavan package (version 0.5-23.1097) in R Studio. Running CFAs and examining fit statistics ensured that the scales that were administered in the academic year of the full-scale data collection had evidence for construct validity. Construct validity means that the scales are measuring the construct that they are meaning to measure and is supported with two pieces of evidence. The first is that the students responded in the think-aloud interviews in a way that was consistent with the construct. The second is that the items are all indicating the same construct to an adequate degree and are not multi-faceted.

A power analysis is needed after determining the functionality of the scales to see if anything meaningful can be said with the sample size obtained through the pilot data collection. Each experiment was implemented twice in the academic year with each implementation containing two sections of up to 24 students per section. Therefore, if all

sections were full and all students consented, the most data collected is 48 students per environment to compare. A power analysis was run, using G*Power (version 3.0.10.) for a two-tailed t-test with an alpha level of 0.05 and a power level of 0.8. This analysis determined that only large effects were able to be detected with 48 students per group. The results of the power analysis are summarized in Table 3.4.

Table 3.4: Power analysis results

	Sample Size per Group Required
Small effect	394
Medium effect	64
Large effect	26

Cognitive data analysis

To assess if their learning goals were met, all faculty members that were interviewed used some form of a laboratory report that students completed after finishing an experiment. These laboratory reports ranged from worksheets to formal laboratory reports. The worksheet format used at University 1 consisted of the students completing a short introduction about the conceptual aspects of the experiment and then answering brief directed questions about their results and conclusions. These were typically used when the students were introduced to a concept for the first time. On concepts that students have experience with, or are more complicated to understand, formal reports were frequently used at the primary institution. A formal report is designed to mimic journal-style articles and have students begin thinking about formal scientific writing (Hofstein, 2004). The student reports were analyzed for how well the students met the learning goals for the

specific experiment regardless of the format of the report since the report is the only assessment of the learning goals used at all institutions involved.

Generating the rubrics

The rubrics used to analyze the reports were generated by using the common learning goals that all professors expressed as the primary focus. An individual instructor's goals were only added if all professors within the same institution agreed that it was an important learning goal. This captured learning goals that were institution-specific but left off learning goals that were unique to individual professors. Generating the rubric in this manner ensured that multiple professors shared the learning goals assessed in the reports and thus increases the likelihood that the findings could be generalized to other professors and institutions. Therefore, the rubrics for each experiment were generated by listing the common learning goals on separate lines of the rubric. See Table 3.5 for an example of what a sample rubric looked like.

Table 3.5: A generic rubric

	DID NOT MEET	**MEETS**
LEARNING GOAL 1		
LEARNING GOAL 2		
LEARNING GOAL 3		

This generic rubric was then filled in with the learning goals that professors had for a given experiment. For example, the common learning goals professors had for a Beer's Law experiment were listed in place of 'Learning Goal 1, 2, 3, etc'. This turned the rubric from the generic rubric in Table 3.5 to an experiment-specific rubric seen in Table 3.6.

Table 3.6: Experiment-specific rubric (Beer's Law)

	DID NOT MEET	MEETS
UNDERSTAND THE RELATIONSHIP BETWEEN ABSORBANCE AND CONCENTRATION		
PREPARE SOLUTIONS		
DETERMINE AN UNKNOWN CONCENTRATION USING THE RELATIONSHIP BETWEEN ABSORBANCE AND CONCENTRATION		

Student laboratory reports were then assessed with the experiment-specific rubrics. For example, if a student wrote in their report that there was a linear, or directly proportional, relationship between absorbance and concentration they were marked as meeting this learning goal. If a student did not mention anything about Beer's Law or the relationships it explains, they were marked as not meeting the learning goal. This process was carried out for all the learning goals of a specific experiment using the experiment-specific rubrics in Appendix B.

Data analysis

Initial data analysis took place after the rubrics were established for each of the selected experiments. An additional researcher, a post-baccalaureate student, was added to the project to assist in the analysis of the student reports. An additional researcher was necessary to establish inter-rater reliability. Inter-rater reliability is defined as the ability for two (or more) independent coders to classify the same object into the same set (Gwet, 2014). In other words, it allows for two independent researchers to assess the level of agreement on which category to mark for the students meeting the learning goal. Having

another researcher also allowed for discussions to take place regarding the clarity of the rubric and learning goals.

The two researchers initially coded a random sample of ten reports per experiment with five of the reports from the students completing the experiment in the hands-on environment and five of the reports from the students completing the experiment in the virtual environment. The two researchers then met to determine the level of initial agreement for the ten reports. If this initial agreement was below 100% for any individual learning goal, the researchers discussed the disagreements. After clarifying the disagreements, the researchers analyzed another seven randomly selected reports and met again. This process continued until the two researchers agreed 100% of the time. A commonly used minimum agreement for the field of chemical education is 70% (Buck et al., 2008; Fay et al., 2007) making the threshold of 100% more than acceptable. Another method of calculating the inter-rater reliability is Cohen's Kappa. (Cohen, 1960) However, when agreement is reached 100% of the time, Cohen's Kappa is equal to percent agreement with a value of 1.

The researchers then coded the remaining reports for each experiment independently after they met the 100% threshold. The researchers met throughout the process to discuss any reports that they were not completely sure about. This process generated a numeric score that represented the level to which the student met the desired goals. For example, if a student fully met the learning objective their report was assigned a score of 1 whereas if they partially met the goal it would. This score was then used in further analysis as a representation of the student's cognitive domain. This was chosen

since the cognitive domain is the domain of content knowledge and the student reports are the primary way students demonstrate their content knowledge at all institutions involved.

Phase 5: Full Data Collection and Analysis

After completing a full year of pilot data collection and analyzing the data, several major changes were made based on two findings. The first finding that came out of the pilot data analysis was that the back-to-back experiments were not as parallel as planned. Although they were covering the same topic (i.e., acid-base titration) the students had very little recognition that they were similar experiments and forgot any skill and conceptual knowledge they learned the week before. Therefore, for the full data collection, the research design was simplified to only looking at one experiment per term rather than continuing with the crossover design. The second finding was that the students were better able to demonstrate their content knowledge when they write a formal report than completing a worksheet. Thus, the three selected experiments were converted to formal reports if they did not already include one. In addition to these findings, there was a need to add to the depth of the information gained from the affective items by adding focus groups and interviews. This was based on the results of phase 4 data analysis. This data analysis showed mixed findings between the two environments and the addition of focus groups and interviews allow for targeted questions about the students' experience to be asked.

Participants

For phase 5, all sections being offered of general chemistry at University 1 were included in the participant pool and therefore all students enrolled in the general chemistry

laboratory course were also included. The sections were selected based on the room they took place in to complete either a selected virtual experiment or the traditional hands-on counterpart. The students in the participant pool were invited to volunteer to be interviewed or participate in a focus group to elaborate on their answers to the items they completed in the survey.

Focus Groups and Interviews

After completing the data analysis for phase 4, it was clear that there was more information to be gained than the survey items could provide. Therefore, in phase 5 students were invited to participate in both focus groups and one-on-one interviews. Focus groups were chosen to capture students' conversations with each other. Focus groups are often able to provide unique perspective since students will react to what other students say in addition to the planned research questions. Students also had the option to volunteer to be interviewed instead of coming to a focus group. While interviews lose the interaction between students, they allow an individual's experience to be well documented. These experiences can give depth to what individual students see as the benefits and drawbacks to the environment in which they completed the selected experiments. All students who were enrolled in the laboratory course were offered to come in for both a focus group and/or an interview.

Both the focus groups and interviews were largely unstructured with the goal of having students elaborate on their experience during the selected experiment. The students already had completed the survey items so there was targeted questions for them to follow-

up on their survey item responses. These focus groups and interviews were then used in phase 5 data analysis to possibly explain some differences between the two environments.

Affective data collection

The affective data collection had to be modified from the pilot data collection because a second experiment was not being conducted. Instead of the students answering the items about both environments, the students were given all items electronically after completing the selected experiment in either the virtual or traditional hands-on environment. Additionally, the VPEQ items were rewritten as generic items rather than about the specific environment since the students no longer experienced both environments. For the new VPEQ items see Appendix A.

Cognitive data collection

The cognitive data collection was similar to the pilot data collection. The one difference was that the worksheets were eliminated and instead the students were asked to write formal reports for the selected experiments that use to have worksheets. The laboratory reports were turned in online via a learning management system and then were downloaded and stored with pseudonyms for future data analysis.

Affective data analysis

The full-scale data collection reached large enough sample sizes that it was possible to run CFA models that separated the data by experiment rather than lumping all the data together to reach a large enough sample. Therefore, the hypothesized models for the scales consisting of four or more items were analyzed using a CFA for each individual

experiment. For the models that showed adequate fit, further analysis was completed to determine what differences in the affective domain exist between the two environments.

For any experiments where initial interest was significantly different between learning environments, a multivariate analysis of covariance (MANCOVA) was ran after achieving model fit. A MANCOVA is used to test if there are differences in group means on multiple dependent variables while also accounting for covariates (Mertler & Reinhart, 2016). This allowed the two environments to be compared in all the affective constructs measured to and see if there are any meaningful differences while accounting for the initial interest in chemistry that the students had. For the experiments where initial interest was not significant, a MANOVA was used instead. The MAN(C)OVA tests were completed using SPSS version 24.0 after checking that all the assumptions for running a MANOVA are valid.

Cognitive data analysis

The rubrics generated from the pilot data analysis were used to code the laboratory reports for the full data collection. This was done by two coders analyzing all of the laboratory reports. The two independent coders coded seven reports independently and then met to establish percent agreement. This process continued until 100% agreement was reached. After it was reached, the independent coders coded all of the laboratory reports and met regularly to clarify any questions. After a rubric score and grade from the laboratory report was generated, the cognitive variables were added to the MANOVA with the affective variables to see if there are any differences between the two environments on both the cognitive and affective variables.

Analyzing student profiles

After classroom observations and interviews in phases 2-4, it was clear that there are groups of students that are experiencing the two environments in drastically different ways. For example, there was one student that is a single parent and needed to find childcare to come to the chemistry lab. This student rejoiced at the idea they could possibly complete experiments at home one day instead of coming in. There was also a student who wanted to work in the pharmaceutical industry that loved coming to chemistry lab every day and resisted the idea of virtual experiments. In addition to the large extremes, many students saw the laboratory as just another class to go through the motions and were not fazed in either environment. With such a varied response to virtual experiments, it was important to capture what student profiles were present in the laboratory and which environment they may be best served in.

In phase 5, student profiles were created by running latent profile analysis. Latent profile analysis is a model-based cluster analysis. The most important choice when doing a latent profile analysis is selecting which clustering variables to use. Since the rubric scores have little variation in scores, the affective variables that were significantly different between environments were used as clustering variables to create affective profiles. Analysis was completed in R version 3.3.0 using the package ‘mclust’. These clusters can then be used as a grouping variable in a MANOVA.

In addition to testing for differences using a MANOVA between environments, a MANOVA can also detect differences between the clusters within an environment.

Therefore, MANOVAs were ran within each environment to determine how the clusters differ on both the cognitive and affective variables. This provides more information on what profile of student made up each cluster and how the type of environment they completed the experiment in influences their affective state.

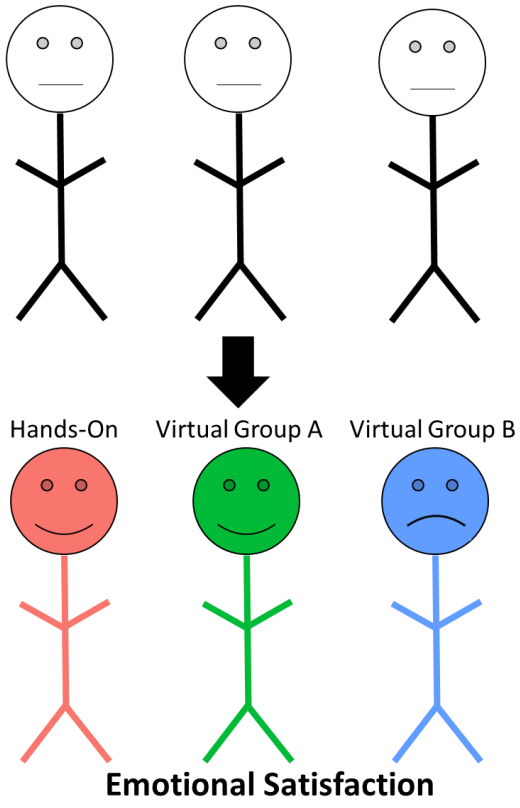
CHAPTER IV: Assessing Affective Differences Between A Virtual General Chemistry Experiment and a Similar Hands-On Experiment

ABSTRACT

To date, few general chemistry laboratory studies have included affective measures despite calls for more research on aspects of this domain. This shortage of studies may be partially due to the scarcity of affective measures that have been designed for, or tested in, the college laboratory setting. To provide measures for use in this environment, several existing affective scales were adapted for this new context. Before data from the scales were utilized to study the environment, evidence was provided for the validity and reliability of the data generated from them. Once sufficient evidence was provided, it was possible to determine affective differences between students completing a Beer's Law experiment in the traditional hands-on laboratory (control group) and a similar experiment in a virtual environment (treatment group). To assess expected differences between environments, scales for anxiety, emotional satisfaction, intellectual accessibility, usefulness of lab, equipment usability, and open-endedness of lab were selected. To account for potential between-student differences, scales for feeling-related initial interest and value-related initial interest were selected. Overall, students who completed the virtual experiment scored significantly lower on the emotional satisfaction, intellectual accessibility, usefulness of lab, and equipment usability scales. However, it was noted that student responses in the virtual environment varied significantly by which

teaching assistant (TA) instructed the section. To test for a possible instructor effect, data from the virtual sections were grouped by TA as 'Virtual Group A' and 'Virtual Group B'. Group A contained the TAs who had sections with lower averages on the emotional satisfaction scale as compared to group B. After controlling for instructor, differences between student responses in the 'Hands-On' sections and 'Virtual Group A' sections were no longer significant while significant differences remained between the responses in the 'Hands-On' and 'Virtual Group B' sections. This outcome indicated that the TA instructing the course may have been more influential on students' affective outcomes than the environment in which the experiment was performed.

GRAPHICAL ABSTRACT



KEYWORDS
General Public, Chemical Education Research, Computer-Based Learning, Distance Learning

INTRODUCTION

Over the past decade, there has been a steady increase in the number of students electing to take college classes from a distance, which typically involves taking online courses (Seaman et al., 2018). While this may provide an acceptable learning experience for courses in many fields, online courses pose a specific challenge to the laboratory component of the chemistry curriculum. Currently, the American Chemical Society (ACS) requires 400 hours of laboratory instruction for a student to earn an ACS certified bachelor's degree (ACS, 2015). This requirement highlights the common belief that laboratory courses are essential to an undergraduate chemistry degree. Thus, universities have sought out varying ways to include a chemistry laboratory experience for students who elect to complete courses in a non-traditional environment. One of the most common approaches to address this challenge is to offer a laboratory course that uses at-home kits (Boschmann, 2003; Hoole & Sithambaresan, 2003; Kennepohl, 2007). These kits allow students to be exposed to laboratory basics, such as glassware and reagents, in a space of their choosing. More recently, institutions have taken advantage of advances in technology to offer alternatives that rely on the use of computers.

The technologic approach can be categorized in two distinct ways. The first category includes experiments that incorporate the use of a remote laboratory environment. Remote laboratories involve a student using a computer interface to control an instrument that is housed at a different institution (Herranz et al., 2018; Ma & Nickerson, 2006). For example, a student controls a robotic system to put a sample into

an instrument to take a physical measurement and data from the instrument is then reported back to the student. This method allows students to collect real-time data from physical instruments without being physically present for the data collection. The second category includes experiments that simulate the entire process, including data collection. This approach, called a virtual laboratory environment, drastically reduces operating costs as there are no physical laboratory spaces, reagents, or instruments to maintain. There have been a number of different approaches to the virtual laboratory ranging from simple simulations (Clark & Chamberlain, 2014; Perkins et al., 2006) to fully-immersive environments (Winkelmann et al., 2017; Winkelmann et al., 2014) and multiple environments in-between (Hawkins & Phelps, 2013; Reece & Butler, 2017; Woodfield et al., 2005; Woodfield et al., 2004).

Regardless of which environment is chosen as an alternative to the physical laboratory, it is imperative that potential differences between environments be evaluated to ensure that students have similar outcomes to the students in the traditional hands-on laboratory across the three domains of learning (cognitive, psychomotor, and affective). Outcomes in the cognitive domain include aspects of the knowledge acquired in an experiment (Bloom et al., 1956), outcomes in the psychomotor domain include the skills acquired from the experience (Simpson, 1971), and outcomes in the affective domain include aspects of students' attitudes and emotions regarding an experiment or the learning environment (Krathwohl et al., 1964). The cognitive domain is frequently measured by administering content-based items such as prelab or postlab quizzes in a laboratory environment or test questions in a lecture environment. The psychomotor

domain is frequently measured in the laboratory setting with the use of laboratory practical exams that measure specific skills students are expected that have learned. The affective domain includes a wide range of constructs including motivation, emotion, interest, values, attitudes, and many more. Within each construct, there may be further subconstructs such as self-efficacy within the broader construct of motivation. These constructs can then be targeted for specific interventions and measured to determine whether an intervention has positive or negative impacts. In the college setting, it has previously been reported that motivation, self-esteem, self-perceptions, feelings of confidence, self-concept of ability, and teacher praise are some of the most important affective constructs to target (Hunt, 1987). More specifically, for the laboratory the construct of general attitude has been proposed as an important affective construct, as one of the goals of the science laboratory is to increase students' attitude toward science (Hofstein, 2017). Aspects of these three domains need to be measured to determine what, if any, differences exist between learning environments.

Differences Between Traditional and Alternative Environments

Several studies have been conducted to examine the advantages and disadvantages of incorporating virtual experiments in various curricula ranging from high school chemistry laboratories (Pyatt & Sims, 2012; Winkelmann et al., 2014) to upper-division college laboratories (Woodfield et al., 2005; Woodfield et al., 2004). These studies generally fall into two categories. The first category includes comparative studies that examine differences between a virtual and a traditional hands-on experiment using treatment and control groups. The second category includes studies that describe the

virtual environment used and the potential advantages and disadvantages they have without the use of a comparative group.

At the general chemistry level, Hawkins and Phelps conducted a comparative study with 84 students completing a virtual electrochemistry experiment and 85 students completing the equivalent traditional hands-on experiment (Hawkins & Phelps, 2013). The students completed a pretest and a posttest content knowledge quiz to determine if there were any cognitive domain differences between the groups. No statistical difference between the two groups was detected. As there were remaining questions if virtual experiments could provide the same psychomotor outcomes as a traditional hands-on laboratory environment, the researchers measured both groups' ability to complete a hands-on practical regardless of which environment they completed the experiment in and again no statistical difference was detected. While the researchers concluded that no differences were detected between environments in the cognitive and psychomotor domains, they did not measure any aspects of the affective domain.

A comparative study that focused on potential differences in the affective domain was conducted by Pyatt and Sims (2012) at the high school level. In their study, students were assessed using a newly created affective instrument, the Virtual and Physical Experimentation Questionnaire (VPEQ), that measured students' attitude towards various aspects of the two environments. This instrument measured aspects of usefulness of computers, anxiety towards computers, equipment usability, open-endedness of lab, and usefulness of lab. Within-person differences were controlled for by using a 2x2 Latin square (crossover) study design in which the 184 students completed both a virtual

experiment and a traditional hands-on experiment (Grant, 1948). Results indicated a higher average for the virtual experiment for the constructs of equipment usability and open-endedness of lab but no detectable difference between environments for the construct of usefulness of lab. The significance of these differences was not tested. In addition to these affective constructs, cognitive domain differences were measured by scoring laboratory reports. The first experiment was scored using a binary scale whereas the second was scored on a four-point scale. No significant difference was detected between groups in the first experiment but a significant difference was detected in the second experiment, with the scores of the students who completed the traditional hands-on experiment being significantly lower than the scores of the students who completed the virtual experiment. The scoring system for the cognitive assessments changed between experiments to allow for more resolution and thus could explain why significance was found for the second experiment but not the first. Future work should be done to elaborate on the affective and cognitive differences seen in this comparative study and how they compare to the collegiate level.

Grove and colleagues measured differences across all three domains of learning for a college hybrid curriculum that incorporated LearnSmart Labs by McGraw-Hill Education (Enneking et al., 2019). Rather than measure differences for a specific experiment, the differences measured were for the entire hybrid curriculum as a whole. In total, 195 students completed the hybrid curriculum across the 2015 calendar year. In this curriculum, students alternated between a virtual experiment and a traditional hands-on experiment. There were no statistically significant differences found when measuring

cognitive outcomes using the ACS General Chemistry Laboratory Assessment or psychomotor outcomes using a hands-on laboratory practical. However, the hybrid students had a significantly lower mean scale score on the affective portion of the Meaningful Learning in the Laboratory Instrument (MLLI), which has 8 items that measure a general affective state. Upon further examination of the items, it was revealed that students in the hybrid curriculum had lower affective aspects with the exception of worrying about completing the lab on time when compared to the hands-on students, however, it was not possible to determine what differences were specific to the virtual environment as the hybrid curriculum included both types of experiments and data was collected only at the beginning and end of the curriculum.

Irby and colleagues focused more narrowly within a hybrid curriculum to better understand if students who completed an electrical conductivity experiment in a virtual environment engaged with the chemistry triplet (Johnstone, 1982), which models the different levels of understanding in chemistry: submicroscopic, macroscopic, and symbolic, to a different degree than students who completed a similar hands-on experiment (Irby et al., 2017). Their study utilized a pretest-posttest alternative treatment with a control group study design (Mack et al., 2019) where there was a control section that did not use the hybrid curriculum and two treatment sections that did. The two treatment sections were staggered so that in any given week, one section was completing a hands-on experiment while the other section was completing a virtual experiment on a different topic. This allowed for the institution to offer three laboratory sections while only having two laboratory rooms available. However, this meant that the control section

and the two treatment sections had a different order of experiments over the course of the term. For example, the "leading" hybrid section completed the conductivity experiment in week seven whereas the "trailing" hybrid section and the hands-on students completed the experiment in week six. After comparing the sections, no statistically significant differences were found on the cognitive outcomes, as measured by pre- and post-assessments, nor students' use of the chemistry triplet. However, the authors note that this could be a function of the small sample size as the students who completed the experiment in the virtual environment appeared to connect between the triplet levels more often despite the difference not being statistically significant.

Winkelmann and colleagues (2017) used the virtual platform Second Life (SL) to measure differences across the three domains of learning for two specific experiments rather than a hybrid curriculum. In this study, 55 students completed both a gas law experiment and a titration experiment in the SL platform while 67 students completed both experiments in a traditional hands-on (control) environment. Both the control and the SL group showed cognitive gains on the post-quiz as compared to the pre-quiz, however, the SL group had significantly higher gains in both experiments. Within the psychomotor domain, students in both groups performed equally well on a follow-up hands-on laboratory practical. To measure differences within the affective domain, students were asked general questions authored by the researchers about their experience in the respective learning environment. These general questions were elaborated on in written responses and focus groups. Students reported that the SL experiments took less time to complete, they felt that they learned more in the traditional hands-on experiment,

and had higher perceived grades in the SL experiment as compared with previous hands-on experiments. Future research can expand on these findings by using affective instruments with measured psychometric properties rather than general self-authored questions.

The second type of commonly conducted study pertaining to alternative labs does not contain a traditional hands-on laboratory control group. Instead, in descriptive studies, the advantages and disadvantages of a specific virtual platform are investigated. Winkelmann and colleagues (2014) conducted a study on SL at the high-school level before conducting the comparative study at the college level. In this study, seven high school students completed a SL experiment as part of a summer class. The students were evaluated on aspects of their attitude using a modified version of the Inquiry Laboratory Attitudes Survey (Chatterjee et al., 2009). However, the scale was developed for inquiry-labs and had not been psychometrically evaluated for the new context. The students were also evaluated on their cognitive outcomes by assessing their laboratory report. The students were able to successfully complete the experiment, as indicated by their laboratory report score, and reported that they felt the SL experiment was shorter than their other experiments in the term. In addition to this study, descriptions of how SL can be used more broadly in chemistry as a discipline has been the subject of multiple articles (Lang & Bradley, 2009; Murray-Rust, 2008).

Woodfield and colleagues created a suite of virtual experiments including one as part of an inorganic laboratory curriculum (Woodfield et al., 2004) and one as part of an organic laboratory curriculum (Woodfield et al., 2005). These studies are descriptive in

nature and thus they did not have a control group. In their inorganic study, they found that students were able to use the virtual environment to complete experiments that would be challenging to complete in a traditional hands-on laboratory setting. For example, the students were given 26 cations and 11 reagents that they could combine, which would have required significant prep time for the traditional hands-on laboratory. Many of the students reported qualitatively that the virtual experiment helped them learn the content. Similarly, in their organic study, they found that the students who had a positive experience with the virtual experiment were more likely to have a higher course grade. Unlike the Hawkins and Phelps study, Woodfield and colleagues focused more on affective differences. They found that students in both the inorganic and organic study preferred the virtual experiment for the adaptability of the environment to perform experiments that are not as feasible to do in traditional hands-on laboratories. However, they used general Likert-type items to ask affective questions that were not rooted in any specific construct and also they did not measure any psychomotor outcomes.

Overall, the comparative and descriptive studies on virtual experiments in the chemistry curricula point to either no difference (Hawkins & Phelps, 2013; Winkelmann et al., 2014) or a slight potential benefit (Pyatt & Sims, 2012; Winkelmann et al., 2017; Woodfield et al., 2005; Woodfield et al., 2004) in the cognitive domain when students complete the experiment virtually. This is in line with studies measuring the cognitive domain when the alternative environment is an at-home kit (Casanova et al., 2006; Kennepohl, 2007) or a remote environment (Corter et al., 2011; Kennepohl et al., 2004; Scanlon et al., 2004; Sonnenwald et al., 2003). In addition to the cognitive domain, there

was no difference in the students' ability to perform the laboratory skill covered in the experiment for the studies that included a psychomotor measure (Hawkins & Phelps, 2013; Winkelmann et al., 2017). However, of the comparative and descriptive studies, only one study used an affective measure that has been psychometrically tested (Pyatt & Sims, 2012). The remaining studies either did not measure the affective domain, or used items that had unknown psychometric properties to measure the affective domain. This is a common challenge across all laboratory studies and not solely on studies involving alternative environments, as there have been few affective measures developed or adapted for chemistry laboratories.

Affective Domain Measures for the Laboratory

In response to a historically heavy focus on cognitive outcomes in research studies, the National Research Council has called for studies in discipline-based education research (DBER) to include the evaluation of outcomes within the affective domain (National Research Council, 2012). Despite this call, a majority of chemistry laboratories studies either provide limited scope to the affective domain or do not include it at all. A primary reason for this has been the lack of affective measures designed for and tested specifically in the laboratory. A recent instrument, the MLLI (Galloway & Bretz, 2015a), was designed to address this issue and it has allowed researchers to study how aspects of the affective domain in the laboratory change over time (Galloway & Bretz, 2015c) and based on an intervention (Flaherty et al., 2017; Schmidt-McCormack et al., 2017). Additionally, it has been used to categorize student profiles (Galloway &

Bretz, 2015d). Another example of an affective instrument developed specifically for the laboratory is the chemistry laboratory anxiety instrument (CLAI) (Bowen, 1999).

While not developed specifically for the laboratory, instruments to measure various aspects of the affective domain of chemistry students include the Metacognitive Activities Inventory (MCAI) (Cooper & Sandi-Urena, 2009), the Chemistry Expectations Survey (CHEMX) (Grove & Bretz, 2007), the Chemistry Self-Concept Inventory (CSCI) (Bauer, 2005), the Attitude toward the Subject of Chemistry (ASCI) (Bauer, 2008) and the revised version (ASCIv2) (Xu & Lewis, 2011), the Colorado Learning Attitudes about Science Survey (CLASS-Chem) (Adams et al., 2008), and the Academic Motivation Scale (AMS-Chem) (Liu et al., 2017). These instruments serve as potential measures for the affective domain in the laboratory setting. However, for any of these instruments to be used in studies of the laboratory environment their functioning in this new context would need to undergo psychometric testing to ensure that there was sufficient evidence of data validity and reliability.

Psychometric Testing

Whenever an instrument is distributed within a different setting than originally developed for, evidence for the validity and reliability of data from the instrument in the new context needs to be provided (Arjoon et al., 2013; Barbera & VandenPlas, 2011; Furr, 2017). For example, it is possible that questions asked in a classroom setting do not function in the same way in a laboratory setting. Thus, the internal structure of the instrument could be different and pose a threat to the validity. Another issue that could arise from adapting an existing instrument to a new context is that students no longer

interpret the questions as they were initially intended. For example, an item asking if the student "feels comfortable" may be interpreted in multiple ways. One way students may interpret the item is if they feel comfortable with big-picture concepts but alternatively, it may be interpreted is if they feel physically comfortable in the laboratory setting itself. Differences like this pose a threat to the response process aspect of validity (Schwartz & Barbera, 2014). Additionally, to accurately compare treatment and control groups with the same instrument, invariance testing needs to take place to show that both groups have the same relationship to the variables being measured (Milfont & Fischer, 2010). If evidence is provided that items are functioning in both contexts, it is possible to use the items to measure meaningful differences between groups.

Research Questions

Previous work on the differences between traditional hands-on and alternative laboratory environments has found very little or no difference on measures in the cognitive domain, however, there have been fewer studies on the differences in the affective domain. This study aims to address this gap by investigating differences between a virtual environment and a traditional hands-on environment within the affective domain of learning. Before differences can be tested, evidence for the validity and reliability of the data generated from the scales used needs to be provided. As such, this research is guided by the following three questions:

1. To what degree can previously developed scales be adapted for use in the laboratory environment?
2. What evidence of validity and reliability supports data collected with the adapted scales?

3. How do various aspects of the affective domain compare when students complete a general chemistry experiment virtually as opposed to a traditional hands-on experiment?

Methodology

Overview

A post-test-only alternative treatment control group research design was chosen to investigate the effect of virtual experiments on affective outcomes (Mack et al.). To address Research Question 1, several affective scales were selected and modified for use in the laboratory setting. Cognitive interviews were used to evaluate the validity of the modified scales. The scales were distributed within the interview format to determine how students responded and to assess any issues with the response process for the selected items. To address Research Question 2, psychometric analysis was performed to address the internal structure of the individual scales, the scales relation to each other, and the internal consistency of the scales. To address Research Question 3, comparative statistics were used to determine what, if any, affective differences existed between environments. As this project has multiple facets, the research questions will be used to organize the methodology and results sections.

Human Subject Research

All parts of this research were approved by the institutional review board at the author's home institution, Portland State University. Participants in the laboratory courses had the option to provide informed consent and only those participants who consented are represented in the data.

Research Question 1: Selection of Scales

Anderson summarized that there are seven central student affective characteristics: values, academic self-esteem, anxiety, interests, locus of control, attitudes, and preferences (Anderson & Bourke, 1981). Anxiety, attitudes, and interests were chosen for inclusion in this study. The selection was limited to three to avoid survey fatigue, thereby reducing the chance of students reporting less thoughtful answers. The Attitude toward the Subject of Chemistry Inventory (ASCI) scales for the cognitive aspect of attitude (intellectual accessibility) and the affective aspect of attitude (emotional satisfaction) were selected based on one of the general goals of the chemistry laboratory; to increase students' attitude toward science (Hofstein, 2017). The original ASCI scales were chosen over the modified version of the scales (ASCIv2) because the scales were being adapted for a new context. Thus, all original items were tested as it was unclear if the modifications made would be the same modifications needed for a different context. Additionally, there may be differences in students' anxiety as those completing an experiment within technology-based alternative environments, such as a virtual environment, do not need to consider the safety hazards of chemicals nor do they have to wear proper personal protective equipment (PPE). Therefore, an anxiety scale based on the Chemistry Laboratory Anxiety Instrument (CLAI) and the ASCI anxiety scale was created. There may also be specific environment differences, such as ease of equipment use, and thus items from the Virtual and Physical Experimentation Questionnaire (VPEQ) were selected. Beyond these differences, there may be differences in students' incoming interest toward chemistry. To control for any incoming differences, an interest scale

previously adapted for chemistry (Ferrell & Barbera, 2015) was selected. The modified version of each scale used in this study is provided in Appendix A.

Research Question 1: Modifying and Adapting Selected Scales

Emotional Satisfaction and Intellectual Accessibility

The ASCI and the modified version (ASCIv2) have been widely used to measure students' attitude changes across the implementation of an intervention (Chan & Bauer, 2015; Mooring et al., 2016; Vishnumolakala et al., 2017). Although the ASCI was originally administered in the laboratory (Bauer, 2008), the items were operationalized to chemistry as a whole. Therefore, the frame of reference of these scales was altered by changing from the original frame of reference of "Chemistry is…" to "This experiment was…". The emotional satisfaction and intellectual accessibility scales are comprised of semantic differential questions. A semantic differential question contains a spectrum between two polar opposite words such as safe and unsafe (Heise, 1970). The administration of the scales was done electronically and as a result, it was possible to change the original seven-point semantic differential scale to a sliding scale from 0 to 100. This change allowed the students to select anywhere along the spectrum. No additional changes were made to the scales.

Anxiety

Students' anxiety levels may differ between laboratory environments and have been shown to be related to student attitude (Bauer, 2008). While the original ASCI did have semantic differential items on anxiety, there was little reasoning given for why some of the word pairs were chosen. For example, it was unclear how the 'disgusting-

attractive' item pair informed aspects of anxiety. Therefore, after careful review of the original ASCI anxiety items, only the 'relaxed-tense' pair was selected for use. In addition to this item, the stems from the twenty items on the Chemistry Laboratory Anxiety Instrument (CLAI) were used to create additional anxiety semantic differential item pairings. For example, multiple CLAI items measured students' nervousness while performing a range of tasks and therefore 'nervous' was chosen as one of the words for the semantic differential. The opposite word for each word pair was decided by the researchers and were later tested for evidence of validity. For the case of nervous, calm was chosen as the opposite word. In total, there were four word pairs selected based on the CLAI stems that, with the original ASCI item, gave a total of five word pairs: nervous-calm, safe-unsafe, anxious-unconcerned, apprehensive-at ease, and relaxed-tense.

Usefulness of Lab, Equipment Usability, and Open-endedness of Lab

The VPEQ was designed to measure specific differences between the virtual environment and the traditional hands-on environment when used in a crossover research design (Pyatt & Sims, 2012). The 39 original items were analyzed to determine which could be modified to fit the current research design in which students complete an experiment in either the treatment or the control environment. For example, item 11 stated "the regular lab experiments worked better than the computer experiments" and item 29 stated "computer simulations work better than regular experiments". These items were combined to ask students if the experiment worked well. After analyzing the original items, ten items were selected that could be generalized or modified for use in

the current research design. These ten items were given on a five-point (strongly agree-strongly disagree) Likert-type scale.

Interest Scale

In addition to expected differences across environments, there may be additional differences between students that could relate to their laboratory experience and outcomes. Previous research found a link between students' incoming interest levels and course performance (Ferrell et al., 2016). Therefore, the two components of initial interest, feeling-related interest and value-related interest, were measured to account for potential differences in incoming ability (Harackiewicz et al., 2008). As these scales were adapted to measure students' incoming general interest in chemistry, no modifications were required. However, they had not been applied to or evaluated within the laboratory environment prior to this study.

Research Question 1: Scale Testing

Response Process Validity

Students were recruited for a response process interview from the general chemistry laboratory sections during the Winter 2017 term. Classroom announcements were made at the beginning of a laboratory period and students were provided a link to sign up for interview slots. Participants were compensated ($20 gift card) for participating in the interview. During the interview, participants completed the items from the anxiety, emotional satisfaction, intellectual accessibility, usefulness of lab, equipment usability, open-endedness of lab, initial interest-feeling, and initial interest-value scales one item at a time. For each item, they were asked to provide their reasoning

for selecting a particular response to determine if there was evidence for response process validity. This type of validity is an evaluation of the respondents understanding of an item to ensure its alignment with the authors intended meaning (American Educational Research Association et al., 1999; Schwartz & Barbera, 2014). If the rationale students provided was not aligned with the researcher's intended meaning, the item was flagged for modification or removal. If students gave a rationale that was aligned with the intended meaning, there was validity evidence that the item was being interpreted as intended and could be used in future implementations of the scale in the given context.

Research Questions 2 and 3: Implementation of Experiments

Population

All sections of the general chemistry laboratory at Portland State University were assigned to complete the selected experiment in either the virtual environment (treatment group) or the traditional hands-on environment (control group). This is an approximately random selection as students enroll in the section that best fits their schedule. Students were not made aware ahead of time which sections would complete the experiment in a virtual environment. A total of 28 sections of the general chemistry laboratory taught by 14 graduate teaching assistants (TAs) were offered in the Fall 2018 term with a total enrollment of 634 students. Sixteen of these sections completed the virtual experiment and twelve sections completed the traditional hands-on experiment. For logistical reasons, the sections were unable to be balanced at fourteen sections for each environment. While all students completed the experiment in their section's assigned

environment, only responses from students who consented to participate in the study were analyzed.

Experiment Selection and Design

The concept of Beer's Law is commonly covered in an experiment during the first term of a general chemistry laboratory sequence and as such was selected for this research study. In this institution's hands-on experiment, students use a known concentration of Allura Red to make a calibration curve and calculate the concentration of Allura Red in red Gatorade. The matching virtual experiment was chosen as the 'Spectrophotometry: Calibration Curves' experiment within LearnSmart Labs by McGraw-Hill Education. In this virtual experiment, students are randomly given either a red, yellow, or blue dye. They are given five prefilled test tubes with known concentration of the selected dye to create the calibration curve and a test tube with unknown concentration. Additionally, since the virtual students did not have to create their own solutions from a stock solution, they were also tasked with first completing the 'Dilute Solutions' experiment in the LearnSmart environment. This experiment had them dilute a stock solution to two different concentrations, which allowed them to still gain practice with the concept of diluting a stock solution.

Data Collection

Students completed either the traditional hands-on or virtual experiment in the laboratory room with their TA and laboratory partner. Once students completed the experiment, the TA prompted them to use the laboratory computers to answer the scale items administered through the Qualtrics program. Included among the scale items was a

'check item'. A check item is a question that asks students to select a specific response option to ensure that students are carefully reading and responding to the items. For this study, the check item read "Please select strongly agree for this question". Therefore, it is assumed any student who did not select strongly agree for the question was not carefully reading the items and thus their data was not used for analysis.

Research Question 2: Analysis

After collecting the affective item responses, the structure of the scales was examined to provide evidence of the internal structural validity of each scale. To do this, confirmatory factor analysis (CFA) was used to evaluate *a priori* models of each scale. The analyses were conducted using version 0.6-3 of the R package lavaan (Rosseel, 2012). To account for any non-normality in the data, the maximum likelihood with Satorra-Bentler corrections (MLM) estimator was used for all CFA models. Additionally, previous research provided links between the feeling and value aspects of interest (Ferrell & Barbera, 2015) and between the constructs of anxiety, emotional satisfaction, and intellectual accessibility (Kurbanoglu & Akim, 2010). Therefore, a two factor and a three-factor model, respectively, were tested to confirm those relations. All scales were reviewed for potential modifications if individual factor loadings were below a cutoff value of 0.4 or if the fit indices were out of range (i.e., CFI below 0.95, SRMR above 0.08, and/or RMSEA above 0.06) as recommend by Hu and Bentler (1999). Additionally, the internal consistency of each scale was tested and modifications were made if any scales had an omega value below the generally accepted cutoff value of 0.70. McDonald's omega is similar to the commonly reported alpha but is more appropriate for

congeneric models, which is a model where factor loadings and error terms are not constrained to be equal (Komperda et al., 2018). Therefore, single-factor congeneric CFA models were tested for each scale.

Research Question 3: Analysis

Once all scale data was deemed to have acceptable model fit, invariance testing took place to examine if both the treatment and control groups responded to the items in a similar fashion. To test for measurement invariance, each model was tested with the data split by group rather than combined into a single data set. If the global model fit is still within the acceptable range when the model is tested by group (with equal loadings and intercepts) or if the CFI changes by less than or equal to 0.01, then measurement invariance is determined and the two groups can be compared on the affective items (Cheung & Rensvold, 2002). A relatively small change in the CFI indicates that specifying the model to have equal loadings and intercepts did not change the model in a meaningful way. This indicates that the two groups were responding in a similar fashion. If the model fit changed drastically then further investigation of the response differences would be warranted. Once invariance was determined, a multivariate analysis of variance (MANOVA) was conducted using version 24 of SPSS to evaluate group means on multiple affective scales to determine if there are any statistical differences in the measured affective domain aspects between the virtual and traditional hands-on environment.

RESULTS

Research Question 1: Response Process Validity

Ten students participated in an interview. Each student completed the items from scales one item at a time and then gave their reasoning for their response selection. For example, on the intellectual accessibility scale, a student selected that the experiment was closer to the 'confusing' side of the word pair 'confusing-clear' because "I think the procedure was a little unclear". This response aligned with the intended interpretation of the word pair and thus provided evidence for response process validity. For all but one item, students gave reasoning for their selected answer that correctly aligned with the intended meaning of the item. This increased the confidence that the newly created anxiety scale was functioning as intended. All ten students struggled to correctly interpret the item, "There is opportunity for me to pursue my own experimental interests" from the VPEQ usefulness of lab scale. All students interviewed were confused about what the term "experimental interests" meant. Therefore, this item was discarded and all other items were retained.

Research Questions 2 and 3: Population

There were 634 students enrolled in the first term general chemistry laboratory in the Fall 2018 term. Of those students, 448 students consented to have their data analyzed. There were 52 students who incorrectly responded to the check item "Please select strongly agree for this question" and thus were removed from the data set leaving 396

students in the final cleaned data set of which 178 of the students completed the traditional hands-on experiment and 218 completed the virtual experiment.

Research Question 2: Reliability and Validity Evidence of Data Provided By Scales

Individual Model Testing and Modifications

Before the difference in means between laboratory environments were analyzed, the scales were tested individually as single-factor, congeneric, CFA models to ensure they functioned as intended. This took place for each scale with four or more indicator items before analyzing the relations between individual scales. The equipment usability and open-endedness of lab scales each consisted of two items and the initial interest-value scale consisted of three items. Thus, it was not possible to test the model fit for these scales as they would not be over-identified models (Duncan, 1975).

Single-factor model results for the anxiety scale indicated that each fit statistic was outside the chosen cutoff criteria (Table 4.1). Additionally, the CFI for the intellectual accessibility scale was outside of the chosen cutoff. All other scales had CFI and SRMR values that were within the acceptable cutoff range and most scales had RMSEA values outside of the cutoff range, as shown in Table 4.1. However, when the degrees of freedom in a model are low (e.g. less than 50), the RMSEA is biased and should be interpreted with caution (Kenny et al., 2015). The degrees of freedom are low for these models and as such the RMSEA was not used as a primary indicator of fit. The lack of model fit for the anxiety and intellectual accessibility scales required further investigation before analysis.

Along with model fit, the internal consistency of each scale with four or more items was tested. Cronbach's alpha is commonly used to measure this reliability, however, as the single-factor models were tested as congeneric, McDonald's omega was more appropriate (Komperda et al., 2018). All scales had an acceptable internal consistency as shown by the omega value in Table 4.1.

As the anxiety scale did not have acceptable fit, the loadings were examined and it was noted that the word pair 'safe-unsafe' loading was 0.165, which was significantly lower than the other items and below the chosen cutoff value of 0.4. Therefore, this item was removed and upon retesting the single-factor model, the anxiety scale had acceptable fit indices, as noted as 'modified anxiety' in Table 4.1, and had satisfactory internal consistency. The loadings of the intellectual accessibility scale did not have any values below the chosen cutoff value, therefore, the modification indices of this scale were investigated to determine if there were feasible relations between variables that would improve model fit. A high modification index between the error terms of the 'hard-easy' and 'challenging-unchallenging' word pairs was detected. Cohen's *w* was calculated to determine the effect of correlating these terms, it was determined that the modification would result in a large effect of 0.46 (Cohen, 1992). Given the large effect and the word pair similarities, it is possible that they did not have independence of errors and thus the error terms for these items were correlated. This modified model showed acceptable model fit with good internal consistency, noted as 'modified intellectual accessibility' in Table 4.1.

Table 4.1: Fit indices and internal consistency values for single-factor models. Indices in italics are outside of the recommended range. Omega values are only shown for congeneric models deemed to have acceptable fit.

	CFI	***SRMR***	***RMSEA***	***df***	***Omega***
Emotional Satisfaction	0.95	0.04	*0.26*	2	0.88
Intellectual Accessibility	*0.91*	0.05	*0.22*	5	---
Modified Intellectual Accessibility	0.99	0.02	*0.10*	4	0.85
Anxiety	*0.85*	*0.09*	*0.21*	5	---
Modified Anxiety	0.99	0.02	0.05	2	0.81
Usefulness of Lab	1.00	0.01	0.05	2	0.85
Initial Interest-Feeling	0.98	0.02	*0.14*	2	0.88

Two-factor and Three-factor Model Testing

The ASCI scales were published with correlations between anxiety, intellectual accessibility, and emotional satisfaction. Therefore, a three-factor model was tested to determine if the newly created anxiety scale correlated to the existing scales in a similar fashion. The four-item anxiety scale, the emotional satisfaction, and the intellectual accessibility with the correlated item error term were tested as a three-factor CFA model, as seen in Figure 4.1. The three-factor model produced acceptable fit indices (CFI: 0.95, RMSEA: 0.08, SRMR: 0.04) allowing correlation comparisons to be made to the previously reported values.

The three-factor model had a strong positive correlation between the intellectual accessibility and emotional satisfaction factors (0.91). Both intellectual accessibility and emotional satisfaction correlated negatively with anxiety (-0.65 and -0.66 respectively), as expected (Bauer, 2008). Bauer reported the correlation between anxiety and emotional satisfaction as -0.72, between anxiety and intellectual accessibility as -0.58, and between emotional satisfaction and intellectual accessibility as 0.62. While the correlation to anxiety is similar, Bauer used different word pairs for the anxiety scale and did not have a correlated error term for intellectual accessibility. Similarly, Xu and Lewis did not have

the correlated error term, had one less item for intellectual accessibility and did not include an anxiety scale in their ASCIv2. Even with the differences, they reported a similar correlation of 0.80 between their version of emotional satisfaction and intellectual accessibility (Xu et al., 2013). These correlations were similar to the previously reported correlations with the exception of the correlation between emotional satisfaction and intellectual accessibility, which was higher than previously reported. This could indicate that the affective and cognitive aspects of attitude may not be as distinct in the laboratory environment using these scales or that the addition of the original fifth item and the correlated error term strengthened the relation between the scales.

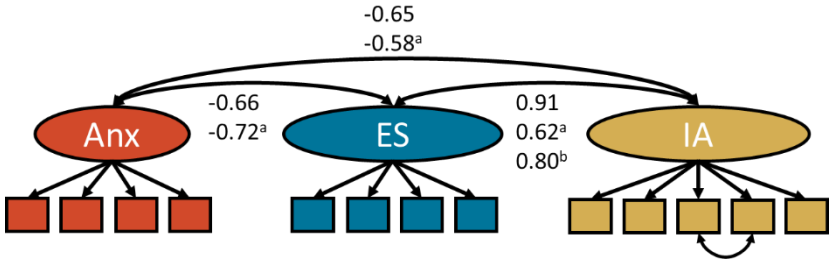


Figure 4.1: Three-factor model showing correlations between ASCI scales including Anxiety (Anx), Emotional Satisfaction (ES), and Intellectual Accessibility (IA). Correlation coefficients from original ASCI (a) and ASCIv2 (b) shown for comparison.

The interest scale had been previously reported in the literature as a two-factor model (Ferrell & Barbera, 2015). Thus, the 'feeling' and 'value' components were tested as a two-factor model and produced an acceptable fit (CFI: 0.98, RMSEA: 0.07, SRMR: 0.03). The correlation between the 'feeling' and 'value' factors was 0.77, which was similar to the reported value of 0.69 as noted in Figure 4.2.

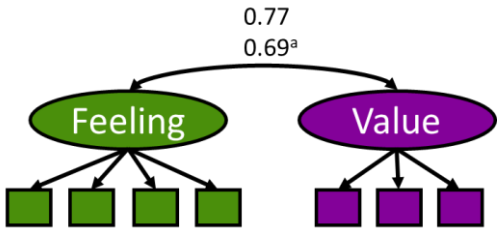


Figure 4.2: Two-factor model and correlations between initial interest scales. Correlation coefficient from original Ferrell and Barbera data shown for comparison.

After testing the models, it was determined that the three-factor model of the anxiety, emotional satisfaction, intellectual accessibility scales and the two-factor model of the initial interest-feeling and initial interest-value scales had acceptable fit in this context and produced similar correlations to their prior setting and format. These results added to the confidence that the scales were functioning as intended, including the initial interest-value scale that could not be evaluated as a single-factor model. However, before the individual scales can be used to compare the two environments, measurement invariance had to be shown to ensure that both groups were responding to the scales in a similar manner.

Invariance Testing

Each single-factor model was retested by group to measure the change in the global fit indices with equal intercepts and loadings. Under these conditions, the CFI for all models only changed slightly except for the one-factor intellectual accessibility model, as shown in Table 4.2. This result indicates that both the virtual and the hands-on students were interpreting the scale items similarly and a comparison between group means could be made for the individual scales. The intellectual accessibility scale had acceptable fit indices, despite the relatively large change in CFI, when tested as a grouped

model (CFI: 0.95, SRMR: 0.07, RMSEA: 0.13) and as such, it was also deemed acceptable for inclusion in the group mean analysis.

Table 4.2: Difference in fit indices for models by group.

	ΔCFI	***ΔSRMR***	***ΔRMSEA***	***df***
Emotional Satisfaction	-0.01	0.01	-0.09	10
Intellectual Accessibility	-0.04	0.05	0.03	16
Anxiety	-0.01	-0.03	-0.03	10
Usefulness of Lab	-0.01	0.04	0.01	10
Initial Interest-Feeling	0.01	0.01	-0.07	10

Research Question 3: Group Comparisons

With evidence of scale invariance established, group comparisons were made. The raw average scale scores for the affective scales are presented in Table 4.3. The anxiety, emotional satisfaction, and intellectual accessibility were collected on a 0 to 100 sliding scale. The remaining scales were on a 5 point Likert-type scale.

Table 4.3: Raw averages for the affective scales

	Anx	**ES**	**IA**	**II-F**	**II-V**	**U**	**EU**	**OE**
Hands-On	32.71	72.28	66.10	3.76	4.32	3.78	4.21	3.54
Virtual	35.68	60.33	57.80	3.69	4.23	3.47	3.75	3.54

Anx: anxiety, ES: emotional satisfaction, IA: intellectual accessibility, II-F: initial interest-feeling, II-V: initial interest-value, U: usefulness of lab, EU: equipment usability, OE: open-endedness of lab

A MANOVA was performed to compare groups using the average scores of the individual scales after checking the assumptions for a MANOVA. The first four assumptions are a function of the study design. All scales chosen were either Likert-type or continuous scales and as such were treated as interval data. It is appropriate to treat composite scores from Likert-type scales as interval data whereas individual item scores should be treated as ordinal data as the differences between responses options are unequal (Boone & Boone, 2012). Students completed the experiment in only one environment and thus there were independent groups. As students can randomly enroll in whichever

section fit their schedule best, there were also independent observations within each group. Additionally, a chi-square test found no statistically significant difference for gender, race, or age between students in the two environments. There were 396 students in the data set, which is a sufficient sample size to conduct the MANOVA.

The last five assumptions are not a function of the study design and need to be checked statistically. Multivariate outliers were tested for using Mahalanobis distance (De Maesschalck et al., 2000). This is a measure of the distance between two points in multivariate space and is used to find rare combinations of variables. For example, students who responded they were simultaneously anxious and comfortable. There were six multivariate outliers found that were above the chi-square value of 26.13 and all six completed the virtual experiment. These six data points were removed before further analysis took place. Multivariate normality was not directly assessed, however, the normality of each scale was evaluated using the skewness and kurtosis values. The skewness for initial interest-value and equipment usability were both below the generally accepted cutoff for normal data of negative one (Potthast, 1993). All other values were within the range of negative one to positive one. A MANOVA is robust to skewness (Olson, 1974) and therefore it is still possible to analyze the data with the skewness in the two scales. To test if there was a linear relationship between groups for each scale, scatterplots by group for each scale were analyzed and there was a linear trend in the scatterplot for all scales. Homoscedasticity is measured to ensure students in both environments had similar variances on the affective scales. Homoscedasticity was assessed through Levene's test and it was found that the scales intellectual accessibility,

emotional satisfaction, usefulness of lab, equipment usability, and open-endedness of lab all had a significant result indicating that the variances were different between groups of students. Similarly to skewness, a MANOVA is robust to homoscedasticity violations and can still be conducted. Lastly, the variance inflation factor (VIF) was tested and all scales had a value greater than one and less than ten, which means that there was no multicollinearity. Multicollinearity is measured to ensure that no two variables are so highly correlated that they are essentially measuring the same construct.

The results of the MANOVA indicated that there were significantly lower averages on emotional satisfaction, intellectual accessibility, usefulness of lab, and equipment usability for the virtual students. The emotional satisfaction and equipment usability scale score differences had a medium effect size, as measured by partial eta squared(Richardson, 2011), while the differences in intellectual accessibility and usefulness of lab scale scores had small effect sizes. The significant differences are represented in bold in Table 4.4 and the full MANOVA results can be found in the Appendix A. No statistical difference between groups was detected for initial interest-feeling, initial interest-value, anxiety, or open-endedness of lab and as such, these scales were not included in further analyses.

Table 4.4: Significance and effect size of the group mean differences

	p-value	Effect Size
Anxiety	0.237	0.004 (small)
Emotional Satisfaction	**<0.001**	0.049 (medium)
Intellectual Accessibility	**0.001**	0.027 (small)
Initial Interest-Feeling	0.466	0.001 (small)
Initial Interest- Value	0.238	0.004 (small)
Usefulness of Lab	**0.001**	0.028 (small)
Equipment Usability	**<0.001**	0.056 (medium)
Open-endedness of Lab	0.971	0.000 (small)

While the initial MANOVA revealed significant differences between environments, there were 14 graduate TAs in charge of teaching the laboratory and the differences detected could be the result of an instructor-effect and not necessarily reflective of the environment itself. To investigate for this possibility, the students' scale scores were plotted by which TA taught their section, an example is seen in Figure 4.3. While it would be possible to conduct a MANOVA with posthoc tests to determine if the averages by TA were statistically different from each other, the sample size by TA in this study do not provide sufficient power to warrant such a test. Given this limitation, the data was examined visually. Upon inspection, it was noted that students who completed the virtual experiment with TAs numbered 11-14 (Figure 4.3) had a lower average for emotional satisfaction than those who completed the virtual experiment with the other TAs (7-10). Therefore, it was possible that the differences in TAs were influencing the significant differences found in the initial MANOVA. To test this, the data from the sections taught by TAs 1-6 were grouped together (Hands-On), by TAs 7-10 were grouped together (Virtual Group A), and by TAs 11-14 were grouped together (Virtual

Group B). The MANOVA was rerun with three groups instead of two to determine if an instructor-effect was leading to the differences between environments.

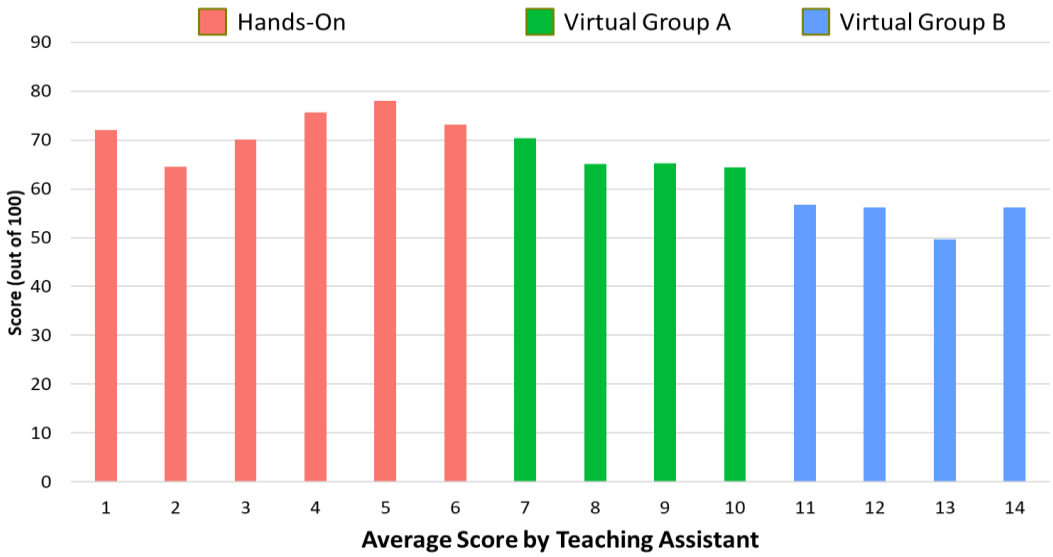


Figure 4.3: Average emotional satisfaction for each TAs sections.

Instructor Effect

As each group had a sufficient sample size for use of Bonferroni corrections as the posthoc test, it was possible to compare the three groups to determine if the visual grouping was meaningful. Virtual Group A and Virtual Group B were compared and significant differences in the emotional satisfaction, usefulness of lab, and equipment usability scales were found (Table 4.5). These differences support the groupings of TAs, although it should be noted that there was no significant difference between the groups for the intellectual accessibility scale. These two groups were then compared separately with the Hands-On group to determine if splitting the virtual TAs changed the initial findings that virtual students had lower averages on the emotional satisfaction, intellectual accessibility, usefulness of lab, and equipment usability scales. When Virtual Group A was compared with the Hands-On group, no significant difference was found

for any of the affective constructs (Table 4.5). However, there was a significant difference for all four scales when Virtual Group B was compared with the Hands-On group. These results suggest an instructor effect is present since the findings were not consistent when each virtual group was separately compared with the Hands-On group.

Table 4.5: p-values for posthoc comparisons between groups.

	Emotional Satisfaction	**Intellectual Accessibility**	**Usefulness of Lab**	**Equipment Usability**
Virtual Group A-Virtual Group B	**0.005**	0.336	**0.006**	**0.002**
Hands-On-Virtual Group A	0.160	0.214	0.883	0.119
Hands-On Virtual Group B	**<0.001**	**0.001**	**<0.001**	**<0.001**

DISCUSSION

With an increase in the use of alternative laboratory environments, it is important to measure differences that exist between the alternative and traditional hands-on environments. However, there has been a lack of affective instruments that have been designed or modified for the laboratory environment to measure these differences. Therefore, an anxiety scale adapted from the CLAI and ASCI, the emotional satisfaction and intellectual accessibility scales from the ASCI, the initial interest-feeling and initial interest-value scales, and the usefulness of lab scale from the VPEQ were modified for the laboratory environment and the psychometric properties were tested. All scales had acceptable fit indices after modification and measurement invariance was established, as noted in Tables 4.1 and 4.2. The three-factor model of anxiety, emotional satisfaction, and intellectual accessibility and the two-factor model of initial interest-feeling and initial interest-value had acceptable fit and produced correlations that were similar to previously reported values. The scales for open-endedness of lab and equipment usability were not

tested with factor models as they had too few items but were included when running the MANOVA. With all scales functioning as intended and measurement invariance established, the group means were compared to determine what affective differences may exist between environments.

The students who completed the virtual Beer's Law experiment had a higher average anxiety score and lower averages on all other affective aspects. These findings were in agreement with previous research (Enneking et al., 2019) that found students experience less favorable affective outcomes when completing a virtual chemistry experiment in the LearnSmart environment as compared with the traditional hands-on environment. However, when instructor effect was accounted for in the present study, the results were split. No significant differences were detected between the scale scores of students in the traditional hands-on environment and students who completed the experiment in the virtual environment with a TA in group A (Hands-On-Virtual Group A in Table 4.5). However, all four scales showed a significant difference when the same comparison was made between students in the traditional hands-on environment and students who had a TA in Virtual Group B (Hands-On-Virtual Group B in Table 4.5). This instructor effect indicates that which TA the students had may be a more decisive factor in their scores on the affect constructs measured than which environment they completed the experiment in.

One possible explanation for the difference seen in students' affective scores could be due to the TAs prior teaching experience. One TA in the Hands-On group, one TA in Virtual Group A, and three TAs in virtual group B were all teaching laboratories

for the first time at this institution. It was possible that the added burden of teaching in a new environment, while still becoming generally comfortable with teaching, could have negatively impacted their section. If the virtual experiment took place after these TAs had gained additional teaching experience, it is possible that the difference between TAs would have been smaller.

CONCLUSIONS

Although previous research on virtual chemistry laboratories has indicated no statistical differences in cognitive and psychomotor outcomes, little research had been presented on affective differences. Therefore, this research adapted and modified existing affective scales for the chemistry laboratory and administered them to students conducting a virtual Beer's law experiment and a traditional hands-on version of the experiment. The scales were tested for the laboratory environment context to ensure they were functioning as intended in the new context. Evidence for the response process validity was provided by student interviews. All scales produced acceptable fit indices when tested with a single-factor, congeneric, CFA model, which supports structural validity. Internal consistency reliability of the scales was supported with acceptable McDonald's omega values. This evidence provides support that it is possible to modify existing affective measures that have been designed for, and tested in, the classroom setting and apply them to the laboratory setting. With functioning affective scales for the laboratory, future studies can increase the body of literature on affective laboratory outcomes.

The functional scales were used to determine what differences in affective constructs may exist between a virtual experiment and a hands-on experiment. After controlling for instructor-effect, statistical differences were split based on which TA the students had. It is possible that the TA's comfort level with teaching the experiment in the environment they were assigned to impacted the students' affective aspects in the laboratory. With previous studies finding no difference in the cognitive and psychomotor domain and this study finding split differences in the affective domain when controlling for instructor effect, laboratory coordinators have evidence that seeking an alternative environment may not significantly harm their student outcomes for this particular experiment. However, the possibility of an instructor-effect should be taken into consideration when considering how students react to an alternative environment.

Recently, there has been a call for future studies to better understand the impact laboratories have on student learning (Bretz, 2019). Future work should be conducted to examine the effect that a TA has on the student's experience. With a lack of uniform TA training or experience, it is possible that students completing identical experiments with a different TA may have drastically different experiences and outcomes. The best practices for training TAs are not well understood (Reeves et al., 2016) and future research should investigate how the instructor-effect can be mitigated through rigorous TA training programs. In this study, these differences were found to impact affective outcomes more than the environment the students completed the experiment in.

CONSIDERATIONS FOR FUTURE RESEARCH

Knowing that there is an instructor effect present in this study, and possibly other studies, on the chemistry laboratory, future research should aim to control for instructor effects as much as possible. Suggestions include increasing the length of training time around the virtual experiment and/or qualitatively observing each classroom to make note of possible differences between TAs. Although it is difficult to ensure different instructors are equal across sections, taking steps early in the research process to control for the possible differences will allow for clarity on if any findings are due to the intervention or the instructor.

LIMITATIONS

While split differences between environments were found in this specific context, the results should not be generalized to other contexts without further testing. The traditional hands-on experiment the students completed was a confirmatory lab in nature and the results may be different from an inquiry-based, project-based, or other types of hands-on curricula. The virtual experiment was completed using a modified procedure in the LearnSmart Labs and the results may be different if a different virtual environment is used. The research took place at Portland State University, a non-traditional urban university in the Pacific Northwest, and the results may be different at different institutions. Future work should focus on testing affective differences between environments in a wide range of contexts to determine how generalizable these findings are.

Additionally, although the students enrolled randomly in sections and did not know which environment or TA they would have at the time of registration, it is still possible that a larger portion of students with negative attitudes towards chemistry enrolled in the four virtual sections with less favorable affective aspects. 'Initial interest' was measured to attempt to capture this difference but may not have adequately captured all incoming differences. As such, it would not be appropriate to use the modified and newly created affective scales to make conclusions about individual TA effectiveness. There are many factors that could influence the differences seen and it is possible that those factors could be outside of the TAs control. Lastly, the analysis was conducted by combining sections to have enough power to detect differences between environments, which leaves the possibility that the findings are not representative of an individual's experience as there may be differences between environments for individual students that were not captured when the data was aggregated.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available in Appendix A and includes administered survey items, initial MANOVA results, and scale scores by TA group.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jbarbera@pdx.edu

ACKNOWLEDGMENTS

We would like to thank the graduate teaching assistants, students, the laboratory coordinator, and stockroom staff that were involved with this work. Without all of them, this study would not have been possible. Additionally, we would like to thank McGraw-Hill Education who generously allowed us to use the LearnSmart Labs platform for this work.

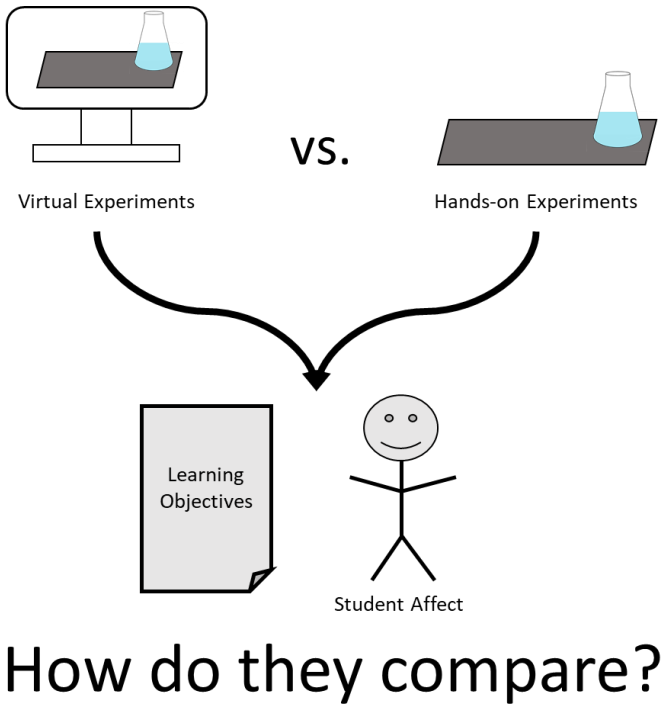
CHAPTER V: Assessing Differences Between Three Virtual General Chemistry Experiments and Similar Hands-On Experiments

ABSTRACT

To date the efficacy of virtual experiments is not well understood. To better understand what differences may exist between a hands-on learning environment and a virtual learning environment, three experiments were chosen for investigation. For each experiment, approximately half of the students completed a hands-on version of the experiment and the other half completed a virtual version. After completing the given experiment, students were compared on: their ability to meet the learning objectives for that experiment, their responses to six affective scales, and their grade on a laboratory report. Differences were found on four learning objectives. Two of these learning objectives were on the Beer's Law experiment and the other two were on the titration experiment whereas the calorimetry experiment had no differences between groups on learning objectives. However, all four differences are likely due to differences in procedures between environments and not due to the environment itself. Additionally, differences were found on two of the affective scales (usefulness of lab and equipment usability) across all three experiments indicating that the students who completed a virtual experiment found the experiment to be less useful and the virtual environment harder to use. Students that completed the virtual version of the titration experiment also

reported that the experiment took less time as indicated by the difference on the open-endedness of lab scale. These differences are not representative of a students' individual experience, however. To capture individual experiences, latent profile analysis was conducted to determine what affective profiles existed within the population. There were three common profiles identified across the three experiments: low affective outcomes, medium affective outcomes, and high affective outcomes. These indicate that while the majority of the students have medium or high affective outcomes and do well on laboratory reports, there is anywhere from four to seventeen percent of the students completing a given experiment, that have low affective outcomes but still do equally well on the laboratory report as the other students. Future work should be conducted to assess why students report low affective outcomes and if a different type of laboratory learning environment or curriculum type would better serve them.

GRAPHICAL ABSTRACT



KEYWORDS

General Public, Chemical Education Research, Computer-Based Learning, Distance Learning

INTRODUCTION

The general chemistry laboratory has historically been a place where students or apprentices learn valuable trade skills for their future career. While scientific thinking and fundamental laboratory skills are still essential for many careers, there has been a drastic increase in the career options students have. This, along with the fact that typically a wide variety of majors enroll in the chemistry sequence, creates a new challenge for

designing a laboratory experience that adequately prepares all students for their future career.

Some institutions have accommodated the differing career goals by creating laboratory sections that cater to different populations of students. For example, students at the University of California San Diego that are pursuing a career in chemistry may opt to enroll in a laboratory course designed specifically for chemistry majors whereas students pursuing nursing at California State University, Sacramento may opt to enroll in a laboratory course with a pre-health focus. However, the ability to create multiple sections catering to different populations of students varies by institution and there is a lack of agreement as to whether non-science majors, or specifically non-chemistry majors, need to take a laboratory that teaches them chemistry-specific skills (Chittleborough et al., 2007; Tro, 2004; Wartell, 1973). In fact, some have gone as far as suggesting that non-majors do not need the laboratory and question why institutions are spending money to teach them laboratory skills (Hawkes, 2004). One challenge in offering multiple types of laboratory experiences is that the number of laboratory sections is often limited by space and staff availability. Some have met this challenge by creating a hybrid curriculum where students complete half of their experiments in a virtual environment and the other half in a traditional hands-on environment (Enneking et al., 2019). While this frees up physical laboratory space, questions remain on the efficacy of virtual experiments.

Previous research on virtual experiments across STEM disciplines have generally found that students perform equally well on cognitive assessments regardless of the type

of learning environment they completed the experiment in (De Jong et al., 2013; Hawkins & Phelps, 2013; Ma & Nickerson, 2006; Pyatt & Sims, 2012; Tatli & Ayas, 2010; Tatli & Ayas, 2013). This trend holds true within chemistry where some studies found no difference in cognitive outcomes (Enneking et al., 2019; Hawkins & Phelps, 2013) and others found that students completing the virtual experiment outperformed those that completed the hands-on experiment (Pyatt & Sims, 2012; Tatli & Ayas, 2010; Winkelmann et al., 2017). Therefore, there exists an established body of evidence that laboratory coordinators can use to make well informed decisions about using virtual experiments. However, cognitive assessments do not measure if students learn the same skills for their career or if they had a positive experience in the laboratory. There has been significantly less research conducted on the psychomotor and affective domains of learning, leaving laboratory coordinators unsure if virtual experiments can truly provide an equal experience for students. Two studies that include a laboratory practical as part of their comparison between virtual and hands-on environments have found that students that learned a skill in the virtual environment are able to successfully perform the skill in person as well (Hawkins & Phelps, 2013; Winkelmann et al., 2017). Despite this, it is possible that students can learn the same content and perform the skill without having a favorable laboratory experience. In fact, one study focused on hybrid laboratories did include affective domain items and found that students who completed the virtual experiment had lower affective outcomes than the control group (Enneking et al., 2019).

This prior work highlights the need to further assess the affective domain when students are completing a virtual experiment. While this study used an instrument (the

MLLI (Galloway & Bretz, 2015a)) that measures the affective domain with eight general items, the affective domain is a broad domain that contains many constructs. As virtual experiments grow in popularity, it is imperative that information about how outcomes on specific constructs compare. This then allows laboratory coordinators to make more informed decisions. One affective construct that has been previously studied in the laboratory and can impact students' experience is anxiety. The Chemistry Laboratory Anxiety Instrument (CLAI) (Bowen, 1999) was developed to specifically measure this construct in the chemistry laboratory environment. It is possible that students who complete the experiment virtually have differing levels of anxiety, as they do not need to worry about personal protective equipment (PPE) or chemical safety. In addition to anxiety, there may be other differences based on the specific environment. The Virtual and Physical Environment Questionnaire (VPEQ) (Pyatt & Sims, 2012) was designed to address specific differences between environments and addresses constructs of equipment usability, usefulness of lab, and open-endedness of lab. These three constructs measure students' feelings towards specific components of the laboratory. In addition to environment-specific differences, there may also be broader affective differences. One of the important goals of any science laboratory and especially chemistry is to improve students' attitude toward chemistry (Hofstein, 2017). It is possible that the ability to improve students' attitude differs based on the learning environment.

In addition to the lack of specific affective constructs studied with regard to virtual laboratories, there is also a need to further study the cognitive outcomes. Despite the number of studies finding no differences between environments, these studies have

relied heavily on the use of multiple-choice test or quiz items to determine performance. Relying on these types of assessments inadequately captures whether students have the scientific thinking needed for many careers. In fact, there has been a recent push to incorporate curricula that focus on scientific writing instead of short post-lab items (Burke et al., 2006; Greenbowe et al., 2007; Stephenson & Sadler-McKnight, 2016). Thus, it is important that rather than compare students on multiple-choice assessments they are compared on their ability to meet the desired cognitive learning objectives of the experiment. However, to date, there is a lack of agreed upon experiment-specific learning objectives that can be used to assess the environments (Hofstein & Lunetta, 2004). With specific learning objectives for each experiment, it would then possible to compare environments and determine if they meet them equally on an experiment by experiment basis.

If evidence is provided that students are meeting the same cognitive objectives and affective outcomes in a new experiment environment (i.e., virtual) as compared with the traditional environment, then laboratory coordinators can select the environment that best matches both the faculty members' goals at that institution and the intended student population for the curriculum. With a wide arsenal of experiments, both virtual and hands-on, that have established and measurable outcomes it would be possible to design multiple laboratory courses that align with the ranges of student expectations and career motivations without the limitation of physical laboratory space.

RESEARCH QUESTIONS

There is a need to better understand if students completing an experiment in an alternative environment, such as the virtual environment, are able to meet the same learning objectives and acquire similar affective outcomes as students in the traditional hands-on environment. The following research questions guided this study:

1. To what degree can experiments in a virtual environment meet the same learning objectives as similar traditional hands-on experiments?
2. How do students' affective outcomes compare when completing an experiment in different learning environments?
3. What effect do individual student differences have on any observed differences in either the affective or learning objective outcomes?

METHODOLOGY

Human Subject Research

This research was approved by the institutional review board at Portland State University. Participants were asked to provide informed consent and only data from those who consented were analyzed.

Selection of Experiments

There is wide variety in the experimental topics covered across different institutions with each institution selecting the topics that they value most. However, there are a number of common topics that are taught at most institutions. Previous work conducted by Reeves and Exton as part of the development of the ACS General Chemistry Laboratory Exam helped gain a better understanding of which experiments are commonly done (Reeves & Exton, 2014). They first compiled a list of laboratory manuals used at a range of institutions, which generated thirty-six unique sources, and

reviewed each for the experimental topics included. After reviewing the manuals, the six most commonly covered topics were:

- Volumetric analysis (titrations)
- Stoichiometry
- Kinetics (determination of rate law)
- Spectrophotometry/Beer's Law
- Properties of Acids and Bases
- Calorimetry

These six experiments range in the level of laboratory skills required to complete the procedure and take place across the entire year of the general chemistry curriculum. To cover a range of skills and chemistry content, one experiment from each term of the general chemistry curriculum was selected for this study. Beer's Law was chosen as the experiment to investigate in the first term, calorimetry was chosen for the second term, and volumetric analysis (titrations) was chosen for the third term of a general chemistry laboratory course based on a quarter system.

Establishing Learning Objectives

Five faculty members from three institutions in the Pacific Northwest were interviewed in a semi-structured format to capture the specific objectives each had for the chosen experiments. Three faculty were from two different community colleges and two were from a doctoral granting university with high research activity (Carnegie Commission, 2015). One of the community colleges used inquiry-based experiments while the other community college and the university used expository experiments. Including different institutions and types of curricula in the targeted population allowed for different perspectives on the learning objectives to be captured. The interviews took place the week prior to the experiment being done at the respective institution.

Participants were asked to explain the procedure for each experiment and what they hoped students would gain by completing the experiment. As participants had different levels of understanding of what "experimental objectives" meant, the question "If students missed today's experiment, what would they miss out on?" was also asked. This question allowed participants to better articulate what important objectives they had for their students. For the full interview protocol, see Appendix B.

For a given experiment, each faculty member's learning objectives were listed and then compared with the others' objectives. With variety in the types of experiments done at institutions over the same topic, it was expected that not all learning objectives would be shared across participants. Therefore, to capture the most salient objectives of each experiment (i.e., those that faculty agreed upon) only the common learning objectives across all participants were used to assess differences in laboratory environments. Once the common objectives were established for each experiment, they were used to generate rubrics to score student's laboratory reports based on how well they met those objectives.

Student Population

Students enrolled in the general chemistry laboratory sequence during the 2018-2019 academic year at Portland State University were the targeted population. This convenience population was chosen as it provided several important features including: 1) the ability to provide significant input to the structure of the laboratory sections, 2) multiple sections that could be easily split by environment type, and 3) the ability to directly work with the university office of information technology to set-up and monitor

the functioning of the virtual experiments on the laboratory computers. The sections of the laboratory courses were split approximately in half for each of the three experiments, with some of the sections completing the traditional hands-on experiment and the other sections completing the experiment in a virtual environment. All enrolled students in a given section conducted the same experiment and generated the same cognitive and affective data as part of their normal laboratory requirements for that day. As students enrolled in whichever section best fit their schedule and did not know ahead of time which sections would conduct the experiment in a virtual environment, there was approximately random grouping. Further information about self-report demographics by grouping can be found in Table B.SI1 in Appendix B. No institutional data was provided by the university. The virtual environment used for all three experiments was the LearnSmart Labs by McGraw-Hill Education. Students completed the virtual environment procedure in their normal laboratory room working with a laboratory partner and with their teaching assistant (TA) present.

Rubrics and Scoring of Laboratory Reports

Students completed a formal laboratory report after each of the experiments included in this study. Identifying information was removed from the reports and each was assigned a random identification number prior to analysis. Coders were not aware of which environment a student completed the experiment in when scoring their report. As the codes were pre-determined based on the faculty members' learning objectives, this was a deductive analysis. A primary and secondary coder individually scored seven student reports at a time for each experiment and then met to discuss their scoring and

calculated a percent agreement. This process repeated until the coders reached 100% agreement. Consensus scoring is one method used to establish inter-rater reliability and with a high consensus score indicates the rubrics were interpreted and utilized in a similar way for the student reports (Stemler, 2004). Initially, the rubric was scored using categories of 'Does not meet' and 'Meets' to mark if a student met the learning objective, however, after preliminary testing of the rubric a third category of 'Partially Meets' was added for cases where students demonstrated only limited evidence of meeting a learning objective. After reaching consensus on a set of reports, the coders individually scored the remaining reports and met regularly to clarify any questions that arose. The rubric scores were then used to compare if students in both environments met the learning objectives to the same degree.

For each experiment, chi-square tests were conducted for individual learning objectives to determine if there were significant differences between rubric scores by learning environment. A 2x3 chi-square test was used to compare scores across two groups (i.e., hands-on and virtual) on a variable with three category options (i.e., meets, partially meets, and does not meet) (Coolican, 2017). A non-significant chi-square test indicated that no statistical differences between learning environments were detected for a given learning objective. Chi-square tests were conducted using version 26 of SPSS.

Measuring Differences in the Affective Domain

Immediately upon the completion of an experiment, six affective scales were administered to students through a Qualtrics survey. The scales measured the constructs of anxiety, intellectual accessibility, emotional satisfaction, equipment usability,

usefulness of lab, and open-endedness of lab. Evidence for the reliability and validity of the data generated by these scales, in these learning environments and with this specific population, has been previously reported (Hensen & Barbera, 2019). The reported validity data included response process validity interviews, conducted to ensure students are interpreting the items in a similar manor as is intended, as well as measurement invariance, establishing that each scale functioned similarly for students in both learning environments. A multivariate analysis of variance (MANOVA) was conducted on the affective scale data from each experiment to detect differences between the learning environments. A MANOVA is an appropriate test to compare two groups of students for multiple outcomes (O'Brien & Kaiser, 1985). Significant findings in the MANOVA would indicate differences in the affective outcomes between learning environments for the given experiment. MANOVA results from the Beer's Law experiment have been previously reported after checking all assumptions for running a MANOVA (Hensen & Barbera, 2019). The assumptions were tested again for the calorimetry and titration data sets as they contain a number of different students than the Beer's Law data set. The MANOVAs were conducted using version 26 of SPSS.

Latent Profile Analysis

Students have many different expectations about the laboratory experience, which have been previously shown to relate to students' affective outcomes (Galloway & Bretz, 2015d). To explore what underlying groups, or profiles, of students were present in this study, a cluster analysis was performed on the data generated for each experiment. The model-based cluster analysis for latent variables is called latent profile analysis or latent

class analysis depending on the type of data used (Vermunt & Magidson, 2002). A model-based approach has the advantage of generating fit indices that are then used to directly compare different models and groupings of the data. Typical fit indices that are reported in latent profile analysis include the Bayesian information criterion (BIC), the Akaike information criterion (AIC), and the log-likelihood.

One of the most important decisions when conducting a cluster analysis is which variables to include. If too many variables are included the resulting profiles have no meaningful interpretation whereas if not enough variables are included then there is not enough variance in the data to detect meaningful profiles. Scores from the emotional satisfaction, intellectual accessibility, usefulness of lab, equipment usability, and open-endedness of lab scales were used as the clustering variables to generate student profiles based on the overall affective outcomes. Anxiety was not included as there were few differences on this scale between environments in all three experiments and as such did not add information toward meaningful profiles. As part of the interviews conducted in a previous study, students in both environments frequently reported that working with chemicals was much less a source of anxiety as compared with the social anxiety of working with other people (Hensen & Barbera, 2019). Thus, it was unsurprising that there were few differences seen between environments on anxiety despite different equipment used. For this study, the latent profile analysis was conducted using the expectation-maximization algorithm and maximum likelihood estimates. The latent profile analyses were conducted using version 5.4.3 of the mclust package in version 3.5.3 of R (Scrucca et al., 2016).

RESULTS AND DISCUSSION

Generating experiment-specific rubrics based on learning objectives

The list of experiment-specific learning objectives, generated through faculty interviews, was analyzed to determine which objectives were shared by the majority of the faculty members interviewed. As seen in Table 5.1, there were three common learning objectives for the Beer's Law experiment, four for the calorimetry experiment, and four for the titration experiment. For more information about individual faculty member's objectives, see Table B.SI2 in Appendix B.

Table 5.1: Common learning objectives across faculty interviewed

		Abbreviation
Beer's Law	Understand the relation between absorbance and concentration	BL1
	Prepare solutions	BL2
	Determine an unknown concentration using the relation between absorbance and concentration	BL3
Calorimetry	Predict the sign of the change in enthalpy for a given reaction	C1
	Determine the enthalpy change for a given reaction	C2
	Understand how to calculate a change in enthalpy from a temperature change	C3
	Understand the difference between endothermic and exothermic and how it relates to the sign of the enthalpy change	C4
Titration	Visually identify a change in pH during a titration using a mixture of indicators	T1
	Identify key points on a titration curve	T2
	Determine the pKa of an unknown analyte using a titration curve	T3
	Determine the molar mass (or mass) of an unknown analyte using a titration curve	T4

These learning objectives were then used to assess the students' ability to demonstrate evidence of meeting them in their laboratory report. To do this, a rubric was generated for each experiment. As an example, the Beer's Law rubric is shown in Table 5.2.

Table 5.2: Rubric used to score student laboratory reports for the Beer's Law experiment in each learning environment

	Does Not Meet	*Partially Meets*	*Meets*
Understand the relation between absorbance and concentration			
Prepare solutions			
Determine an unknown concentration using the relation between absorbance and concentration			

Student Population

For the Beer's Law experiment, 174 students completed the hands-on experiment and 216 students completed the virtual experiment. The following term for the calorimetry experiment, 129 students completed the hands-on experiment and 152 students completed the virtual experiment. Finally, in the last term for the titration experiment, 72 students completed the hands-on experiment and 117 students completed the virtual experiment. For more information on the student population and demographics see Table B.SI1 in Appendix B.

Assessing Differences in Learning Objectives

The laboratory reports of study participants were carefully read and the coders looked for any evidence of the students meeting the stated learning objectives noted on each rubric. For the first Beer's Law objective (BL1), an example of a student report that received a score of 'Meets' is "*A substances concentration and it's absorbance are directly proportional. A high-concentration solution absorbs more light and a low-*

concentration solution absorbs less light". This student demonstrated that they fully understood the relation. For comparison, a student report that received a score of 'Partially Meets' is "*Beer's law, which states A=ebc, lets one use the relationship between absorbance to create a calibration curve*". This student seems to have some understanding of how to use the relation but does not provide further evidence that they understand it and do not simply just understand the experimental steps. The score 'Does Not Meet' was given for any report that provided no evidence of understanding the relation. The three scoring categories were used in a similar fashion for all other learning objectives. Table 5.3 contains the results of scoring the reports and the significance of the chi-square results when comparing an objective between environments. The N/A category was used when students did not include a relevant section in the report as there was no way of judging a missing section.

Table 5.3: Percentage of students meeting learning objectives and chi-square results for all learning objectives by environment type. *Significant at p=0.05, **Significant at p=0.001

	Hands-On					*Virtual*				
	N	Does Not Meet (%)	Partially Meets (%)	Meets (%)	N/A (%)	N	Does Not Meet (%)	Partially Meets (%)	Meets (%)	N/A (%)
BL1	137	19.7	6.6	69.3	4.4	176	21.0	6.8	72.2	0.0
BL2*		0.0	0.0	5.1	94.9		0.0	0.0	0.0	100.0
BL3*		26.3	0.0	73.7	0.0		17.0	0.0	83.0	0.0
C1	110	76.4	0.0	23.6	0.0	140	84.3	0.7	15.0	0.0
C2		0.9	0.0	99.1	0.0		2.1	0.0	97.9	0.0
C3		12.7	2.7	84.5	0.0		15.0	10.0	75.0	0.0
C4		13.6	10.9	75.5	0.0		12.1	6.4	81.4	0.0
T1**	64	0.0	0.0	37.5	62.5	90	0.0	0.0	0.0	100.0
T2		9.4	25.0	65.6	0.0		12.2	25.6	62.2	0.0
T3		34.4	6.3	59.4	0.0		28.9	3.3	67.8	0.0
T4**		7.8	1.6	90.6	0.0		58.9	3.3	37.8	0.0

Most of the learning objectives showed no statistical difference between environments. However, as noted in Table 5.3 with asterisks, there were significant differences on four learning objectives. Two of these, BL2 and T1, were skill-based objectives that explicitly addressed a procedural step. Thus, it is not surprising that only a few students in the traditional hands-on environment included evidence of meeting these objectives and none of the students in the virtual environment included evidence of meeting them, as students typically do not include details about specific procedural steps in their report. Interestingly, the majority of hands-on students that did meet BL2 did not meet BL1. It is likely that these students were only able to summarize the procedural steps they conducted rather than understand and document why they conducted them. Thus, the score of 'Meets' on these skill-based objectives may not be an indication of if students learned the skill but rather their ability to write a complete laboratory report.

Students also differed on their ability to meet learning goal T4. This difference could be due to a function of the design of the LearnSmart Labs. In the traditional hands-on titration experiment students started with an unknown solid and were asked to identify the unknown by calculating the molar mass. However, in the virtual environment students started with an unknown solution and were asked to identify the unknown by calculating the pKa and then asked to calculate how much mass was initially dissolved to make the solution. While students in both environments were asked to use the equation: $Molar\ Mass = \frac{grams}{Molarity * volume}$, the virtual students frequently did not provide evidence of calculating the initial mass dissolved. Instead, the students stopped once they were able to get the identity of the acid with the pKa, as only that finding had to be reported to the TA before they could leave for the day.

The fourth objective that students differed on was their ability to use Beer's Law to calculate the unknown concentration (BL3), with a higher proportion of students that completed the experiment in the virtual environment meeting this objective. It was observed by the first author and the TAs that students in the virtual environment had more time to do the calculations, as the experiment itself did not take as long as the hands-on counterpart did. Therefore, extra time students in the virtual environment had to work on the calculations with their lab partner and/or TA could explain this higher percentage. It is possible that if each student had an equivalent amount of time to work on the calculations with assistance from a partner and/or TA that this difference would be minimized. Additionally, this finding was not significant using the stricter p-value of 0.01 to correct for multiple comparisons.

Assessing Differences in Affective Outcomes

In addition to the learning objectives, affective outcomes were also compared across environments. After checking the assumptions for running a MANOVA, there were normality and homoscedasticity violations. However, MANOVAs are robust to violations in these assumptions (Olson, 1974). For the skewness and kurtosis values see Table B.SI3 in Appendix B. MANOVAs were conducted to compare the scale scores for the anxiety, emotional satisfaction, intellectual accessibility, usefulness of lab, equipment usability, and open-endedness of lab scales. Table 5.4 consists of the results of these MANOVAs and the respective effect sizes as measured by partial eta squared. A bolded p-value indicates a significant result. A partial eta of 0.01 represents a small effect, a value of 0.06 represents a medium effect, and a value of 0.14 represents a large effect (Cohen, 1992). See Table B.SI4 in Appendix B for the averages of all six scales by experiment and environment type.

As seen in Table 5.4, for the Beer's law experiment, many of the affective scale outcomes were significantly different between environments and both the emotion satisfaction and equipment usability scales were approaching a medium effect size. The differences highlighted in orange indicate that the hands-on students had the significantly higher average whereas the difference highlighted in purple indicates the virtual students had the significantly higher average. However, this Beer's Law data was previously analyzed (Hensen & Barbera, 2019) and an instructor-effect was detected.

Table 5.4: MANOVA results of affective differences across laboratory environments

	Beer's Law		***Calorimetry***		***Titration***	
	p-value	**Effect Size**	**p-value**	**Effect Size**	**p-value**	**Effect Size**
Anxiety	0.237	0.004	0.512	0.002	0.477	0.003
Emotional Satisfaction	**<0.001**	0.049	0.478	0.001	0.110	0.003
Intellectual Accessibility	**0.001**	0.027	0.681	0.002	0.489	0.014
Usefulness of Lab	**0.001**	0.028	**0.013**	0.022	**0.017**	0.030
Equipment Usability	**<0.001**	0.056	**<0.001**	0.043	**<0.001**	0.067
Open-endedness of Lab	0.971	0.000	0.194	0.006	**0.034**	0.024

In a previously reported analysis of the Beer's Law data, Hensen and Barbera (Hensen & Barbera, 2019) noted that four TAs that taught the virtual experiment had sections with much lower averages on the emotion satisfaction scale than the other four TAs that taught the virtual experiment. As part of their analysis, a MANOVA was run with three groupings (Hands-on, Virtual A - higher emotional satisfaction, and Virtual B - lower emotional satisfaction) instead of just by learning environment. With these TA groupings, none of the affective scale results were significantly different between students in the hands-on sections and the Virtual A group. However, the emotional satisfaction, intellectual accessibility, usefulness of lab, and equipment usability scales were significantly different between students in the hands-on sections and those in the Virtual B group. No evidence of an instructor effect was found for the calorimetry or titration experiment. As both the calorimetry and titration experiments take place in later terms and the Virtual B group consisted of mostly first-year TAs with limited teaching experience, the instructor effect could have been minimized as the TAs gained experience. However, no generalizations about the effect of teaching experience can be

made from this study as TAs rotate in and out of teaching general chemistry laboratories throughout the academic year and each quarter consisted of a different combination of TAs.

For both the calorimetry and titration experiments, data from the affective scales of usefulness of lab and equipment usability showed differences between environments with the traditional hands-on students reporting higher averages for both scales (noted in orange in Table 5.4). For all experiments, the effect size of the usefulness of lab was small but the effect size of equipment usability was medium indicating that the students had minor differences on how *useful* they thought the experiment was but larger differences on their *perceived ability* to use the equipment. However, when accounting for multiple comparisons, the usefulness of lab differences are not significant at a corrected p-value of 0.01 and thus there is not enough power in this sample to make definitive conclusions about that scale. It is possible that if students utilized the virtual environment more often that they may begin to feel more comfortable using it as it does take time to get oriented with the program.

Additionally, the open-endedness of lab scale was significantly different with a small effect size for the titration experiment and the virtual students having a higher average (noted in purple in Table 5.4). However, similar to the usefulness of lab differences, when accounting for the multiple comparisons made, this finding was not significant at the stricter p-value of 0.01.

Latent Profile Analysis

The affective comparisons noted above do not evaluate differences between specific students but rather differences between environments. Therefore, latent profile analyses were conducted to investigate what groupings of students existed based on their affective characteristics. These analyses indicated that the Beer's Law and calorimetry data had four profiles (groupings of students) and the titration data had three, as shown in Table 5.5. Each analysis was run ten times, with a random order of the data, to ensure that the solutions were stable (Scrucca & Raftery, 2015). The profiles were named based on the defining characteristics of the affective scale scores. More detailed information on the process of selecting the best fitting profiles using mclust is contained in Table B.SI5 in Appendix B.

Table 5.5: Number of students in each profile

		N
Beer's Law	Low	83
	Medium	209
	High	78
	Mixed	20
Calorimetry	Low	22
	Medium	67
	High	111
	Very High	81
Titration	Low	33
	Medium	100
	High	56

For each experiment, there were three similar groupings: low, medium, and high affective outcomes. For more information on the scale averages by grouping see Table B.SI6 in Appendix B. These groupings are similar to previous cluster analysis results

found by Galloway and Bretz (Galloway & Bretz, 2015d). There were also groupings that were unique to an experiment. For the Beer's Law experiment, there was a grouping of students that had low averages on the emotional satisfaction and intellectual accessibility scales but high averages on the usefulness of lab, equipment usability, and open-endedness of lab scales. This indicated mixed outcomes where the students thought the experiment worked well and was useful but still found it to not be accessible or emotionally satisfying. Also, as noted earlier, the calorimetry experiment had a 'very high' profile. It was unsurprising that many students reported high affective outcomes for the calorimetry experiment because both the hands-on and virtual versions of this experiment involved relatively few experimental steps and were shorter than other experiments conducted that term.

While there was a range of affective outcomes across each experiment, interestingly, as seen in Figure 1, the average report score across profiles was consistent indicating that it may not be possible to identify which students had poor affective outcomes solely based on their academic performance in the laboratory. In other words, a student that did very well on the laboratory report may still have had low affective outcomes and vice versa. To investigate this further, the profiles were examined by individual learning objective rather than an overall grade for better resolution.

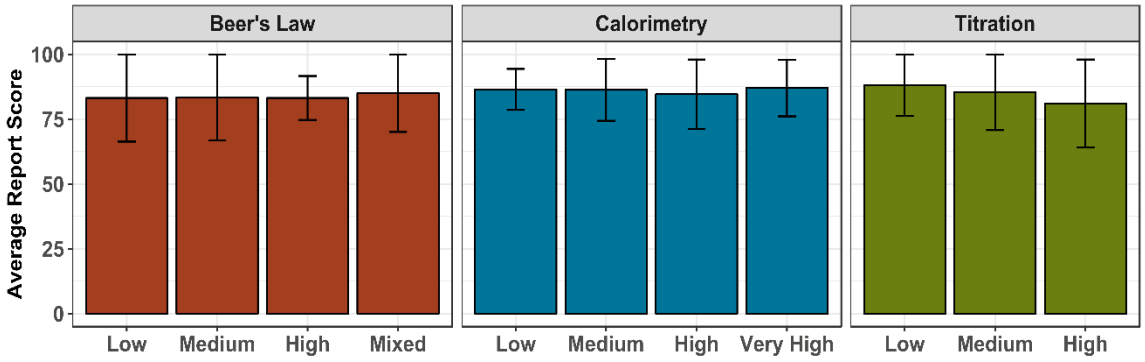


Figure 5.1: Average report score by profile and experiment

The percent of students in each rubric category for each learning objective are shown in Figure 2. Learning objectives BL2 and T1 were not included in their respective analyses as no students in the virtual environment, and few students in the traditional hands-on environment met them. Despite providing a more detailed view of the cognitive outcomes, the lack of differences in learning objectives was similar to the lack of differences seen in the report scores, adding more evidence that it was not possible to identify which students were in each grouping based on their laboratory reports. For example, on objective C3 (understand how to calculate a change in enthalpy from a temperature change) there were approximately equal percentages of students that either met or partially met the objective despite differences in affective outcomes. Similarly, there were no major differences between the students' ability to meet the learning objectives for the majority of the objectives (BL1, C2, C3, C4, T2, and T3).

While the majority of the learning objectives had no differences based on learning profile, three differences were observed. As can been seen in Figure 2 for objective BL3 (determining an unknown concentration), the 'low' affective group had the highest percentage of students meeting the objective. However, this difference may be an artifact

of the low grouping itself having a higher percentage of virtual students, seen in Figure B.SI1. As noted earlier and seen in Table 5.3, the virtual experiment was observed to be shorter, which led to the virtual students meeting this goal more often. The fact that the virtual students may have had more time to work through the calculations with their lab partner and/or TA, combined with the fact that more virtual students are in the low profile, provide a possible reason for the higher percentage of students in the low-profile meeting BL3. Similarly, the virtual students did not meet learning objective T4 (determining molar mass) as often as the hands-on students. As seen in Table 5.3, the majority of virtual students failed to meet this learning objective whereas the majority of hands-on students did meet this objective due to procedural differences. Thus, the higher percentage of virtual students in the low affective group (as seen in Figure B.SI1) explains a possible reason why the low affect group did not meet this learning objective as often as the other groups. The third difference was seen with learning objective C1 (predict the sign of the change in enthalpy for a given reaction). This difference is likely a function of the small sample size for the low affective profile. There were only 22 students in the low affective profile for this experiment, which means that each student represents 4.5% of the data plotted in Figure 2. Given the lack of differences on the majority of objectives and laboratory report scores, the difference seen in learning goal C1 is most likely contributed to sample size limitations. It is possible that the difference may not be observed in studies with a larger sample size.

Overall, the majority of learning objectives, in addition the laboratory report scores, showed no difference between affective groups. This highlighted that solely

relying on differences in cognitive outcomes to determine if an intervention is successful fails to differentiate between students with low and high affective outcomes. While there is a body of literature that has found students in the virtual experiment are able to meet the cognitive outcomes similarly (Enneking et al., 2019; Hawkins & Phelps, 2013) or outperform the hands-on students (Pyatt & Sims, 2012; Tatli & Ayas, 2010; Winkelmann et al., 2017), future research should ensure that affective outcomes are measured and focus on how to identify the students in the low affective profiles in order to target laboratory interventions and ensure all students are having a positive laboratory experience.

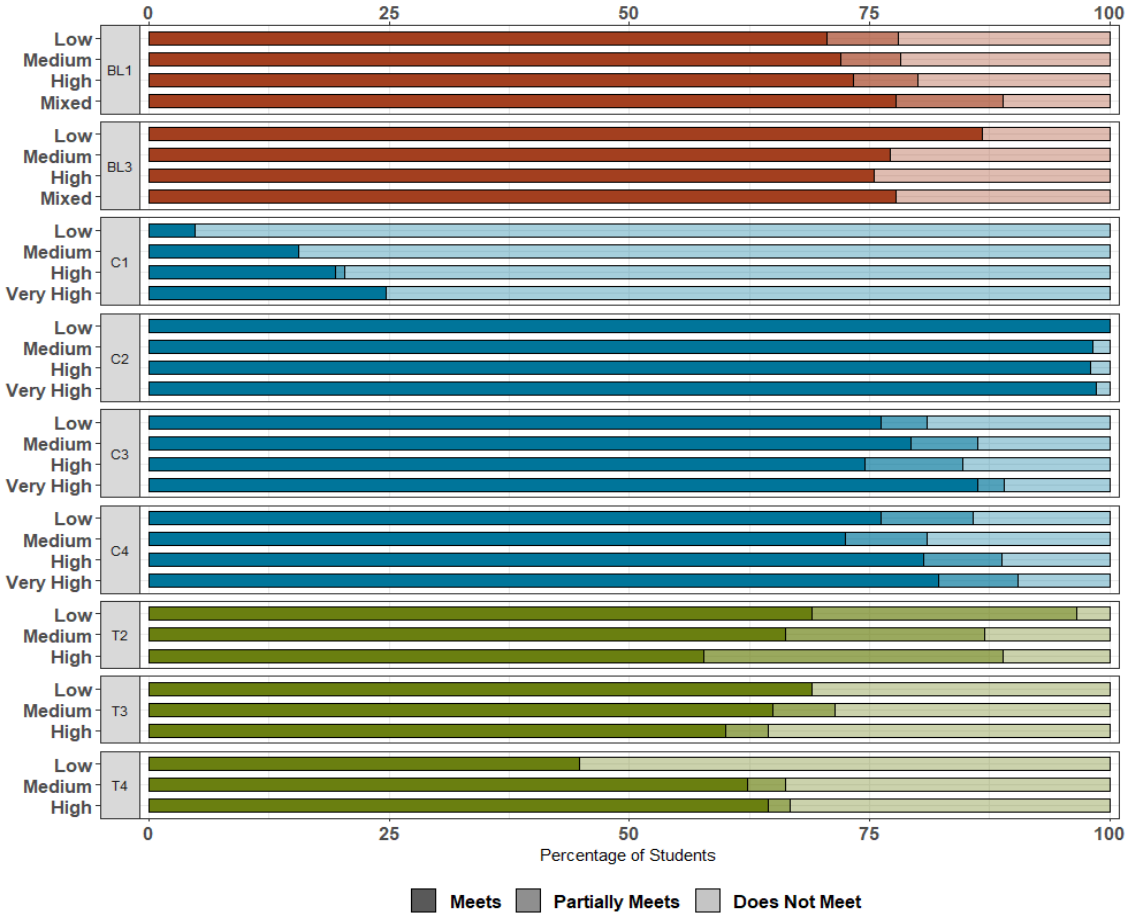


Figure 5.2: Percent of students in each rubric category by learning objective

CONCLUSION

After comparing student outcomes across three experiments conducted in both a virtual and a hands-on environment, differences were detected on four of the eleven common learning objectives. Two of the differences were on skill-based objectives and the other two were on objectives related to the outcome of specific calculations within an experiment. While statistically different outcomes were detected, the results are likely due to alignment issues with experimental procedures and report requirements. For the Beer's Law experiment, differences were seen on learning objectives BL2 and BL3. The differences seen on objective BL2 could be contributed to the report requirements as this was a skill-based objective and the report requirements did not include having students explicitly write about the procedural steps they completed. Additionally, in the Beer's Law experiment students in the virtual environment were observed to take less time to complete the experiment which freed up more time to work through the analysis of data with their lab partner and/or TA and thus they met learning objective BL3 more often. Similar to the Beer's Law experiment, students in the titration experiment also did not provide evidence of meeting a skill-based objective, T1, and an objective that had procedural differences between the two learning environments, T4. Overall, the students in the virtual environment consistently struggled to provide evidence of meeting skill-based learning objectives and outcomes designed around specific procedural steps due to differences in the procedures between environments and the report requirements. This result is similar to previous findings that specific differences between procedures in the learning environments account for the differences observed (Hawkins & Phelps, 2013).

Therefore, careful design of the experiments and assessments should take place to ensure that students have the opportunity to equally meet the desired experimental learning objectives. If students cannot meet the learning objective in a given learning environment, then that environment should not be used for that experiment. Overall, if students had equal time to work on processing the data and identical procedures in both environments, the differences found in these learning objectives would be greatly minimized.

In addition to investigating differences in the learning objectives between environments, six different affective outcomes (anxiety, intellectual accessibility, emotional satisfaction, equipment usability, usefulness of lab, and open-endedness of lab) were monitored. For most, no detectable differences were found. However, across all three experiments, students in the virtual environment reported lower averages on the equipment usability and usefulness of lab scales. This finding provides unique insight into what differences may exist between learning environments. Historically, the affective domain is understudied, and thus, past studies have focused on cognitive outcomes. While very minor cognitive differences were found, the affective differences highlight larger discrepancies between the learning environments. While an instructor effect was found for the Beer's Law experiment, no evidence of an instructor effect was found for either the calorimetry or titration experiments. This effect was either minimized as TAs gained experience or that the instructor effect was specific to individual TAs who did not teach laboratory sections in subsequent terms.

As the result of a latent profile analysis, there were also differences on affective outcomes based on individual students regardless of which environment they completed the experiment in. For each experiment, the majority of students had medium and high affective outcomes. However, there were still a fair number of students that had low affective outcomes. This may be a function of the wide variety of students that enroll in the chemistry laboratory with many different backgrounds and career paths. The one-size-fits-all approach may work for a large majority of the students but it is possible that select students may benefit from different types of laboratories.

Recently, there has been a call to conduct more research on the laboratory environment and what the role of the laboratory is (Bretz, 2019). A unique challenge of the laboratory is that it often consists of multiple sections taught by multiple instructors and the student population is made of diverse majors. Thus, to study the laboratory effectively requires researchers to carefully consider how to control for a wide range of confounding variables that exist in the natural setting of a laboratory course rather than conduct controlled studies that rely on volunteers that do not necessarily represent the average student population. Once more research that carefully controls for the confounding variables present in the laboratory setting is conducted, there may be a better sense of which students benefit from the current model of the laboratory and which do not.

LIMITATIONS

To minimize changes to the curriculum at Portland State University, learning objectives were assessed using the assessments already in place. This meant that the

tactile learning objectives were evaluated using the laboratory report instead of a laboratory practical. It is possible that the differences seen in the skill-based learning objectives would be different if a laboratory practical was utilized. Additionally, McGraw-Hill generously allowed us to use the LearnSmart Labs as the virtual platform. However, this meant that there was no control over the elements of the procedural design in the virtual environment. It is possible that a different virtual environment made to specifically target desired learning outcomes could produce different findings. The ability (or lack of it) to control procedural design could also impact learning objectives that are specific to an institution. For example, one institution uses nanomaterials for their Beer's Law experiment and has learning objectives directly related to using nanomaterials. Thus, if the institution uses a virtual environment that is not customizable it may not be possible for students to meet institution-specific learning objectives.

Beyond experimental limitations, this research took place at Portland State University, an urban Pacific Northwest university, and as such the findings should not be generalized to other settings without future work being conducted. Future studies would benefit from the inclusion of other settings, such as community college populations or those where students have more exposure to the virtual learning environment. While two of the five faculty interviewed were professors at the institution the data was collected at, the learning objectives reported by the five faculty members are not comprehensive, therefore, it is possible that faculty members at other institutions place different value on the objectives presented. Additionally, previous work found an instructor effect existed for the affective outcomes in the Beer's Law experiment. While this effect may be

minimized as TAs gain experience, it is also possible that it was specific to individual TAs. Therefore, instructor effect should be examined or controlled for in future research to ensure that outcomes are not a result of who is teaching the section.

With the current sample it was not possible to further investigate the characteristics of the profiles based on demographics. With a more adequate sample size, it would possible to evaluate for measurement invariance by demographic group to ensure that members across groups of interest are interpreting the items in a similar fashion. Once measurement invariance is established, the profiles could be further compared on the demographic variable of interest.

IMPLICATIONS FOR FUTURE RESEARCH

This study expanded on previous research to investigate the learning objectives and affective outcomes for a range of experiments. Based on the findings, there is a need to conduct future studies using laboratory practical exams to investigate the tactical learning objectives. Previous research (Hawkins & Phelps, 2013; Winkelmann et al., 2017) has found no differences on the students' ability to complete skill-based learning objectives but more work in this area is warranted. As a possible instructor effect was found with one experiment but not the other experiments, future research should choose research designs that allow for a true treatment-control study to be conducted where the same instructor is teaching in both environments. Additionally, there is a need for qualitative studies to further investigate the affective grouping of students. These studies could help identify the nature of the defining characteristics within the groupings. With this information, curriculum reform could then take place to target these groupings to

ensure that more students have a laboratory experience that produces positive affective, cognitive, and psychomotor outcomes.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available in Appendix B. It includes the interview protocol, list of learning objectives, demographics, skew and kurtosis values, affective averages by environment and experiment, BIC indicies, and affective averages by profile and experiment.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jbarbera@pdx.edu

ACKNOWLEDGMENTS

We would like to thank Medina Glenn and Amanda Schmidt for their help with this study. We would also like to thank the students, graduate teaching assistants, the laboratory coordinator, and stockroom staff involved with this work. Without these people, this study would not have happened. Also, many thanks to McGraw-Hill Education who generously allowed the use of the LearnSmart Labs platform for this study.

CHAPTER VI: Conclusions, Implications, and Future Research

Conclusions

This dissertation research addressed three distinct gaps in the literature. The first major gap is that differences in affective outcomes between virtual and hands-on learning environments have previously gone largely unstudied. To address this, existing affective scales were modified and adapted for use in the laboratory. These newly created scales were then used to compare the two learning environments on affective outcomes. The second gap in the literature is that the majority of existing studies comparing the two learning environments only focus on a single experiment leaving outstanding questions about if the findings from single-experiment studies were specific to the experiment topic studied or if the specific procedure mattered when making the comparison. This was addressed by including three distinct experiments across the General Chemistry curriculum to detect what, if any, differences existed between experiments on cognitive and affective outcomes. The final gap in the literature that was address was measuring cognitive domain differences through learning objectives. While there is existing literature that found little to no differences in the cognitive domain (Hawkins & Phelps, 2013; Pyatt & Sims, 2012; Winkelmann et al., 2017; Winkelmann et al., 2014; Woodfield et al., 2005; Woodfield et al., 2004), the existing studies rely on self-authored tests and quizzes rather than learning objectives. Learning objectives can give a more detailed picture into if students are meeting the overall goals of the course rather than answering a simple multiple-choice item correctly. As such, student cognitive outcomes were measured through learning objectives to determine similarities and differences between

the two learning environments. This dissertation research addresses these three important gaps in the current literature through five integrated questions.

Research Question 1: To what degree can previously developed scales be adapted for use in the laboratory environment?

The lack of affective studies on virtual general chemistry experiments was addressed by measuring six specific constructs that could potentially be different between the environments. Three of these constructs, anxiety, intellectual accessibility, and emotional satisfaction, had been previously measured using items from the Attitude towards the Subject of Chemistry Inventory (ASCI) (Bauer, 2008). Additionally, another instrument focusing on anxiety previously existed in the literature, the Chemistry Laboratory Anxiety Instrument (CLAI), and items from this instrument were also used to inform the anxiety scale used in this research (Bowen, 1999). Minor wording changes were made to the previously established items to change the frame of reference on the existing scales to the laboratory environment. The remaining three constructs, usefulness of lab, equipment usability, and open-endedness of lab, were measured with select items from the Virtual and Physical Experimentation Questionnaire (VPEQ) (Pyatt & Sims, 2012). Items addressing the six affective constructs chosen for this research were selected by modifying items from the existing ASCI, CLAI, and VPEQ instruments. These items then underwent further psychometric testing to ensure functionality in the laboratory setting.

Research Question 2: What evidence of validity and reliability supports data collected with the adapted scales?

After selecting and modifying existing affective items for use in the laboratory context, further psychometric analysis was done. Whenever scales are adapted for a new context, there needs to be sufficient evidence of validity and reliability presented to support the use of the items. To start, response process validity interviews were conducted with the modified items of the existing scales to determine if the way in which students were interpreting the items was consistent with the original intended meaning. Based on these interviews, one item was removed. The final set of items (found in Appendix A) was administered to students after they completed either a virtual experiment or the equivalent hands-on experiment. Students' responses to these items were analyzed using confirmatory factor analysis (CFA) to determine if there was evidence of structural validity. Structural validity is an indication that the items are informing the intended constructs and elicits their relation with other constructs. Fit indices are used as an indication of how well the data generated by the items fit the proposed model. After making minor modifications, the CFI and SRMR fit indices for individual one-factor CFA models as well as two- and three-factor CFA models were within the acceptable range and thus indicated sufficient evidence for structural validity. All correlations, with the exception of the correlation between emotional satisfaction and intellectual accessibility, between factors in the two- and three-factor models closely matched previous findings indicating that the modifications and the adaption to a new context had a limited effect on the validity of the items. The correlation between the

emotional satisfaction and intellectual accessibility scales was 0.91, which suggested that there was little separation between subconstructs. However, these were kept separate rather than collapsing into a single attitude scale due to the previous theoretical justification put forth by Xu and Lewis (2011). Finally, McDonald's Omega was measured to indicate how well the items within a construct were similar to one another, all omega values were sufficiently high indicating evidence towards the single-administration reliability of the scales.

Research Question 3: How do various aspects of the affective domain compare when students complete a general chemistry experiment virtually as opposed to a traditional hands-on experiment?

Once evidence was provided that the items from the six scales were functioning in the laboratory environment, the items were administered after three common chemistry laboratory experiments. The experiments chosen were a Beer's Law experiment, a calorimetry experiment, and a titration experiment, thereby spanning content across the three terms of General Chemistry. When comparing students that completed a virtual Beer's Law experiment with students that completed a hands-on Beer's Law experiment, the students that completed the virtual Beer's Law experiment had lower averages on the emotional satisfaction, intellectual accessibility, usefulness of lab, and equipment usability scales. However, an instructor effect was detected and therefore it was possible that who the student had as their instructor influenced their responses to the affective scales. At Portland State University, all instructors were graduate teaching assistants

(TAs) and it was determined that the majority of first-year TAs that taught the virtual experiment had students with lower affective averages. While it cannot be definitively concluded that their lack of teaching experience contributed to the lower averages, it is one possibility. Previous research has found that one of the most important indicators of an effective TA is that they are knowledgeable not only in the content but also in how to teach (Herrington & Nakhleh, 2003). Establishing an effective training program can help give first year TAs the tools they need to be an effective teacher (Marbach-Ad et al., 2012). The fact that an instructor effect was not detected in the second and third terms could indicate that as TAs gain experience this effect is minimized, or it could instead be an artifact of specific TAs that did not teach in the second or third terms. The teaching assignments change term by term and thus not all teaching assistants taught all three terms.

In the calorimetry experiment, the students that completed the virtual experiment had lower averages on the usefulness of lab and equipment usability scales, similar to the Beer's Law experiment, however, they had higher averages on the open-endedness of lab scale. This indicated that while they found the experiment to be less useful and the equipment harder to use, on average, they enjoyed that the procedure took less time to complete than it would have taken to complete a typical hands-on experiment. Like the other two experiments, students that completed a virtual titration experiment in the third term also reported lower averages on the usefulness of lab and equipment usability scales. With all three experiments having a similar outcome, it is likely that the students at this

institution found the virtual experiments, in general, to be less useful and harder to use than their hands-on counterpart and that it was not specific to an experiment type.

Overall, across the three experiments no difference between learning environments was detected for the anxiety, intellectual accessibility, and emotional satisfaction scales after accounting for an instructor effect while there was a difference between learning environments detected for the usefulness of lab and equipment usability scales.

Research Question 4: To what degree can experiments in a virtual environment meet the same learning objectives as similar traditional hands-on experiments?

In addition to the affective scale data collected, cognitive outcomes were also measured. Student laboratory reports were collected for each experiment and were scored using a rubric to determine whether the student did not meet, partially met, or met common learning objectives. These objectives were for a given experiment regardless of which environment they completed the experiment in and were based on learning objectives stated during interviews with five faculty members from three different institutions. After scoring the reports for the three different experiments, differences on four learning goals were found. The differences on the second learning objective for the Beer's Law experiment (BL2), which was "prepare solutions", and the first objective for the titration objective (T1), which was "visually identify a change in pH during a titration using a mixture of indicators", were seen because these objectives were skill-based and students struggled to provide evidence of meeting these learning objectives in a written

report. It is possible that these differences would not be present if the objectives were measured through a laboratory practical instead of the report. The difference on the fourth learning objective on the titration experiment (T4), which was determining the molar mass, was due to differences in the procedures between the two environments and thus it is possible that if the procedures matched identically that this difference would be minimized. The difference on the third learning objective for the Beer's Law experiment (BL3), which was determining an unknown concentration using Beer's Law, was a function of the virtual experiment taking less time and thus the students had more time to work on processing the data. Had both environments had the same amount of time for processing the data, it is possible this difference would also be minimized. These four detected differences were all minor in nature and could be minimized in future studies by making careful considerations in the experimental and course designs. There were no major differences detected between the two environments regarding students meeting learning objectives, indicating that generally, students were able to meet, or not meet, the learning objectives to the same degree in both environments.

Research Question 5: To what degree do student characteristics explain differential outcomes in a general chemistry laboratory course?

To better understand what groups of students were present in the laboratory, latent profile analysis was conducted using five affective scales as the clustering variables. This type of analysis allows individual differences to be further investigated as it was possible that there were groups of students that had an experience that was different from the

average experience. The profile analysis revealed three groupings of students that were consistent across all three experiments. There were student groups with high, medium, or low affective outcomes regardless of which learning environment they completed the experiment in. This indicated while the majority of students are in the medium and high affective profile, there was anywhere from four to seventeen percent of the students that left a given experiment having low affective outcomes regardless of which learning environment they completed the experiment in. Further analysis revealed that the average laboratory report score across the three groupings was consistent and thus indicated it was not possible to identify the students in the low affective profile by cognitive measures alone. Not only were the scores consistent, the students' ability to meet the learning objectives was largely consistent regardless of which profile the student belonged to. This highlights the need to not solely rely on cognitive measures to determine if an intervention is successful or not but instead affective measures should be included in intervention studies to ensure all students are having a positive experience in the laboratory.

Future Research

As institutions face challenges meeting the demands of increasing enrollment and decreasing budgets in General Chemistry laboratories, some have implemented virtual experiments as a way of increasing capacity and/or decreasing budget costs. However, it remained unclear in the literature if students could have similar outcomes from the virtual environment as opposed to coming into the traditional environment to complete the experiment. Specifically, if outcomes in both the affective and cognitive domains were

similar. To measure these affective outcomes, it was necessary to have affective items that have been adapted for the laboratory. Few laboratory-specific affective instruments existed prior to this research and thus items covering six constructs were adapted for the laboratory and then psychometrically evaluated. These items can be used in future studies for a wide range of contexts to measure students' affective outcomes in the laboratory when comparing an intervention, such as using argument-driven inquiry experiments (Walker et al., 2011), to the traditional laboratory experiments.

Future research would benefit from further examining the relationship between the emotional satisfaction and intellectual accessibility scales. Previous research has made a case for separating out the scales as part of a larger model but the correlations found in this research suggest that in future work it could be beneficial to collapse the items from the two scales into one scale titled 'Attitude'. Alternatively, if the behavioral and cognitive components of attitude are important for future research, additional items could be added to theoretically justify the separation of the scales.

In addition to affective outcomes, one major finding from this research is that only minor differences were detected in the cognitive outcomes across learning environments, and with careful consideration in the implementation, those differences could be further minimized. Institutions can use this finding to better inform their decisions and researchers can use this to build future studies. In the affective domain, students reported that the virtual experiments were less useful and harder to use than the hands-on experiments. To minimize these differences in future implementations, institutions should consider how to make the virtual experiments feel more connected to

the students' real-life and increase the training opportunities on the software. For example, a virtual experiment could include analyzing water downstream from a factory for contaminants. Future studies should consider measuring the skill-based objectives explicitly with laboratory practical exams rather than reports, to directly measure how students in the two learning environments compare. Researchers should also take extra precaution to make the procedures in the two learning environments as similar as possible. Previous work on virtual experiments (Hawkins & Phelps, 2013), found that minor differences between procedures can influence students' perception of the learning environments as well as their ability to reach the desired outcome.

A minor finding that researchers should also consider in future studies is the possibility of an instructor effect. One challenge with studying the laboratory is that at many institutions the sections are taught by multiple graduate students (TAs) with varying levels of interest and experience (Abraham et al., 1997). While TA training and weekly meetings are often conducted to ensure all sections have a similar experience, it is possible that differences between instructors have an effect on student outcomes. For example, student A may have the same course number at the same time in the room next to student B but they may leave their section with very different affective and cognitive outcomes based on their TA. While important to study, one issue a large number of studies run into is that there is not a sufficient sample size to have enough power to detect statistically significant differences by class section or TA. Thus, future studies should include a research design that ensures there is a large enough sample size to capture differences between TAs and how those differences are influencing student outcomes in

the course. If large enough sample sizes are not possible, well thought out qualitative studies can be conducted to gain a deeper understanding of what differences exist between TAs and how this might affect the students' outcomes. In future work it would still be possible to minimize the instructor effect by having the same instructor teach one section with virtual experiments and one with the hands-on experiments. In that design, it would be likely that any differences seen between sections could be contributed to differences in the learning environment and not differences between TAs.

Recommendations for Implementation

If institutions choose to implement virtual experiments in the General Chemistry laboratory, there are several recommendations for best practice based on this study. First, it is recommended that the virtual experiment implemented is based on real-world examples. This can be done two different ways. First, in cases where no context is provided within the platform hosting the virtual environment, the instructor of the laboratory could provide relevant context for the experiment and procedure and how it relates to a health concern, industrial application, or other real-world context of interest relevant to their students. Alternatively, the virtual experiment could have these real-world applications naturally built into the context of the experiment. Such as analyzing water for contaminates. Situating the experiment within a broader context allows students to connect with the material rather than feel that the virtual environment is not a useful experiment. This becomes increasingly important when virtual experiments are utilized at institutions that use problem-based experiments or other curricula for the hands-on experiments that situate experiments in real-life scenarios. One major advantage virtual

experiments have is their ability to complete experiments in situations that cannot be easily replicated in the laboratory. This should be taken advantage of and experiments should be selected that allow for traditionally difficult situations to be analyzed such as repeating experiments that use expensive reagents or that are time intensive.

The second recommendation is that students be given ample opportunities to learn the virtual software. The virtual software may not be intuitive for all students and thus there is a need to have the students complete practice experiments where they can get comfortable with the software without the added pressure of also finishing the experiment.

The final recommendation is that institutions wishing to implement virtual experiments should only do so with TAs that are informed about the virtual experiments and have had adequate pedagogical training. How a TA responds to a question about a virtual experiment can significantly influence how the students feel about the experiment. For example, if a student asks a question about the data they got from the experiment and the TA responds with “don’t worry about that, it is not a real experiment”, this could give the student the impression that the TA does not think it is worth their time. However, if the TA responds with “That is a great question! The data looks that way because….”, the impression could instead be that the TA is excited about the virtual experiment and views it as valuable. It is important to note that while implementation may vary and may not directly involve at TA, that any personnel that interact with students about the virtual experiment should be informed about the virtual environment and understand the potential benefits. It is the author’s belief that with experiments situated in real-world

examples, ample time to learn the software, and extensive TA training, the differences seen in the affective domain would decrease and the two learning environments would be approximately equivalent at the General Chemistry laboratory for the affective and cognitive domain outcomes tested in this study.

Overall, this research has begun to address the efficacy of virtual experiments. Institutions can use this research to make better-informed decisions about if and how they want to implement virtual experiments. Researchers now have access to affective items on six constructs that have been adapted for the laboratory environment (virtual or not) and have been psychometrically evaluated. See Appendix A for a copy of the six scales used for this study. These items can be applied to a wide range of studies to better understand the affective domain in the laboratory environment. Specifically, for virtual experiments, researchers can build upon this work and further explore differences between learning environments in the affective and cognitive domains as well as add information to the literature about potential differences in the psychomotor domain.

REFERENCES

Abell, T. N., & Bretz, S. L. (2018). Dissolving Salts in Water: Students' Particulate Explanations of Temperature Changes. *Journal of Chemical Education, 95*(4), 504-511.

Abouserie, R. (1994). Sources and levels of stress in relation to locus of control and self esteem in university students. *Educational Psychology, 14*(3), 323-330.

Abraham, M. R., Craolice, M. S., Graves, A. P., Aldhamash, A. H., Kihega, J. G., Gal, J. G. P., & Varghese, V. (1997). The nature and state of general chemistry laboratory courses offered by colleges and universities in the United States. *Journal of Chemical Education, 74*(5), 591-594.

ACS. (2015). ACS Guidelines for Chemistry in Two-Year College Programs. In. Washington, DC: Fensham 1992.

ACS. (2019). Chemistry Careers. Retrieved from https://www.acs.org/content/acs/en/careers/college-to-career/chemistry-careers.html

Adams, G. R., & Fitch, S. A. (1983). Psychological environments of university departments: Effects on college students' identity status and ego stage development. *Journal of Personality and Social Psychology, 44*(6), 1266.

Adams, W. K., Wieman, C. E., Perkins, K. K., & Barbera, J. (2008). Modifying and validating the Colorado Learning Attitudes about Science Survey for use in chemistry. *Journal of Chemical Education, 85*(10), 1435-1439.

Alkan, F., & Koçak, C. (2015). Chemistry laboratory applications supported with simulation. *Procedia-Social and Behavioral Sciences, 176*, 970-976.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational Psychological Testing. (1999). *Standards for educational and psychological testing*: Amer Educational Research Assn.

Anderson, L. W., & Bourke, S. F. (1981). *Assessing affective characteristics in the schools*: Routledge.

Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education, 90*(5), 536-545.

Ashley, K., Cordell, D., & Mavinic, D. (2011). A brief history of phosphorus: from the philosopher's stone to nutrient recovery and reuse. *Chemosphere, 84*(6), 737-746.

Astin, A. W. (1977). *Four Critical Years. Effects of College on Beliefs, Attitudes, and Knowledge*. California: Jossey-Bass Publishers.

Bandura, A. (1997). *Self-efficacy: The exercise of control*: Macmillan.

Banerjee, P. P., Luciano, C. J., Lemole Jr, G. M., Charbel, F. T., & Oh, M. Y. (2007). Accuracy of ventriculostomy catheter placement using a head-and hand-tracked high-resolution virtual reality simulator with haptic feedback. *Journal of Neurosurgery*(107), 515-521.

Barbera, J., & VandenPlas, J. (2011). All assessment materials are not created equal: the myths about instrument development, validity, and reliability. In *Investigating classroom myths through research on teaching and learning* (Vol. 1074, pp. 177-193).

Bauer, C. F. (2005). Beyond "student attitudes": Chemistry self-concept inventory for assessment of the affective component of student learning. *Journal of Chemical Education, 82*(12), 1864-1870.

Bauer, C. F. (2008). Attitude toward chemistry: a semantic differential instrument for assessing curriculum impacts. *Journal of Chemical Education, 85*(10), 1440-1445.

Beran, J. A. (2010). *Laboratory manual for principles of general chemistry*: John Wiley & Sons.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain*. New York: David McKay.

Bogner, A., & Menz, W. (2009). The theory-generating expert interview: epistemological interest, forms of knowledge, interaction. In *Interviewing Experts* (pp. 43-80). London: Palgrave Macmillian.

Böhmová, H., & Šulcová, R. (2007). Chemistry experiment in distance education. *Problems of Education in the 21st Century, 2*.

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of Extension, 50*(2), 1-5.

Boschmann, E. (2003). Teaching chemistry via distance education. *Journal of Chemical Education, 80*(6), 704-708.

Bowen, C. W. (1999). Development and score validation of a chemistry laboratory anxiety instrument (CLAI) for college chemistry students. *Educational and Psychological Measurement, 59*(1), 171-185.

Brandriet, A. R., Xu, X., Bretz, S. L., & Lewis, J. E. (2011). Diagnosing changes in attitude in first-year college chemistry students with a shortened version of Bauer's semantic differential. *Chemistry Education Research and Practice, 12*(2), 271-278.

Bretz, S. L. (2019). Evidence for the Importance of Laboratory Courses. *Journal of Chemical Education, 96*(2), 193-195. doi:10.1021/acs.jchemed.8b00874

Bretz, S. L., Fay, M., Bruck, L. B., & Towns, M. H. (2013). What faculty interviews reveal about meaningful learning in the undergraduate chemistry laboratory. *Journal of Chemical Education, 90*(3), 281-288.

Brewer, S. E., Cinel, B., Harrison, M., & Mohr, C. L. (2013). First year chemistry laboratory courses for distance learners: Development and transfer credit acceptance. *The International Review of Research in Open and Distributed Learning, 14*(3), 488-507.

Brinson, J. R. (2015). Learning outcome achievement in non-traditional (virtual and remote) versus traditional (hands-on) laboratories: A review of the empirical research. *Computers & Education, 87*, 218-237.

Bruck, A. D., & Towns, M. (2013). Development, implementation, and analysis of a national survey of faculty goals for undergraduate chemistry laboratory. *Journal of Chemical Education, 90*(6), 685-693.

Bruck, L. B., Towns, M., & Bretz, S. L. (2010). Faculty perspectives of undergraduate chemistry laboratory: Goals and obstacles to success. *Journal of Chemical Education, 87*(12), 1416-1424.

Buck, L. B., Bretz, S. L., & Towns, M. H. (2008). Characterizing the level of inquiry in the undergraduate laboratory. *Journal of College Science Teaching, 38*(1), 52-58.

Burke, K., Greenbowe, T. J., & Hand, B. M. (2006). Implementing the science writing heuristic in the chemistry laboratory. *Journal of Chemical Education, 83*(7), 1032-1038.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research, 9*(3), 240-244.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.

Canpolat, N., Pinarbasi, T., & Sözbilir, M. (2006). Prospective teachers' misconceptions of vaporization and vapor pressure. *Journal of Chemical Education, 83*(8), 1237-1242.

Carnegie Commission. (2015). *The Carnegie Classification of Institutions of Higher Education*. Retrieved from Bloomington, IN:

Casanova, R. S., Civelli, J. L., Kimbrough, D. R., Heath, B. P., & Reeves, J. H. (2006). Distance learning: A viable alternative to the conventional lecture-lab format in general chemistry. *Journal of Chemical Education, 83*(3), 501-507.

Chan, G. C.-Y., & Chan, W. (2001). Beer's law measurements using non-monochromatic light sources—a computer simulation. *Journal of Chemical Education, 78*(9), 1285-1288.

Chan, J. Y., & Bauer, C. F. (2014). Identifying at-risk students in general chemistry via cluster analysis of affective characteristics. *Journal of Chemical Education, 91*(9), 1417-1425.

Chan, J. Y., & Bauer, C. F. (2015). Effect of peer-led team learning (PLTL) on student achievement, attitude, and self-concept in college general chemistry in randomized and quasi experimental designs. *Journal of Research in Science Teaching, 52*(3), 319-346.

Chatterjee, S., Williamson, V. M., McCann, K., & Peck, M. L. (2009). Surveying students' attitudes and perceptions toward guided-inquiry and open-inquiry laboratories. *Journal of Chemical Education, 86*(12), 1427-1432.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255.

Chini, J. J., Madsen, A., Gire, E., Rebello, N. S., & Puntambekar, S. (2012). Exploration of factors that affect the comparative effectiveness of physical and virtual manipulatives in an undergraduate laboratory. *Physical Review Physics Education Research, 8*(1), 010113:010111-010112.

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

Choi, A., Hand, B., & Greenbowe, T. (2013). Students' written arguments in general chemistry laboratory investigations. *Research in Science Education, 43*(5), 1763-1783.

Clark, T. M., & Chamberlain, J. M. (2014). Use of a PhET interactive simulation in general chemistry laboratory: Models of the hydrogen atom. *Journal of Chemical Education, 91*(8), 1198-1202.

Cobb, S., Heaney, R., Corcoran, O., & Henderson-Begg, S. (2009). The learning gains and student perceptions of a Second Life virtual lab. *Bioscience Education, 13*(1), 1-9.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

CollegeBoard. (2013). AP Chemistry Guided Inquiry Experiments: Applying the Science Practices Student Manual. In: College Board New York, NY.

Coolican, H. (2017). *Research Methods and Statistics in Psychology*. New York: Psychology Press.

Cooper, M. M. (1994). Cooperative chemistry laboratories. *Journal of Chemical Education, 71*(4), 307.

Cooper, M. M. (2005). *Cooperative Chemistry Lab Manual*: McGraw-Hill Science, Engineering & Mathematics.

Cooper, M. M., & Sandi-Urena, S. (2009). Design and validation of an instrument to assess metacognitive skillfulness in chemistry problem solving. *Journal of Chemical Education, 86*(2), 240-245.

Corter, J. E., Esche, S. K., Chassapis, C., Ma, J., & Nickerson, J. V. (2011). Process and learning outcomes from remotely-operated, simulated, and hands-on student laboratories. *Computers & Education, 57*(3), 2054-2067.

Crocker, J., Luhtanen, R., Blaine, B., & Broadnax, S. (1994). Collective self-esteem and psychological well-being among White, Black, and Asian college students. *Personality and Social Psychology Bulletin, 20*(5), 503-513.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Dalgarno, B., Bishop, A. G., Adlong, W., & Bedgood, D. R. (2009). Effectiveness of a virtual laboratory as a preparatory resource for distance education chemistry students. *Computers & Education, 53*(3), 853-865.

Davenport, J., Rafferty, A., Timms, M., Yaron, D., & Karabinos, M. (2012). *Chemvlab+: evaluating a virtual lab tutor for high school chemistry.* Paper presented at the International Conference of the Learning Sciences (ICLS).

De Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and virtual laboratories in science and engineering education. *Science, 340*(6130), 305-308.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems, 50*(1), 1-18.
Duncan, O. D. (1975). *Introduction to structural equation models*. New York, New York: Academic Press, Inc.
Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040.
Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*(2), 256.
Eliot, C. W., & Storer, F. H. (1876). *A manual of inorganic chemistry*. Chicago: Ivision, Blakeman, Taylor & Company.
Enneking, K. M., Breitenstein, G. R., Coleman, A. F., Reeves, J. H., Wang, Y., & Grove, N. P. (2019). The Evaluation of a Hybrid, General Chemistry Laboratory Curriculum: Impact on Students' Cognitive, Affective, and Psychomotor Learning. *Journal of Chemical Education, 96*(6), 1058-1067. doi:10.1021/acs.jchemed.8b00637
Ercan, O. (2014). Effect of 5E learning cycle and V diagram use in general chemistry laboratories on science teacher candidates 'attitudes, anxiety and achievement. *International Journal of Social Sciences and Education, 5*(1), 161-175.
Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1-4.
Farrokhi, F., & Mahmoudi-Hamidabad, A. (2012). Rethinking Convenience Sampling: Defining Quality Criteria. *Theory & Practice in Language Studies, 2*(4).
Fay, M. E., Grove, N. P., Towns, M. H., & Bretz, S. L. (2007). A rubric to characterize inquiry in the undergraduate chemistry laboratory. *Chemistry Education Research and Practice, 8*(2), 212-219.
Feisel, L. D., & Rosa, A. J. (2005). The role of the laboratory in undergraduate engineering education. *Journal of Engineering Education, 94*(1), 121-130.
Felder, R. M., Felder, G. N., & Dietz, E. J. (2002). The effects of personality type on engineering student performance and attitudes. *Journal of Engineering Education, 91*(1), 3-17.
Ferrell, B., & Barbera, J. (2015). Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry. *Chemistry Education Research and Practice, 16*(2), 318-337.
Ferrell, B., Phillips, M. M., & Barbera, J. (2016). Connecting achievement motivation to performance in general chemistry. *Chemistry Education Research and Practice, 17*(4), 1054-1066.
Finkelstein, N. D., Adams, W. K., Keller, C., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., . . . LeMaster, R. (2005). When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical Review Special Topics-Physics Education Research, 1*(1), 010103.
Flaherty, A., O'Dwyer, A., Mannix-McNamara, P., & Leahy, J. (2017). Evaluating the Impact of the "Teaching as a Chemistry Laboratory Graduate Teaching Assistant"

Program on Cognitive and Psychomotor Verbal Interactions in the Laboratory. *Journal of Chemical Education, 94*(12), 1831-1843.
Fraser, B. J., McRobbie, C. J., & Giddings, G. J. (1993). Development and cross-national validation of a laboratory classroom environment instrument for senior high school science. *Science Education, 77*(1), 1-24.
Furr, R. M. (2017). *Psychometrics: an introduction*. Thousand Oaks, CA: Sage Publications.
Galloway, K. R., & Bretz, S. L. (2015a). Development of an assessment tool to measure students' meaningful learning in the undergraduate chemistry laboratory. *Journal of Chemical Education, 92*(7), 1149-1158.
Galloway, K. R., & Bretz, S. L. (2015b). Measuring meaningful learning in the undergraduate chemistry laboratory: a national, cross-sectional study. *Journal of Chemical Education, 92*(12), 2006-2018.
Galloway, K. R., & Bretz, S. L. (2015c). Measuring meaningful learning in the undergraduate general chemistry and organic chemistry laboratories: a longitudinal study. *Journal of Chemical Education, 92*(12), 2019-2030.
Galloway, K. R., & Bretz, S. L. (2015d). Using cluster analysis to characterize meaningful learning in a first-year university chemistry laboratory course. *Chemistry Education Research and Practice, 16*(4), 879-892.
Georgiou, J., Dimitropoulos, K., & Manitsaris, A. (2007). A virtual reality laboratory for distance education in chemistry. *International Journal of Social Sciences, 2*(1), 34-41.
Grant, D. A. (1948). The latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin, 45*(5), 427-442.
Greenbowe, T. J., Poock, J. R., Burke, K., & Hand, B. M. (2007). Using the science writing heuristic in the general chemistry laboratory to improve students' academic performance. *Journal of Chemical Education, 84*(8), 1371-1379.
Groth, T., Falk, H., & Westgard, J. (1981). An interactive computer simulation program for the design of statistical control procedures in clinical chemistry. *Computer Programs in Biomedicine, 13*(1-2), 73-86.
Grove, N., & Bretz, S. L. (2007). CHEMX: An instrument to assess students' cognitive expectations for learning chemistry. *Journal of Chemical Education, 84*(9), 1524-1529.
Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282.
Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*: Advanced Analytics, LLC.
Hagay, G., & Baram-Tsabari, A. (2015). A strategy for incorporating students' interests into the high-school science classroom. *Journal of Research in Science Teaching, 52*(7), 949-978.
Hammersley, M. (1987). Some notes on the terms 'validity'and 'reliability'. *British Educational Research Journal, 13*(1), 73-82.
Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal

relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100*(1), 105-122.
Hawkes, S. J. (2004). Chemistry Is Not a Laboratory Science. *Journal of Chemical Education, 81*(9), 1257. doi:10.1021/ed081p1257
Hawkins, I., & Phelps, A. J. (2013). Virtual laboratory vs. traditional laboratory: Which is more effective for teaching electrochemistry? *Chemistry Education Research and Practice, 14*(4), 516-523. doi:10.1039/C3RP00070B
Hechter, R. P. (2013). Hockey, iPads, and projectile motion in a physics classroom. *The Physics Teacher, 51*(6), 346-347.
Heise, D. R. (1970). The semantic differential and attitude research. *Attitude Measurement, 4*, 235-253.
Hensen, C., & Barbera, J. (2019). Assessing Affective Differences Between A Virtual General Chemistry Experiment and a Similar Hands-On Experiment. *Journal of Chemical Education, 96*(10), 2097. doi:10.1021/acs.jchemed.9b00561
Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177-189.
Herranz, J., Carro, G., Contreras, A., Camacho, E. M., Garcia-Loro, F., & Gil, M. C. (2018). *From a Hands-on Chemistry Lab to a Remote Chemistry Lab: Challenges and Constrains.* Paper presented at the Online Engineering & Internet of Things, Columbia University.
Herrington, D. G., & Nakhleh, M. B. (2003). What defines effective chemistry laboratory instruction? Teaching assistant and student perspectives. *Journal of Chemical Education, 80*(10), 1197.
Herrmann, N. (1995). The creative brain. . *Canada: Quebecor Printing Book Group*.
Hofstein, A. (2004). The laboratory in chemistry education: Thirty years of experience with developments, implementation, and research. *Chemistry Education Research and Practice, 5*(3), 247-264.
Hofstein, A. (2017). The role of laboratory in science teaching and learning. In *Science Education* (pp. 357-368): Springer.
Hofstein, A., & Lunetta, V. N. (1982). The role of the laboratory in science teaching: Neglected aspects of research. *Review of Educational Research, 52*(2), 201-217.
Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education, 88*(1), 28-54.
Holmberg, R. G., & Bakshi, T. S. (1982). Laboratory work in distance education. *Distance Education, 3*(2), 198-206.
Hoole, D., & Sithambaresan, M. (2003). Analytical chemistry labs with kits and CD-based instructions as teaching aids for distance learning. *Journal of Chemical Education, 80*(11), 1308-1310.
Hou, H.-T., & Lin, Y.-C. (2017). *The Development and Evaluation of an Educational Game Integrated with Augmented Reality and Virtual Laboratory for Chemistry Experiment Learning.* Paper presented at the 2017 6th IIAI International Congress on Advanced Applied Informatics.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Hunt, W. C. (1987). *Teaching and Learning in the Affective Domain: A Review of the Literature*. Retrieved from Olympia, WA:

Irby, S. M., Borda, E. J., & Haupt, J. (2017). Effects of Implementing a Hybrid Wet Lab and Online Module Lab Curriculum into a General Chemistry Course: Impacts on Student Performance and Engagement with the Chemistry Triplet. *Journal of Chemical Education, 95*(2), 224-232.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98.

Jeschofnig, P. (2004). *Effective laboratory experiences for distance learning science courses with self-contained laboratory kits.* Paper presented at the Proceedings of the 20th Annual Conference on Distance Teaching & Learning.

Johnstone, A. H. (1982). Macro- and micro-chemistry. *School Science Review*(64), 377-379.

Jones, M. G., Howe, A., & Rua, M. J. (2000). Gender differences in students' experiences, interests, and attitudes toward science and scientists. *Science Education, 84*(2), 180-192.

Kelly, O. C., & Finlayson, O. E. (2007). Providing solutions through problem-based learning for the undergraduate 1st year chemistry laboratory. *Chemistry Education Research and Practice, 8*(3), 347-361.

Kendall, L. (1964). The effects of varying time limits on test validity. *Educational and Psychological Measurement, 24*(4), 789-800.

Kennepohl, D. (2007). Using home-laboratory kits to teach general chemistry. *Chemistry Education Research and Practice, 8*(3), 337-346.

Kennepohl, D., Baran, J., & Currie, R. (2004). Remote instrumentation for the teaching laboratory. *Journal of Chemical Education, 81*(12), 1814-1816.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486-507.

Kim, J.-O., & Mueller, C. W. (1978). *Factor Analysis: Statistical Methods and Practical Issues*. California: Sage.

Kim, J. S. (2005). The effects of a constructivist teaching approach on student academic achievement, self-concept, and learning strategies. *Asia Pacific Education Review, 6*(1), 7-19.

King, J. H. T., Wang, H., & Yezierski, E. J. (2018). Asymmetric Aldol Additions: A Guided-Inquiry Laboratory Activity on Catalysis. *Journal of Chemical Education, 95*(1), 158-163.

Kohn, M. (1951). Some important moments in the development of organic elementary analysis. *Analytica Chimica Acta, 5*, 337-344.

Komperda, R., Pentecost, T. C., & Barbera, J. (2018). Moving beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for

Quantitative Chemistry Education Research. *Journal of Chemical Education, 95*(9), 1477-1491.
Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). Taxonomy of educational objectives, handbook ii: Affective domain. New York: David McKay Company. *Inc. ISBN 0-679-30210-7, 0-582-32385, 1*.
Kurbanoglu, N. I., & Akim, A. (2010). The relationships between university students' chemistry laboratory anxiety, attitudes, and self-efficacy beliefs. *Australian Journal of Teacher Education, 35*(8), 48-59.
Lang, A. S., & Bradley, J.-C. (2009). Chemistry in Second Life. *Chemistry Central Journal, 3*(1), 14. doi:10.1186/1752-153x-3-14
LePine, J. A., LePine, M. A., & Jackson, C. L. (2004). Challenge and hindrance stress: relationships with exhaustion, motivation to learn, and learning performance. *Journal of Applied Psychology, 89*(5), 883-891.
Lewis, S. E. (2014). Examining evidence for external and consequential validity of the first term general chemistry exam from the ACS Examinations Institute. *Journal of Chemical Education, 91*(6), 793-799.
Lindberg, B. S. (2015). Proceedings of the Upsala Medical Society: How it all started 150 years ago. *Upsala Journal of Medical Sciences, 120*(2), 65-71.
LindenLab. (2003). The Second Life® virtual world.
Linn, M. C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science, 347*(6222), 1261757:1261751-1261756.
Liu, Y., Ferrell, B., Barbera, J., & Lewis, J. E. (2017). Development and evaluation of a chemistry-specific version of the academic motivation scale (AMS-Chemistry). *Chemistry Education Research and Practice, 18*(1), 191-213.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635-694.
Lucas, C. J. (1994). *American higher education: A history*: Springer.
Lyall, R., & Patti, A. T. F. (2010). Taking the Chemistry Experience Home—Home Experiments or "Kitchen Chemistry". In *Accessible Elements: Teaching Science Online and at a Distance* (pp. 83-108). Athabasca University: AU Press.
Ma, J., & Nickerson, J. V. (2006). Hands-on, simulated, and remote laboratories: A comparative literature review. *ACM Computing Surveys (CSUR), 38*(3), 1-24.
Mack, M. R., Hensen, C., & Barbera, J. Metrics and Methods Used To Compare Student Performance Data in Chemistry Education Research Articles. *Journal of Chemical Education, 96*(3), 401-413.
Mack, M. R., Hensen, C., & Barbera, J. (2019). Metrics and Methods Used To Compare Student Performance Data in Chemistry Education Research Articles. *Journal of Chemical Education, 96*(3), 401-413. doi:10.1021/acs.jchemed.8b00713
Marbach-Ad, G., Schaefer, K. L., Kumi, B. C., Friedman, L. A., Thompson, K. V., & Doyle, M. P. (2012). Development and evaluation of a prep course for chemistry graduate teaching assistants at a research university. *Journal of Chemical Education, 89*(7), 865-872.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391.

Martínez-Jiménez, P., Pontes-Pedrajas, A., Climent-Bellido, M., & Polo, J. (2003). Learning in chemistry with virtual laboratories. *Journal of Chemical Education, 80*(3), 346-352.

McCarthy, W. C., & Widanski, B. B. (2009). Assessment of Chemistry Anxiety in a Two-Year College. *Journal of Chemical Education, 86*(12), 1447-1449.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Mertler, C. A., & Reinhart, R. V. (2016). *Advanced and multivariate statistical methods: Practical application and interpretation*: Routledge.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

MHEducation. (2014). LearnSmart Labs. Retrieved from http://www.mhlearnsmart.com/flow/flowjs.html?isbn=0077818172&name=lms

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of psychological research, 3*(1), 111-130.

Miller, M., & Lu, M.-Y. (2003). Serving non-traditional students in e-learning environments: Building successful communities in the virtual campus. *Educational Media International, 40*(1-2), 163-169.

Miller, T. J., McGlashan, T. H., Rosen, J. L., Cadenhead, K., Ventura, J., McFarlane, W., . . . Woods, S. W. (2003). Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. *Schizophrenia Bulletin, 29*(4), 703-715.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100.

Mooring, S. R., Mitchell, C. E., & Burrows, N. L. (2016). Evaluation of a Flipped, Large-Enrollment Organic Chemistry Course on Student Attitude and Achievement. *Journal of Chemical Education, 93*(12), 1972-1983.

Moser, S., Zumbach, J., & Deibl, I. (2017). The effect of metacognitive training and prompting on learning success in simulation-based physics learning. *Science Education, 101*(6), 944-967.

Muenjohn, N., & Armstrong, A. (2008). Evaluating the structural validity of the multifactor leadership questionnaire (MLQ), capturing the leadership factors of transformational-transactional leadership. *Contemporary Management Research, 4*(1), 3-14.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research, 22*(3), 267-305.

Murray-Rust, P. (2008). Chemistry for everyone. *Nature, 451*(7179), 648-651.

Mutambuki, J., & Fynewever, H. (2012). Comparing chemistry faculty beliefs about grading with grading practices. *Journal of Chemical Education, 89*(3), 326-334.
Nakhleh, M. B. (1994). Student's models of matter in the context of acid-base chemistry. *Journal of Chemical Education, 71*(6), 495-499.
National Research Council. (2012). Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. Retrieved from https://doi.org/10.17226/13362
National Student Clearinghouse Research Center. (2017). *Current Term Enrollment Estimates*. Retrieved from https://nscresearchcenter.org
Nguyen, T. (2015). The effectiveness of online learning: Beyond no significant difference and future horizons. *Journal of Online Learning and Teaching, 11*(2), 309-319.
Nielsen, S. E., Scaffidi, J. P., & Yezierski, E. J. (2014). Detecting art forgeries: a problem-based Raman spectroscopy lab. *Journal of Chemical Education, 91*(3), 446-450.
Novak, A. (1963). Scientific inquiry in the laboratory. *The American Biology Teacher, 25*(5), 342-346.
Novak, J. D. (2002). Meaningful learning: The essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners. *Science Education, 86*(4), 548-571.
NRC. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*: National Academies Press.
O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: an extensive primer. *Psychological Bulletin, 97*(2), 316-333.
Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association, 69*(348), 894-908.
Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-Cultural Universals of Affective Meaning* (Vol. 1): University of Illinois Press.
Pavelich, M. J., & Abraham, M. R. (1979). An inquiry format laboratory program for general chemistry. *Journal of Chemical Education, 56*(2), 100-103.
Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C., & LeMaster, R. (2006). PhET: Interactive simulations for teaching and learning physics. *The physics teacher, 44*(1), 18-23.
Pickering, M. (1993). The teaching laboratory through history. *Journal of Chemical Education, 70*(9), 699-700.
Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research, 31*(6), 459-470.
Pintrich, P. R., & Garcia, T. (1991). *Student goal orientation and self-regulation in the college classroom* (Vol. 7). Greenwich, CT: JAI Press.
Pintrich, P. R., & Schrauben, B. (1992). *Students' motivational beliefs and their cognitive engagement in classroom academic tasks*.
Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of mathematical and statistical psychology, 46*(2), 273-286.

Pyatt, K., & Sims, R. (2012). Virtual and physical experimentation in inquiry-based science labs: Attitudes, performance and access. *Journal of Science Education and Technology, 21*(1), 133-147.

Reece, A. J., & Butler, M. B. (2017). Virtually the Same: A Comparison of STEM Students Content Knowledge, Course Performance, and Motivation to Learn in Virtual and Face-to-Face Introductory Biology Laboratories. *Journal of College Science Teaching, 46*(3), 83-89.

Reeves, J., & Kimbrough, D. (2004). Solving the laboratory dilemma in distance learning general chemistry. *Journal of Asynchronous Learning Networks, 8*(3), 47-51.

Reeves, J. H., & Exton, D. (2014). Developing the first online general chemistry laboratory exam. *Innovative Uses of Assessments for Teaching and Research, 1182*, 181-191.

Reeves, T. D., Marbach-Ad, G., Miller, K. R., Ridgway, J., Gardner, G. E., Schussler, E. E., & Wischusen, E. W. (2016). A conceptual framework for graduate teaching assistant professional development evaluation and research. *CBE—Life Sciences Education, 15*(2), 1-9.

Reid, N., & Shah, I. (2007). The role of laboratory work in university chemistry. *Chemistry Education Research and Practice, 8*(2), 172-185.

Ricci, R. W., & Ditzler, M. A. (1991). Discovery chemistry: A laboratory-centered approach to teaching general chemistry. *Journal of Chemical Education, 68*(3), 228-231.

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135-147.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software, 48*(2), 1-36.

Rotgans, J. I., & Schmidt, H. G. (2011). Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction, 21*(1), 58-67.

Rusek, M., Beneš, P., & Carroll, J. (2018). Unexpected Discovery: A Guided-Inquiry Experiment on the Reaction Kinetics of Zinc with Sulfuric Acid. *Journal of Chemical Education*, 1018-1021.

Rutherford, F. J. (1964). The role of inquiry in science teaching. *Journal of Research in Science Teaching, 2*(2), 80-84.

Sambell, K., Brown, S., & McDowell, L. (1997). "But Is It Fair?": An Exploratory Study of Student Perceptions of the Consequential Validity of Assessment. *Studies in Educational Evaluation, 23*(4), 349-371.

Scanlon, E., Colwell, C., Cooper, M., & Di Paolo, T. (2004). Remote experiments, re-versioning and re-thinking science learning. *Computers & Education, 43*(1-2), 153-163.

Schmidt-McCormack, J. A., Muniz, M. N., Keuter, E. C., Shaw, S. K., & Cole, R. S. (2017). Design and implementation of instructional videos for upper-division undergraduate laboratory courses. *Chemistry Education Research and Practice, 18*(4), 749-762.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323-338.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*(3-4), 207-231.

Schwab, J. J. (1958). The teaching of science as inquiry. *Bulletin of the Atomic Scientists, 14*(9), 374-379.

Schwartz, P., & Barbera, J. (2014). Evaluating the content and response process validity of data from the chemical concepts inventory. *Journal of Chemical Education, 91*(5), 630-640.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal, 8*(1), 289-317.

Scrucca, L., & Raftery, A. E. (2015). Improved initialisation of model-based clustering using Gaussian hierarchical partitions. *Advances in data analysis and classification, 9*(4), 447-460.

Seaman, J. E., Allen, I. E., & Seaman, J. (2018). Grade Increase: Tracking Distance Education in the United States. *Babson Survey Research Group.* Retrieved from http://www.onlinelearningsurvey.com/highered.html

Senese, F., & Bender, C. (2000). *The Internet chemistry set: web-based remote laboratories for distance education in chemistry.* Paper presented at the EdMedia: World Conference on Educational Media and Technology.

Simpson, E. (1971). Educational Objectives in the Psychomotor Domain. In *Behavioral Objectives in Curriculum Development: Selected Readings and Bibliography*. New Jersey: Educational Technology Publications.

Simpson, R. D. (1978). Relating Student Feelings to Achievement in Science. In *What Research Says to the Science Teacher* (Vol. 1, pp. 40-54). Washington. D.C.: ERIC Information Analysis Center for Science.

Slepchenko, B. M., Schaff, J. C., Macara, I., & Loew, L. M. (2003). Quantitative cell biology with the Virtual Cell. *Trends in Cell Biology, 13*(11), 570-576.

Snider, J. G., & Osgood, C. E. (1969). *Semantic differential technique; a sourcebook*: Aldine Pub. Co.

Sommer, R. (1931). The Liebig Laboratory and Liebig Museum in Giessen. *Journal of Chemical Education, 8*(2), 211-222.

Sonnenwald, D. H., Whitton, M. C., & Maglaughlin, K. L. (2003). Evaluating a scientific collaboratory: Results of a controlled experiment. *ACM Transactions on Computer-Human Interaction, 10*(2), 150-176.

Stegall, S. L., Grushow, A., Whitnell, R., & Hunnicutt, S. S. (2016). Evaluating the effectiveness of POGIL-PCL workshops. *Chemistry Education Research and Practice, 17*(2), 407-416.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173-180.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1-19.
Stephenson, N., & Sadler-McKnight, N. (2016). Developing critical thinking skills using the science writing heuristic in the chemistry laboratory. *Chemistry Education Research and Practice, 17*(1), 72-79.
Strauss, A. L. (1987). *Qualitative analysis for social scientists*: Cambridge University Press.
Suits, J. (2014). Lab Manual.
Tamir, P., & Lunetta, V. N. (1981). Inquiry-related tasks in high school science laboratory handbooks. *Science Education, 65*(5), 477-484.
Tatli, Z., & Ayas, A. (2010). Virtual laboratory applications in chemistry education. *Procedia-Social and behavioral sciences, 9*, 938-942.
Tatli, Z., & Ayas, A. (2013). Effect of a virtual chemistry laboratory on students' achievement. *Journal of Educational Technology & Society, 16*(1), 159-170.
Tro, N. J. (2004). Chemistry as general education. *Journal of Chemical Education, 81*(1), 54-57.
Tüysüz, C. (2010). The Effect of the Virtual Laboratory on Students' Achievement and Attitude in Chemistry. *International Online Journal of Educational Sciences, 2*(1), 37-53.
Ural, E. (2016). The effect of guided-inquiry laboratory experiments on science education students' chemistry laboratory attitudes, anxiety and achievement. *Journal of Education and Training Studies, 4*(4), 217-227.
Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis, 11*, 89-106.
Vishnumolakala, V. R., Southam, D. C., Treagust, D. F., Mocerino, M., & Qureshi, S. (2017). Students' attitudes, self-efficacy and experiences in a modified process-oriented guided inquiry learning undergraduate chemistry classroom. *Chemistry Education Research and Practice, 18*(2), 340-352.
Walker, J. P., Sampson, V., & Zimmerman, C. O. (2011). Argument-driven inquiry: An introduction to a new instructional model for use in undergraduate chemistry labs. *Journal of Chemical Education, 88*(8), 1048-1056.
Wartell, M. (1973). A new general chemistry laboratory scheme: Observation, deduction, reportage. *Journal of Chemical Education, 50*(5), 361-362.
Washbon, J. L. (2012). Learning and the new workplace: Impacts of technology change on postsecondary career and technical education. *New Directions for Community Colleges, 2012*(157), 43-52.
Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research, 19*(1), 231-240.
Wengraf, T. (2001). *Qualitative research interviewing: Biographic narrative and semi-structured methods*. Thousand Oaks, California: Sage.

Wheeler, L. B., Clark, C. P., & Grisham, C. M. (2017). Transforming a Traditional Laboratory to an Inquiry-Based Course: Importance of Training TAs when Redesigning a Curriculum. *Journal of Chemical Education, 94*(8), 1019-1026.

Whitney, T. M., Rodé, F., & Tung, C. C. (1972). The'Powerful Pocketful': an Electronic Calculator Challenges the Slide Rule. *Hewlett-Packard Journal, 23*(10), 2-9.

Williams, S. M., & Hmelo, C. E. (1998). Guest Editors' Introduction. *Journal of the Learning Sciences, 7*(3-4), 265-270. doi:10.1080/10508406.1998.9672055

Winkelmann, K., Keeney-Kennicutt, W., Fowler, D., & Macik, M. (2017). Development, Implementation, and Assessment of General Chemistry Lab Experiments Performed in the Virtual World of Second Life. *Journal of Chemical Education, 94*(7), 849-858.

Winkelmann, K., Scott, M., & Wong, D. (2014). A study of high school students' performance of a chemistry experiment within the virtual world of second life. *Journal of Chemical Education, 91*(9), 1432-1438.

Woodfield, B. F., Andrus, M. B., Andersen, T., Miller, J., Simmons, B., Stanger, R., . . . Allen, R. (2005). The virtual ChemLab project: A realistic and sophisticated simulation of organic synthesis and organic qualitative analysis. *Journal of Chemical Education, 82*(11), 1728-1735.

Woodfield, B. F., Catlin, H. R., Waddoups, G. L., Moore, M. S., Swan, R., Allen, R., & Bodily, G. (2004). The virtual ChemLab project: a realistic and sophisticated simulation of inorganic qualitative analysis. *Journal of Chemical Education, 81*(11), 1672-1678.

Woodrow, J. E. (1994). The development of computer-related attitudes of secondary students. *Journal of Educational Computing Research, 11*(4), 307-338.

Xu, X., & Lewis, J. E. (2011). Refinement of a chemistry attitude measure for college students. *Journal of Chemical Education, 88*(5), 561-568.

Xu, X., Villafane, S. M., & Lewis, J. E. (2013). College students' attitudes toward chemistry, conceptual knowledge and achievement: structural equation model analysis. *Chemistry Education Research and Practice, 14*(2), 188-200.

Zacharia, Z. C., & De Jong, T. (2014). The effects on students' conceptual understanding of electric circuits of introducing virtual manipulatives within a physical manipulatives-oriented curriculum. *Cognition and Instruction, 32*(2), 101-158.

Zacharia, Z. C., & Michael, M. (2016). Using physical and virtual manipulatives to improve primary school students' understanding of concepts of electric circuits. In *New Developments in Science and Technology Education* (pp. 125-140): Springer.

Zacharia, Z. C., Olympiou, G., & Papaevripidou, M. (2008). Effects of experimenting with physical and virtual manipulatives on students' conceptual understanding in heat and temperature. *Journal of Research in Science Teaching, 45*(9), 1021-1035.

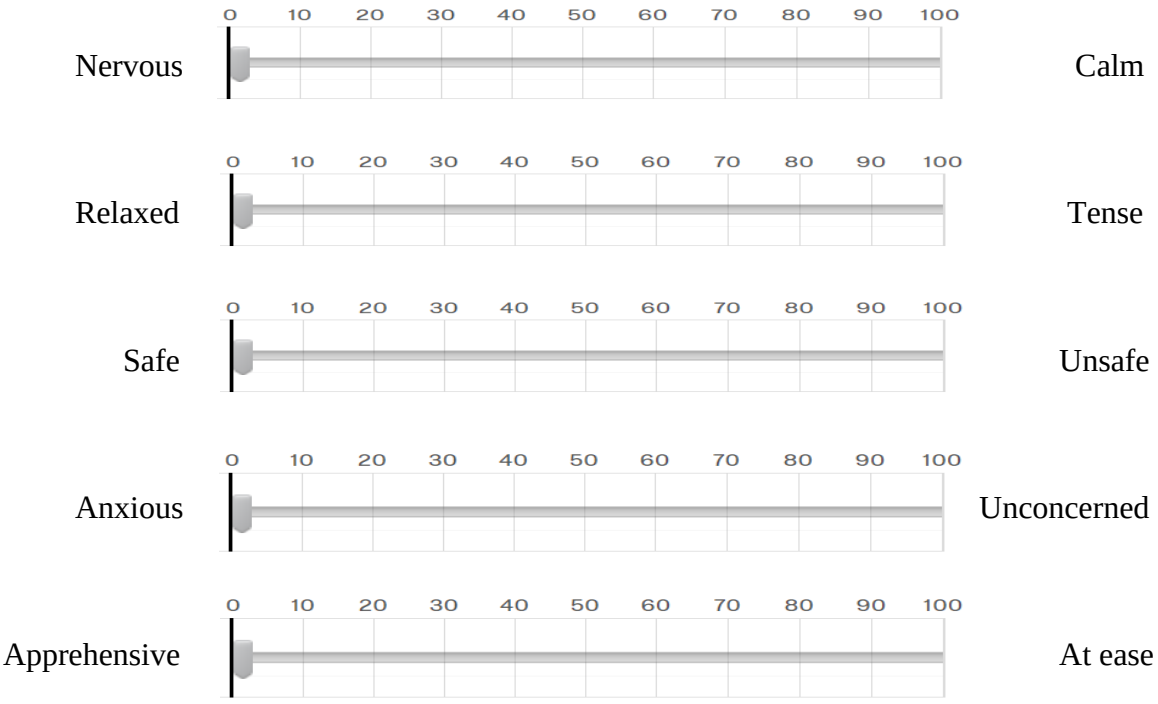
Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education, 25*(9), 1081-1094.

APPENDIX A: Supporting Information for Chapter IV

Survey items as administered:

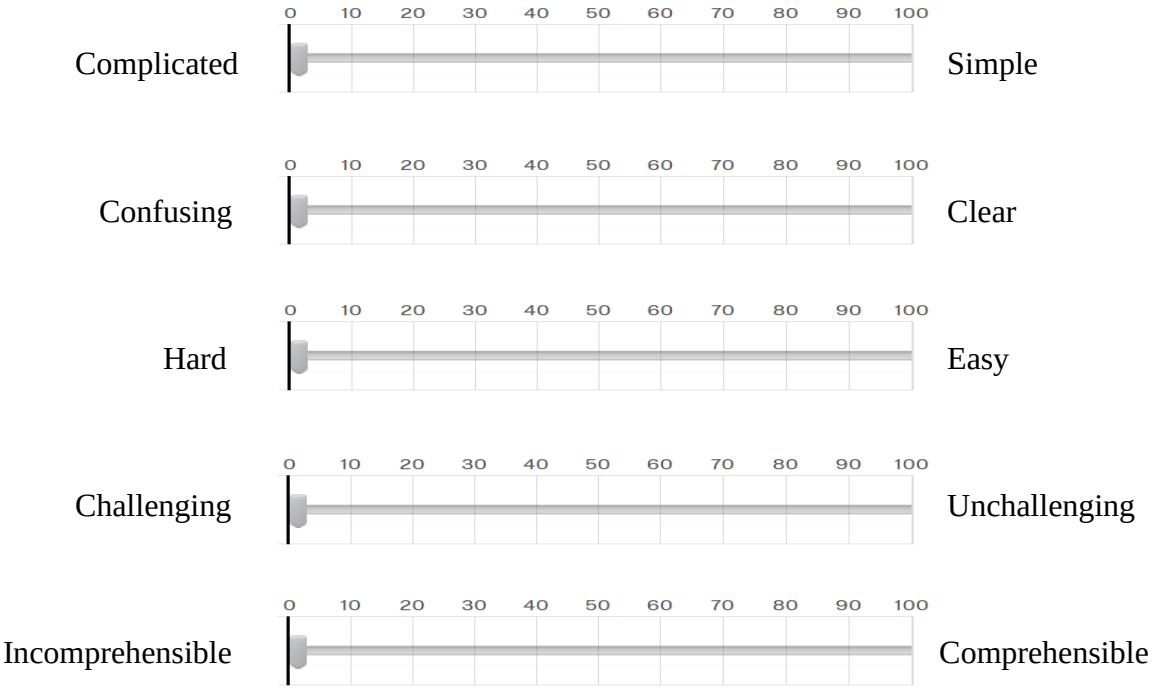
anxiety:

During this experiment I felt:



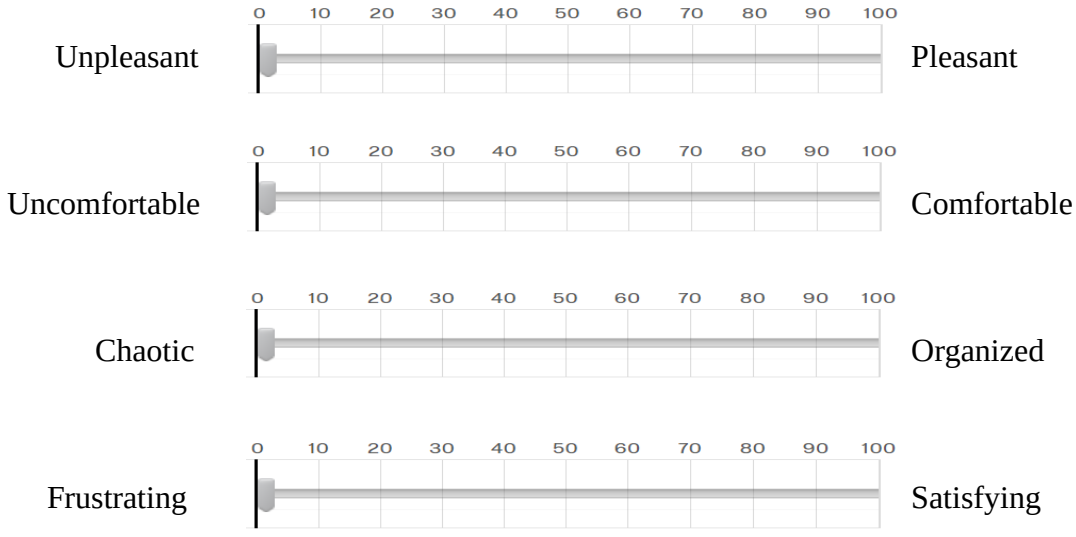
intellectual accessibility:

This experiment was:



emotional satisfaction

This experiment was:



initial interest-feeling

For the following question please indicate the level to which you agree with the statement.

I am fascinated by chemistry

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

I chose to take general chemistry because I'm really interested in the topic

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

I am really excited about taking this class

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

I am really looking forward to learning more about chemistry

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

initial interest-value

I think the field of chemistry is an important discipline

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

I think that what we will study in General Chemistry will be important for me to know

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

I think what we will study in General Chemistry will be worthwhile for me to know

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

usefulness of lab

This experiment allowed me to study complex problems

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

I liked performing this experiment

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

This experiment was interesting to me

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

I learned from the experiment

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

equipment usability

This experiment was easy to set-up and operate

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

This experiment worked well

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

open-endedness of lab

It was easy to explore in this experiment

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

This experiment was too long

a) Strongly agree
b) Somewhat agree
c) Neither agree nor disagree
d) Somewhat disagree
e) Strongly disagree

Table A.SI1: Initial MANOVA results.

	Type III Sum of Squares	*F*	*Sig.*	*Partial Eta Squared*
initial interest-feeling	0.460	0.532	0.466	0.001
initial interest-value	0.794	1.397	0.238	0.004
anxiety	847.545	1.401	0.237	0.004
intellectual accessibility	6636.629	10.666	0.001	0.027
emotional satisfaction	13751.823	19.931	0.000	0.049
usefulness of lab	9.158	11.239	0.001	0.028
equipment usability	20.577	22.870	0.000	0.056
open-endedness of lab	0.001	0.001	0.971	0.000

Table A.SI2: Average Scale Scores by Teaching Assistant.

TA	*AN*	*ES*	*IA*	*II-F*	*II-V*	*UoL*	*EU*	*OeoL*
1	37.32	72.08	61.49	3.81	4.24	3.75	4.15	3.60
2	35.98	64.52	58.00	3.91	4.38	3.92	4.19	3.52
3	32.56	70.04	62.69	3.65	4.21	3.71	4.07	3.32
4	27.05	75.62	68.71	3.62	4.38	3.71	4.19	3.46
5	32.14	77.99	73.04	3.97	4.40	3.89	4.31	3.54
6	32.20	73.12	72.98	3.67	4.29	3.72	4.38	3.81
7	26.89	70.33	66.54	3.97	4.30	4.34	4.48	4.11
8	30.81	65.11	63.83	3.62	4.21	3.06	3.68	3.52
9	32.63	65.26	57.05	3.79	4.21	3.59	4.04	3.72
10	39.30	64.41	56.78	3.66	4.19	3.76	3.76	3.24
11	40.30	56.79	55.92	3.79	4.27	3.41	3.65	3.23
12	37.55	56.14	57.95	3.58	4.26	3.47	3.80	3.62
13	41.28	49.71	52.96	3.79	4.41	3.09	3.26	3.30
14	36.73	56.15	53.51	3.45	4.02	3.16	3.39	3.48

AN: anxiety, ES: emotional satisfaction, IA: intellectual accessibility, II-F: initial interest-feeling, II-V: initial interest-value, U: usefulness of lab, EU: equipment usability, OE: open-endedness of lab

Note: Data from TAs 1-6 were grouped as 'Hands-On', data from TAs 7-10 were grouped together as 'Virtual Group A', and data from TAs 11-14 were grouped together as 'Virtual Group B'

APPENDIX B: Supporting Information for Chapter V

Interview Protocol:

1) I first wanted to just remind you of who I am and what my dissertation project is on. I am Cory Hensen and I am currently working on my Ph.D. under Jack Barbera. My dissertation project is looking at the efficacy of virtual laboratories. We are currently starting year 1 of the preliminary data collection before we move on to starting with students, we want to first understand where faculty are coming from through these interviews. Before I can begin looking at virtual laboratories, I first want to understand the learning objectives for the specific experiments I am interested in. Currently you are teaching (coordinating) Chem [course number] which covers the [experiment name] experiment in which I am interested.
    a. If you are okay with being interviewed, I would like to go over the informed consent [informed consent details].
    b. Thank you for signing that form. I am now going to turn on the audio recorder if you are okay with that.
2) I first want to start with asking how long you have been a faculty member at this institution?
3) How many of those years have you been involved in the general chemistry laboratory?
    a. In what capacity are you involved in the general chemistry laboratory?
4) Now I wanted to get into asking about a specific laboratory experiment. This term the students are doing an experiment over [topic]. Here is a copy of the procedure in case you need it. I wanted to ask you what learning objectives, or things you want your students to get out of this lab, you have?
    a. How many of these are assessed?
    b. If students missed today's experiment, what would they miss out on?

Table B.SI1: Demographics

		Beer's Law	*Calorimetry*	*Titration*
Total Enrollment (N)		630	484	355
Consented (N)	Hands-on	174	129	72
	Virtual	216	152	117
**Female (%)*	Hands-on	61.5	55.0	56.0
	Virtual	55.0	57.2	65.8
**White (%)*	Hands-on	57.5	49.6	44.0
	Virtual	49.1	57.2	49.6
**Biology Major (%)*	Hands-on	36.2	40.3	41.3
	Virtual	25.9	35.5	42.7

*****These categories were the majority for all experiments and sections for both the consented and overall course populations

Table B.SI2: List of overarching learning goals and experiment-specific learning objectives by faculty member

Faculty	**Overarching Goals**	**Beer's Law Objectives**	**Calorimetry Objectives**
	After completing this course, students will be able to do:	After doing this experiment, students will be able to:	
A	○ Graphical analysis	○ Understand and use the relationship between absorbance and concentration ○ Prepare solutions from both a stock solution and a solid ○ Calculate the molarity of a given solution	○ Experimentally determine and feel enthalpy changes ○ Use Hess's Law to predict the enthalpy change for a given reaction ○ Understand the relationship between energy and enthalpy at a constant pressure ○ Understand the relationship between energy and temperature

B	○ Error analysis ○ Measurement	○ Visualize concentration strength in a serial dilution ○ Derive graphically the relationship between absorbance and concentration ○ Use the relationship between absorbance and concentration to solve for an unknown concentration	○ Experimentally determine the thermal energy (q) for a given reaction ○ Use thermal energy to calculate the enthalpy change of a given reaction ○ Describe the relationship between a measured temperature change and an enthalpy change
C	○ Comparison with literature values ○ Unit analysis	○ Graphically determine the relationship between absorbance and concentration ○ Determine an unknown concentration using the relationship between absorbance and concentration ○ Successfully prepare a calibration curve ○ Prepare standard solutions from a stock solution	○ Experimentally determine the enthalpy of neutralization of phosphoric acid ○ Compare the experimental value with the literature value and determine percent error ○ Apply and understand the first law of thermodynamics
D	○ Graphing ○ Collaboration	○ Determine graphically the relationship between absorbance and concentration ○ Use the relationship to solve for an unknown concentration ○ Understand how light interacts with matter to produce the maximum wavelength ○ Understand real-world applications of spectroscopy	○ Experimentally determine the enthalpy of dissolution ○ Predict the sign of the change in enthalpy from a temperature change ○ Calculate heat energy by using a temperature change ○ Relate enthalpy changes to bond formation

Faculty	Overarching Goals	Beer's Law Objectives	Calorimetry Objectives
E	○ Graphing	○ Prepare calibration standard solutions ○ Understand the relationship between absorbance and percent transmittance ○ Understand the interaction of light and matter at the nano level ○ Use a calibration curve to determine an unknown concentration	○ Experimentally determine the change in enthalpy given a temperature change ○ Understand the relationship between mass and heat energy ○ Understand the difference between exothermic and endothermic reactions ○ Predict the sign of the change in enthalpy from a temperature change

The faculty members were not asked explicitly about any broad learning goals, however, some learning goals were still mentioned in the course of the interview. These were noted separately and were not included in any analysis as this study was focused on experiment-specific learning objectives.

Faculty	Titration Objectives
	After doing this experiment, students will be able to:
A	○ Successfully preform a titration ○ Identify key points on a titration curve ○ Use a pH titration curve to determine the concentration of a solution containing an acid ○ Identify the Brønsted-Lowry acids and bases present in solution and which of these substance(s) control the pH
B	○ Visually identify a change in pH during a titration ○ Use a titration curve to identify the molar mass and pKa of an unknown analyte
C	○ Identify key points on a titration curve ○ Determine the pKa and molar mass of an unknown anlatye using a titration curve ○ Visualize pH changes using a mixture of indicators
D	○ Determine the pKa and identify of an unknown acid using a titration curve ○ Predict the pH at the equivalence point ○ Identify key points on a titration curve ○ Predict which acid-base species are present at various points throughout a titration

E	○ Identify key points on a titration curve ○ Identify the unknown analyte using the calculated pKa value ○ Understand the reaction of a weak acid with a strong base ○ Understand real-world applications of titrations

Table B.SI3: Skew and Kurtosis values

		Hands-On		*Virtual*	
		Skewness	Kurtosis	Skewness	Kurtosis
Beer's Law	Anxiety	0.466	-0.648	0.329	-0.602
	Emotional Satisfaction	**-1.099**	**1.074**	-0.503	-0.688
	Intellectual Accessibility	-0.759	0.158	-0.251	-0.786
	Usefulness of Lab	-0.625	0.160	-0.536	-0.598
	Equipment Usability	**-1.277**	**2.077**	-0.764	-0.235
	Open-endedness of Lab	-0.488	0.163	-0.395	-0.127
Calorimetry	Anxiety	0.724	-0.657	0.903	-0.174
	Emotional Satisfaction	**-1.580**	**1.823**	**-1.399**	**1.217**
	Intellectual Accessibility	**-1.487**	**1.501**	**-1.571**	**1.796**
	Usefulness of Lab	-0.875	1.009	-0.622	-0.126
	Equipment Usability	**-1.009**	**0.329**	**-1.487**	**2.905**
	Open-endedness of Lab	-0.296	-0.744	-0.400	-0.553
Titration	Anxiety	0.798	-0.177	0.311	-0.765
	Emotional Satisfaction	-0.976	0.640	-0.548	-0.551
	Intellectual Accessibility	-0.802	0.436	-0.624	-0.411
	Usefulness of Lab	-0.335	-0.208	-0.459	-0.615
	Equipment Usability	**-1.404**	**2.539**	**-1.010**	0.232
	Open-endedness of Lab	0.153	0.019	-0.464	-0.169

Table B.SI4: Affective averages by environment and experiment

		Anx	**ES**	**IA**	**U**	**EU**	**OE**
Beer's Law	Hands-On	32.71	72.28	66.10	3.78	4.21	3.54
	Virtual	35.68	60.33	57.80	3.47	3.75	3.54
Calorimetry	Hands-On	23.56	78.12	77.32	3.88	4.66	4.07
	Virtual	21.72	75.83	78.56	3.62	4.41	3.95
Titration	Hands-On	32.08	69.10	69.58	3.73	4.29	3.23
	Virtual	33.12	63.50	68.25	3.37	3.76	3.50

Scales on a 0-100 semantic differential scale: Anx: anxiety, ES: emotional satisfaction, IA: intellectual accessibility

Scales on a 0-5 point Likert-type scale: U: usefulness of lab, EU: equipment usability, OE: open-endedness of lab

Latent Profile Analysis:

Once the clustering variables were selected as: emotional satisfaction, intellectual accessibility, usefulness of lab, open-endedness of lab, and equipment usability, the R package mclust was used to conduct a latent profile analysis. The anxiety scale was not selected as a clustering variable. A latent profile analysis has an advantage over traditional distance-based cluster analysis as it allows competing models to be compared with a fit index to determine the best clustering for the data. There are fourteen different types of models compared and each of these types had nine sub-models that were used to determine the number of profiles. There were four different categories that the models could be different on: the distribution of the data within each grouping, the volume of the grouping, the shape of the grouping, and the orientation of the grouping. The first letter of the model represents whether the volume was forced to be equal between the groupings (E) or if there was variation allowed in the volume (V). The second letter of the model indicates whether the shape of the model was forced to be equal between the groupings (E) or if there was variation allowed in the shape (V). The third letter of the model specifies whether the orientation of the model was on the coordinate axes (I), forced to be equal between groups (E), or allowed to vary (V). There are two models that do not follow this lettering. EII is for spherical groups with equal volume and equal shape and VII is for spherical groupings with variable volume and equal shape.

For the Beer's Law data, the r function mclustBIC was used to compare all the models on the BIC fit index:

Table B.SI5: BIC indices for all possible models

	EII	**VII**	**EEI**	**VEI**	**EVI**	**VVI**	**EEE**
1	-16501.4	-16501.4	-10475.4	-10475.4	-10475.4	-10475.4	-9655.02
2	-15153	-15076.6	-9880.17	-9755.95	-9877.55	-9755.02	-9620.04
3	-14574.2	-14422.3	-9798.39	-9594.37	-9737.39	NA	-9484.54
4	-14237.5	-14144.4	-9605.86	-9501.16	-9643.88	NA	-9490.44
5	-13978.2	-13905.1	-9603.75	-9497.37	-9645.41	NA	-9505.78
6	-13909.1	-13792.4	-9613.43	-9461.04	-9639.73	NA	-9551.32
7	-13765.9	-13675.8	-9568.55	-9493.52	-9686.25	NA	-9577.47
8	-13674.6	-13450	-9576.26	-9501.44	NA	NA	-9587.03
9	-13616.4	-13322.9	-9612.14	-9492.6	NA	NA	-9622.77
	EVE	**VEE**	**VVE**	**EEV**	**VEV**	**EVV**	**VVV**
1	-9655.02	-9655.02	-9655.02	-9655.02	-9655.02	-9655.02	-9655.02
2	-9547.8	-9458.01	-9468.32	-9571.09	-9484.68	-9572.07	-9489.55
3	-9573.11	-9461.59	-9430.9	-9582.6	-9485.44	-9611.83	-9520.89
4	-9540.49	**-9408.39**	-9445.8	-9586.7	-9484.16	-9625.05	-9560.75
5	-9562.34	-9408.89	-9433.96	-9641.8	-9551.22	-9689.47	-9551.73
6	-9587.87	-9435.61	-9467.29	-9661.99	-9576.45	-9775.14	-9626.42
7	-9649.66	-9453.16	-9511.17	-9720.75	-9631.47	-9855.15	-9721.86
8	-9663.83	-9484.48	-9547.29	-9807.52	-9711.39	-9853.02	-9788.98
9	-9672.88	-9492.94	-9567.88	-9840.68	-9729.3	-9987.25	NA

The best fitting model is the one that produces the highest BIC since BIC is calculated to be maximized in mclust. Therefore, the best fitting model was VEE with 4 profiles, as shown in bold in Table SI5. The grouping with five profiles had a similar fit but ultimately four was chosen as it was higher and is the simpler case. The more profiles that are selected, the harder it is to make meaningful comparisons between the profiles. This means that the groups were ellipsoidal with varying volume but equal shape and orientation. This process repeated in a similar fashion for the other two experiments. For the calorimetry experiment, the solution of five profiles had the highest BIC but after looking at the profiles, two profiles had very similar characteristics and were collapsed into one profile resulting in four profiles. For the titration experiment, the solution of three profiles had the highest BIC and was selected as the best fitting.

Table B.SI6: Affective averages by profile and experiment

		Anx	**ES**	**IA**	**U**	**EU**	**OE**
Beer's Law	Low	53.73	38.26	37.76	2.59	2.45	2.67
	Medium	31.45	71.98	66.93	3.70	4.16	3.54
	High	16.57	92.85	82.99	4.35	4.84	4.35
	Mixed	53.64	7.35	19.59	3.95	4.60	3.98
Calorimetry	Low	59.28	5.95	13.18	3.82	4.41	4.09
	Medium	31.54	57.48	66.10	3.15	4.03	3.36
	High	18.03	88.08	86.61	3.67	4.54	3.82
	Very High	8.57	99.38	98.38	4.57	5.00	5.00
	Very High	12.14	96.19	92.35	4.22	4.92	4.71
Titration	Low	54.98	26.21	46.75	2.69	2.23	2.73
	Medium	34.42	64.80	66.93	3.45	4.12	3.18
	High	17.90	90.36	84.98	4.10	4.70	4.18

Scales on a 0-100 semantic differential scale: Anx: anxiety, ES: emotional satisfaction, IA: intellectual accessibility

Scales on a 0-5 point Likert-type scale: U: usefulness of lab, EU: equipment usability, OE: open-endedness of lab

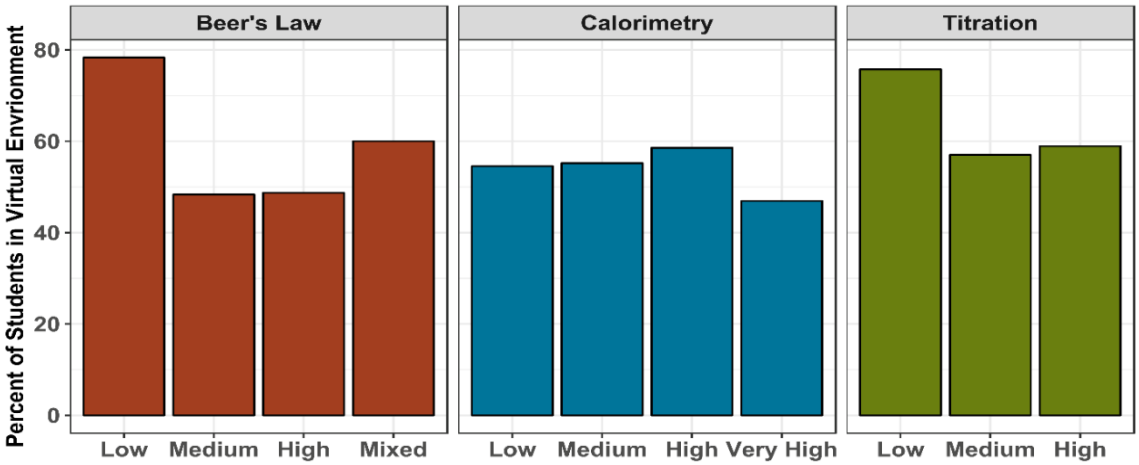


Figure SI1: Percent of students that completed the experiment in the virtual environment by profile