PERCEIVED USABILITY EVALUATION OF HANDS-ON AND VIRTUAL SCIENCE LABORATORIES: USING THE SYSTEM USABILITY SCALE (SUS) TO DETERMINE ADULT LEARNERS' PREFERRED AT-HOME LABORATORY EXPERIENCE

Amiee M. Wagner

A Dissertation

Submitted to Franklin University in partial fulfillment of
the requirements for the degree of

DOCTORATE OF EDUCATION IN ORGANIZATIONAL LEADERSHIP

July 2023

Committee:

Dr. Valerie A. Storey, Committee Chair

Dr. Bora Pajo, Committee Member

Dr. Patrick Bennett, Committee Member

Franklin University
This is to certify that the dissertation prepared by

Amiee Wagner

"Perceived Usability Evaluation of Hands-on and Virtual Science Laboratories: Using the System Usability Scale (SUS) to Determine Adult Learners' Preferred At-Home Laboratory Experience"

Has been approved by the committee as satisfactory completion of the dissertation requirements for the degree of

Doctor of Education in Organizational Leadership

Valerie A. Storey
Valerie A. Storey (Jul 27, 2023 17:25 EDT)

07/27/2023

Dr. Valerie A. Storey, Committee Chair and Adjunct Faculty
Franklin University

Patrick Bennett

07/31/2023

Dr. Patrick Bennett, Committee Member and VP, Academic Quality & Planning, Franklin University

Bora Pajo (Jul 31, 2023 10:22 EDT)

07/31/2023

Dr. Bora Pajo, Committee Member and Dir. of Dissertation Process, Franklin University

Rachel Althof
Rachel Althof (Jul 31, 2023 14:04 EDT)

07/31/2023

Dr. Rachel Althof, EdD Program Chair
Franklin University

Wendell Seaborne

07/31/2023

Dr. Wendell Seaborne, Dean, Doctoral Studies
Franklin University

FRANKLIN UNIVERSITY
Doctoral Studies

(THIS PAGE WAS INTENTIONALLY LEFT BLANK)

Abstract

Despite an extensive amount of research examining students' perceptions of alternative laboratory experiences, little focus has been aimed at the perceptions of adult learners. This two-phase mixed-methods study investigated the preferred at-home laboratory experience, and the factors that played a significant role in the user experiences, of adult learners in an online, undergraduate science course. A modified version of Brooke's (1996) System Usability Scale (SUS) was utilized in Phase I (quasi-experimental crossover phase); students self-selected between an at-home hands-on (AHHO) and virtual (VL) laboratory experience in Phase II. The SUS is a normed, validated, product-agnostic questionnaire that measures the quality of the user experience based on factors that support adult learners' needs (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy). Both phases showed no significant difference in preference for the AHHO or VL experience; and both the quantitative and qualitative strands showed that effectiveness, engagement, and ease of use were the most important factors for a good laboratory experience. The findings imply that the laboratory medium (AHHO or VL) is not as important to adult learners as the clarity of the laboratory instructions and the usability of the laboratory materials; and that a blend of AHHO and VL experiences would be the most beneficial to learners. Moreover, as the self-selection and crossover data arrived at the same conclusion, the modified SUS proved to be an effective instrument to measure the perceived usability, its sub-factors, and the quality of the user experience for laboratory experiments.

Keywords: online science laboratories, hands-on, virtual, student perceptions, perceived usability, system usability scale (SUS), adult learners

This dissertation is dedicated to my family. Your patience and understanding made this journey possible; your love and encouragement fueled the ride; and your faith and sacrifice helped me to cross the finish-line.

Acknowledgments

First and foremost, I give my thanks to God. It is through His love and grace that all things are possible. Thank you for making the way and for carrying me when I was weak.

I would like to thank my committee members for their support and encouragement throughout this journey. To Dr. Valerie A. Storey, thank you for not only modeling what it means to be a scholar-practitioner, but also serving as a guide and mentor throughout this doctoral program and the publication process. I am so blessed to graduate with two conference presentations and a journal publication with you. To Dr. Bora Pajo, thank you for your amazing (and timely!) feedback. This science geek definitely needed your help navigating the transition from the natural sciences to the social sciences. To Dr. Patrick Bennett, thank you for keeping me grounded in the leadership aspects of the project (and for all of your APA corrections!). I needed the reminder to not only look at the research problem through the lens of an educator, but also through the lens of a leader.

I would also like to thank the SCIE students for taking the time to complete the survey; the SCIE 211 instructors for facilitating the research study; and my colleagues for cheering me on. I am also grateful for the permissions from John Brooke and Rebecca Reck for the use of their instruments.

Special thanks to my tribe: Bridget, Andy, Latania, Ty-Juan, and Kasi. Meeting with you every week kept me sane (and on track) through this process. To Gina, I am so glad that we got to travel on this journey (and to Denver!) together. To the many others that shared this journey with me, I was surprised and delighted by the depth of the relationships we formed in this online dissertation program. Last but not least, to my family, thank you for gifting me the necessary time and space to complete this personal goal.

Table of Contents

List of Figures

List of Tables

Table Page

Chapter 1 – Introduction

Hands-on laboratory experiences have long been considered to be the cornerstone of science education (Clough, 2002; Deboer, 1991; Hofstein & Lunetta, 2003; Magin & Kanapathipallai, 2000). Balancing this with the rising demand for online courses (National Center for Educational Statistics [NCES], 2021d) continues to pose a significant challenge for science educators around the world (Bhute et al., 2021; Faulconer & Gruss, 2018; Raman et al., 2022; Reeves & Crippen, 2021). To meet their students' distance needs, instructors have developed a variety of alternative laboratory experiences (Bhute et al., 2021; Brinson, 2015; Faulconer & Gruss, 2018). While these alternative laboratory experiences have been extensively studied in traditional student populations (Attardi et al., 2016; Chen et al., 2014; Cossovich et al., 2020; Finne et al., 2022; Johnson & Barr, 2021; Klahr et al., 2007; Lindsay & Good, 2005; Nolen & Koretsky, 2018; Olympiou & Zacharia, 2011; Reece & Butler, 2017; Youngblood et al., 2022), there is a paucity of research on their efficacy within adult student populations.

This chapter will provide a brief overview of the background and purpose of the study, the research question and hypotheses, and its significance to the existing body of literature. The chapter will conclude with an extensive list of definitions for the key terms associated with alternative laboratory experiences, adult learning, and perceived usability.

Background of the Study

Alternative Laboratory Experiences

Research widely supports that hands-on laboratory activities play a critical role in the construction of scientific knowledge and development of practical skills (Clough, 2002; Deboer, 1991; Hofstein & Lunetta, 2003 Hofstein & Mamlok-Naaman, 2007; Magin & Kanapathipallai, 2000). However, science educators have increasingly incorporated alternative laboratory experiences into their courses to expand access to distance learners (Martinez et al., 2011;

Moosvi et al., 2019; Sindelar & Witkowski, 2021; Vogt et al., 2013) and to alleviate issues of safety (Tatli & Ayas, 2013), over-crowding (Reece & Butler, 2017), and the high costs to maintain and staff a traditional laboratory (Hofstein & Mamlok-Naaman, 2007; Ma & Nickerson, 2006; Reece & Butler, 2017). Some examples of the alternative laboratory approaches that have emerged since the early 2000s include simulations in which students manipulate virtual laboratory equipment and/or analyze simulated data (Chen et al., 2014; Klahr et al., 2007; Reece & Butler, 2017); remote labs in which students use an interface to manipulate real laboratory equipment from a distance (Corter et al., 2007; Meintzer et al., 2017; Tho & Yeung, 2018); and at-home, hands-on laboratories in which students utilize home supplies or materials provided in a kit (Al-Soufi et al., 2020; Cossovich et al., 2020; Orozco, 2017).

As hands-on laboratory activities have consistently been identified as a critical factor in both generating and sustaining student interest in STEM fields of study (Clough, 2002; VanMeter-Adams et al., 2014; Vennix et al., 2018; Young et al., 2016), factors such as learning outcomes achievement, student satisfaction, and student perception of effectiveness, engagement, and/or usability in the alternative laboratory formats have been extensively studied (Faulconer & Gruss, 2018; Heradio et al., 2016; Raman et al., 2022; Reeves & Crippen, 2021; Tsihouridis et al., 2018). While students largely reported they were able to learn key concepts in nearly all of the laboratory experiences, perceived engagement was consistently greater in modalities that incorporated the opportunity to interface with or physically manipulate laboratory equipment (Kelley, 2020; Meintzer et al., 2017; Stuckey-Mickell & Stucky-Danner, 2007).

However, the user experiences widely reported in the available literature (Achuthan et al., 2021; Barthet, 2021; Dalgarno et al., 2009; Erasmus et al., 2014; Paxinou et al., 2020; Rosen & Kelley, 2020; Sotelo et al., 2022) may not hold true for adult learners. Research has shown that learning activities that are considered to be highly effective in traditional student populations may not be as effective (or even ineffective) with learners that are 25 years or older (Arghode et al., 2017; Chen, 2014; Knowles, 1984; Lewis & Bryan, 2021; Remenick & Goralnik, 2019; Taylor & Hamdy, 2013). This implies that these student sub-populations have different characteristics, learning needs, and expectations.

Andragogy

Malcolm Knowles (1980, 1984) defined andragogy as the "the art and science of helping adults learn" (Knowles, 1980, p. 43). He observed that adult learners tend to be more independent, self-directed, and self-motivated than traditional students (Knowles, 1980, 1984). Due to the challenges of a full-time work schedule and extensive family obligations, adult learners have a limited amount of time to dedicate to their studies (Bowers & Bergman, 2016; Osam et al., 2016). As such, they value relevant, active, real-world learning activities that are directly tied to learning objectives (Arghode et al., 2017; Chen, 2014; Knowles, 1980; Loeng, 2018; Pratt, 1993; Taylor & Hamdy, 2013). In other words, they prefer activities that are not perceived as busy work or a waste of their valuable time (Gutruf et al., 2021; Johnson & Barr, 2021; Moosvi et al., 2019; Osam et al., 2016). To keep learners engaged (and to avoid wasting their time), it is important to measure the quality of the students' experiences.

Perceived Usability

Perceived usability measures the quality of a product by the quality of the user experience (Iwarsson & Ståhl, 2003; Quesenbery, 2001; Shackel, 2009). In this framework, the user experience is measured by variables that are consistent with the attributes of adult learners (Arghode et al., 2017; Chen, 2014; Knowles, 1980, 1984) and with the variables that have been measured in the available literature. The variables measured are Quesenbery's (2001) 5E's of perceived usability (effectiveness, efficiency, engagement, error tolerance, and ease of use) and Bandura's (1997) self-efficacy.

Furthermore, perceived usability can be quickly and effectively measured using Brooke's (1996) System Usability Scale (SUS). The SUS is one of the most frequently used instruments in usability studies (Bevan, 1995; Brooke, 2013; Lacerda & von Wangenheim, 2018; Lewis, 2018a;). The strengths of the SUS, include: (a) it is simple to use and can be completed in just a few minutes; (b) it has been standardized, validated, normalized, and shown to be very reliable; (c) minor modifications to the questions does not affect validity or reliability of the scale; and, (d) the results are reliable even with sample sizes as small as 8 (Bangor et al., 2008; Bevan, 1995; Brooke, 1996; Brooke, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016; Shackel, 2009). Moreover, the SUS has proven to be a particularly effective instrument when comparing the user experiences for two versions of the same 'application' (Brooke, 2013).

Statement of the Problem

While andragogy emphasizes learning through hands-on activities (Lewis & Bryan, 2021), there is a paucity of research on adult learners' preferred laboratory experience. Given that adults over the age of 25 accounts for 25% of total student enrollment in higher education in the United States (NCES, 2021b) and that the number of students taking distance education classes has nearly tripled since 2019 (NCES, 2021a), this represents a significant gap in the literature.

Purpose of the Study

The purpose of this study was to determine adult learners preferred at-home laboratory experience in an online, undergraduate science course for non-majors in order to make informed instructional design decisions that are both student-centered and cost-effective. A modified version of Brooke's (1996) System Usability Scale (SUS) was used to measure the perceived usability, effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy of at-home, hands-on (AHHO) and at-home, virtual laboratory (VL) experiments. A comparison of mean perception data was used to determine students' preferred laboratory experience.

The SUS is a product-agnostic questionnaire (Bevan, 1995; Brooke, 2013; Lewis, 2018a) that has been used to measure the perceived usability of a wide range of products (Lacerda & von Wangenheim, 2018; Lewis, 2018b; Lewis & Sauro, 2019). Although this tool has been used to evaluate students' perceptions of various educational technologies (Granić & Ćukušić, 2011; Harrati et al., 2016; Orfanou et al., 2015; Pal & Vanijja, 2020; Vlachogianni & Tselios, 2021, 2022), to the researcher's knowledge, the SUS has never been used to evaluate user experiences for laboratory experiments. As such, the secondary and tertiary purposes of this study were to investigate the applicability of the SUS in the context of perceived usability evaluation of laboratory experiments and to determine which (if any) of Quesenbery's (2001) 5E's of perceived usability (effectiveness, efficiency, engagement, error tolerance, and/or ease of use) and Bandura's (1997) self-efficacy played a significant role in adult learners' user experiences.

Investigating the applicability of the SUS supports the primary purpose of this study. Science educators have conducted comparison studies for the last two decades (Bhute et al., 2021; Raman et al., 2022; Reeves & Crippen 2021; Wei et al., 2019); however, there are still no clear answers as to which laboratory experience is the most effective, most engaging, or preferred by learners (Tsihouridis et al., 2018). Further complicating matters, nearly all of these studies utilized self-developed questionnaires/surveys/interview guides, home-grown assignments, or final grades to evaluate the laboratory experiences for their specific students at their specific institutions (Faulconer & Gruss, 2018). As such, it is nearly impossible to generalize these findings, even to the same course at another institution (Gopalan et al., 2020). This highlighted the need for a normed and validated "product-agnostic" instrument that can be used to evaluate student perception data across a myriad of laboratory experiences; a tool that can be used by science program chairs to better manage their curriculum and costs for laboratory resources.

Research Questions and Hypotheses

Adult learning theories emphasize the importance of active, hands-on learning for this student population (Chen, 2014; Greene & Larsen, 2018; Lewis & Bryan, 2021; Knowles, 1980, 1984). However, adult learners' perceptions are largely missing from the available literature. To ascertain if adult learners preferred at-home, hands-on laboratory activities or at-home, virtual laboratory activities (and why), the overarching mixed-methods research questions were:

(i) Will adult learners in an online, undergraduate, science course for non-majors at a private, professionally-focused university in the Midwest prefer at-home, hands-on laboratory experiences or at-home, virtual laboratory experiences?

(ii) Which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) play a significant role in participants' user experiences?

To provide context for the research questions and hypotheses, this study occurred in two phases. In Phase I (quasi-experimental crossover phase), students completed the SUS and open-ended survey questions that were modified from Reck et al. (2019) after completing an AHHO laboratory experience and after completing a VL experience. In Phase II (self-selection phase), students were given the opportunity to self-select between the AHHO and VL experiences for their last laboratory assignment; participants then completed a survey in which they identified which laboratory experience they selected and why.

Quantitative Research Questions and Hypotheses

RQ1: Will adult learners in an online, undergraduate, science course for non-majors at a private, professionally-focused university in the Midwestern United States prefer at-home, hands-on laboratory experiences or at-home, virtual laboratory experiences?

Preference Based Upon Perceived Usability Data: Two-sample t-test ($\alpha=0.05$)

$H_{01\text{-}1}$: The difference in mean perceived usability will not be statistically significant for this population of adult learners ($\mu_{AHHO} - \mu_{VL}=0$).

$H_{1\text{-}1}$: The difference in mean perceived usability will be statistically significant for this population of adult learners ($\mu_{AHHO} - \mu_{VL}\neq 0$).

Preference Based Upon Sub-factors: Analysis of Variance (ANOVA) ($\alpha=0.05$)

$H_{01\text{-}1'}$: The difference in means ($\mu_{AHHO} - \mu_{VL}$) for each of the sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) will be equal.

$H_{1\text{-}1'}$: At least one of the differences in means ($\mu_{AHHO} - \mu_{VL}$) will be different.

Preference Based Upon Self-Selection Data: Two-proportion Z-test ($\alpha=0.05$)

$H_{01\text{-}2}$: The proportion of participants that self-select the AHHO experience will be the same as the proportion that self-select the VL experience ($\hat{p}_{AHHO} = \hat{p}_{VL}$)

$H_{1\text{-}2}$: The proportion of participants that self-select the AHHO experience will not be the same as the proportion that self-select the VL laboratory experience ($\hat{p}_{AHHO} \neq \hat{p}_{VL}$)

RQ2: Is there evidence to support that one or more of the factors (effectiveness, efficiency, engagement, ease of use, error tolerance, self-efficacy) plays a significant role in this population of students' user experience?

Multiple Regression Analysis ($\alpha=0.05$) – AHHO

H_{02a}: None of the factors ($\beta_i = 0$) play a significant role in this population of students' user-experience for the AHHO laboratory experience.

H_{2a}: At least one of the factors ($\beta_i \neq 0$) plays a significant role in this population of students' user-experience for the AHHO laboratory experience.

Multiple Regression Analysis ($\alpha=0.05$) – VL

H_{02b}: None of the factors ($\beta_i = 0$) play a significant role in this population of students' user-experience for the VL experience.

H_{2b}: At least one of the factors ($\beta_i \neq 0$) plays a significant role in this population of students' user-experience for the VL experience.

Qualitative Research Questions

RQ3: Which at-home laboratory experience do participants say they prefer (hands-on or virtual)? Why do they say they prefer this laboratory experience?

RQ4: Which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, self-efficacy, or other) played an important role in students' user experiences?

RQ5: Do the same factors identified by user experience responses play a role in the students' self-selection process?

Significance of the Study

Although the enrollment for adult learners has outpaced that of traditional-aged students for the last decade (Chen, 2014; NCES, 2021b), most courses are still designed for traditional students (Chen, 2014; Galustyan, et al., 2019; Greene & Larsen, 2018; Lewis & Bryan, 2021; Remenick & Goralnik, 2019). As students' perceptions affect how and what they learn (Greene & Larsen, 2018; Ramsden, 1979; Tinto, 2017), this can have a negative impact on adult learners' laboratory experiences (Arghode et al., 2017; Chen, 2014). However, adult learners' perceptions of alternative laboratory experiences are largely missing from the existing body of literature.

This study utilized a modified version of Brooke's (1996) System Usability Scale (SUS) as a framework to evaluate adult learners' perceptions of and preferences for AHHO and VL experiences. Ultimately, the results of this study can be used to add adult learners' perceptions of at-home laboratory experiences to the existing body of literature and to make cost-effective, data-driven decisions for course design.

Laboratories are expensive, whether they are performed at home or in a traditional campus laboratory. For at-home laboratories, the cost (on average) for a standard laboratory kit is \$150-\$300 per student (Orozco, 2017); the cost for a good quality simulator is \$50-\$100 per simulation per student (Senapati, 2022). While it is important to control costs for the University and for students (especially for adult learners), the quality of the laboratory experience should not be limited by the lack of home supplies or by the poor quality of the free simulators. Therefore, understanding adult learners preferred at-home laboratory experiences can help lead faculty and program chairs at primarily online universities better manage their curriculum and the costs for laboratory resources.

Assumptions

Previous studies largely compared the traditional, hands-on laboratory experience, which was performed in an actual science laboratory, with an alternative laboratory experience, that was performed at home, alone, with limited access to instructors and peers (Achuthan et al., 2021; Chen et al., 2014; Finne et al., 2022; Kapici et al., 2019; Klahr et al., 2007; Koretsky et al., 2011; Moosvi et al., 2019; Olympiou & Zacharia, 2011; Rosen & Kelley, 2020; Schultz et al., 2020; Stokes & Silverthorn, 2021). While previous studies revealed that students tended to prefer hands-on laboratory experiences (Johnson & Barr, 2021; Kelley, 2020; Meintzer et al., 2017; Stuckey-Mickell & Study-Danner, 2007), it is difficult to discern if students preferred the ability to manipulate objects physically or if they preferred the 'laboratory setting' with easy access to an instructor and the social aspects of working with peers.

This study assumed the at-home laboratory settings were the same for both the hands-on and virtual experiments. However, the study's design cannot guarantee that was the case. The laboratories were completed in any space the student chose for their "at-home" location. This is a fair assumption given that students in an online science class are not likely to have access to a brick-and-mortar science laboratory where they perform the experiment with lab partners under the direct supervision of an instructor.

Furthermore, the laboratory materials (background materials, procedures, goals/objectives, and pre-/post-lab questions) were identical for each set of laboratories in the crossover phase of this study. The only difference between the laboratories was that the hands-on experiment utilized physical manipulatives while the virtual experiment utilized virtual manipulatives. As such, it was assumed that student perception data and self-selection data reflected their perceptions of and preference for manipulating physical equipment versus virtual equipment. Again, the study's design cannot guarantee this was the case. Students' lack of access to the physical equipment necessary to complete the hands-on experiment or technical issues with the simulators for the virtual experiment may have impacted their perceptions of the experience. However, this was a fair assumption given that the only physical equipment needed are coins, ruler, and a balloon (common household items) and that, per university policy, the students registered in the online course should have the technical capabilities to operate the simple simulators.

Definitions of Key Terms

For this study, the following terms are defined as:

Adult learner. An adult learner is defined as a non-traditional student that is aged 25 or older (Bowers and Bergman, 2016; Lewis & Bryan, 2021; Osam et al., 2016).

Alternative laboratory (AL). An alternative (also known as a non-traditional, online, or distance) laboratory experience is one in which all or part of the traditional, hands-on experience is replaced with a virtual, remote, or at-home experiences (Faulconer & Gruss, 2018).

Ease of use (EU). According to Quesenbery (2001), ease of use is the extent to which specified users are able to use the product without intervention (i.e. ability to use the product autonomously).

Effectiveness (E): Effectiveness is defined as the extent to which a product can be used by specified users to achieve specified goals (ISO 9241-11:2018 section 3.1.12).

Efficiency (F). Efficiency is defined as the extent to which specified users can complete assigned tasks within the allotted timeframe (ISO 9241-11:2018 section 3.1.13).

Engagement (N). According to Quesenbery (2001), engagement is the extent to which the product was pleasant and satisfying to use (i.e. specified users enjoyed the experience).

Error tolerance (ET). According to Quesenbery (2001), error tolerance is the extent to which the product helps users recover from errors.

Non-traditional student. A non-traditional student is defined as one who exhibits one or more of the following characteristics: delayed enrollment in post-secondary education by more than a year after high school graduation, completed high school with a GED, attends part-time, works full-time, is financially independent, has dependents other than a spouse, is a single parent, and is age 25 or older (NCES, n.d.; Osam et al., 2016).

Satisfaction. Satisfaction is defined as the extent to which a product meets the user's needs and expectations (ISO 9241-11:2018 section 3.1.14).

Self-efficacy (SE). Bandura (1997) defined self-efficacy as one's belief in their ability to perform a specific task or achieve a specific goal.

Traditional laboratory (TL). A traditional science laboratory experience is defined as one that is performed in a fully-equipped, brick-and-mortar campus laboratory; laboratory experiments are performed under the supervision of an instructor and/or teaching assistants; and students engage in hands-on experiences with physical equipment, instruments, and/or chemical reagents (Clough, 2002; Hofstein & Mamlok-Naaman, 2007; Seery, 2020).

Traditional student. A traditional student is defined as one who enrolled in post-secondary education within a year of high school graduation, has a high school diploma, is registered full-time, works part-time (if at all), is dependent on parents for financial support, and is less than 25 years old (NCES, n.d.; Osam et al., 2016).

Usability: The International Organization for Standardization (2018) defined usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 9241-11:2018 section 3.1.1).

Organization of this Dissertation

This chapter provided an overview of the background and purpose of the study, the research question and hypotheses, its significance to the existing body of literature, and an extensive list of definitions for the key terms associated with alternative laboratory experiences, adult learning, and perceived usability.

The remaining chapters will be organized as follows: Chapter 2 will include a systematic review of the available literature published on the efficacy of alternative laboratory experiences; Chapter 3 will describe the mixed-methods design and instruments used to collect data; Chapter 4 will report the results of the data analysis; and, Chapter 5 will provide a summary of the study’s findings (and their limitations), recommendations for future research, and implications for practice.

Chapter 2 – Literature Review

Enrollment in distance education classes has been on the rise for the last two decades (NCES, 2021d). For science educators, the increased demand for online courses has dramatically changed where and how science laboratories are performed (Raman et al., 2022; Reeves & Crippen, 2021). Examples of the alternative laboratory experiences designed for distance learners, include: virtual labs (VL) in which students engage with immersive simulations (Avci, 2022; Crandall et al., 2015; Gnesdilow & Putambekar, 2022; Reece & Butler, 2017), virtual manipulatives (Attardi & Rogers, 2015; Barbeau et al., 2013; Klahr et al., 2007; Zacharia et al., 2008), live-streamed or pre-recorded video demonstrations (Barbeau et al, 2013; Chen, 2022; Finne et al., 2022; Schultz et al., 2020), and/or the analysis of simulated or pre-collected datasets (Hsu & Rowland-Goldsmith, 2020; Johnson & Barr, 2021; Vogt et al., 2013); remote labs (RL) in which students use a computer interface to manipulate real laboratory equipment from a distance (Achuthan et al., 2021; Meintzer et al., 2017; Tho & Yeung, 2018) or to receive real-time data from a data-pack or equipment from remote location (Childers & Jones, 2017; Erasmus et al., 2014; Stokes & Silverthorn, 2021); at-home, hands-on laboratories (AHHO) in which students utilize home supplies and/or materials provided in a kit (Brewer at al., 2013; Moosvi et al., 2019; Nguyen & Keuseman, 2020; Reuter, 2009; Youngblood et al., 2022); and blended labs (B) in which hands-on activities are supplemented with virtual or remote elements (Bortnik et al, 2017; Chang et al, 2022; Davies, 2019; Paxinou et al., 2020). Table 2.1 provides a few examples from the literature for each type of alternative laboratory experience (VL, RL, AHHO, and B); each is organized by the decade it was published.

Table 2.1

Examples of Alternative Laboratory Experiences by Decade of Publication

Alternative Laboratory Experiences	2000s	2010s	2020s
Virtual Labs (VL)			
Simulations	Lindsay & Good (2005); Corter et al. (2007); Stuckey-Mickell & Stuckey-Danner (2007)	Koretsky et al. (2011); Chen (2014); Reece & Butler (2017); Gunawan et al. (2019)	Klein et al. (2021); Avci (2022); Desa et al. (2022); Gnesdilow & Putambekar (2022)
Virtual manipulatives	Klahr et al. (2007); Zacharia et al. (2008)	****	****
Live-streamed/pre-recorded video demonstrations	****	Barbeau et al. (2013); Attardi et al. (2016)	Johnson & Barr (2021); Chen (2022)
Analysis of pre-collected datasets and images	****	Vogt et al. (2013)	Hsu & Rowland-Goldsmith (2020); Klein et al. (2021); Finne (2022)
Remote Labs (RL)			
Manipulated/received real-time data from remote equipment	Lindsay & Good (2005); Corter et al. (2007)	Lowe et al. (2013); Erasmus et al. (2014); Meintzer et al. (2017); Zidny et al. (2019)	Achuthan et al. (2021)
At-Home, Hands-On Labs (AHHO)			
Kits	Reuter (2009)	Mawn et al. (2011); Brewer et al. (2013); Orozco (2017);	Rosen & Kelley (2020); Barthet (2021); Rayment et al. (2022)
Home supplies	****	Vogt et al. (2013); Moosvi et al. (2019)	Al-Soufi et al. (2020); DeChenne-Peters et al. (2022)
Blended Laboratories (B)			
Traditional lab blended with VL	Jara et al. (2009); Toth et al. (2009)	Saitta et al. (2011); Bortnik et al. (2017); Davies (2019)	Paxinou et al. (2020); Heng et al. (2022); Chang (2022)

Note. **** = there were no examples from the pool of studies selected for this review

While acceptance of online learning has significantly increased in the last twenty years (Shachar & Neumann, 2010; Tsihouridis et al., 2018), the same cannot be said for these nontraditional laboratory experiences (Raman et al., 2022; Reeves & Crippen, 2021; Spencer, 2021). However, science educators' concerns about moving laboratory experiences out of a fully-equipped campus laboratory are justified (Bhute et al., 2021; Wei et al., 2019). Research widely supports that hands-on laboratory activities play a critical role in the construction of scientific knowledge (Clough, 2002; Hofstein & Mamlok-Naaman, 2007; Pheeney, 1997), generate and sustain interest in STEM fields (VanMeter-Adams et al., 2014; Vennix et al., 2018; Young et al., 2016), and aid in practical skills development (Davies, 2019; de Jong et al., 2013; Heng et al., 2022; Paxinou et al., 2020).

As the laboratory experience is often considered to be the cornerstone of STEM education (Clough, 2002), it should not be surprising that the efficacy of alternative laboratory experiences has been extensively studied. Since the early 2000s, there have been more than 20,000 publications (journal articles, editorials, books, and conference proceedings) in which science educators from nearly all STEM disciplines (Biology, Chemistry, Physics, and Engineering), from nearly all educational levels (K-12, undergraduate, and graduate), and from more than 25 countries have grappled with slightly different variations of the same two questions: Are non-traditional laboratory experiences as effective and engaging as the traditional laboratory experience? Which laboratory experience do students prefer?

This chapter provides a systematic review of the available literature published on the efficacy of alternative laboratory experiences. The sections that follow will outline the search strategy utilized to investigate the available literature, provide a general overview of the selected studies, and discuss the themes that emerged from their analysis.

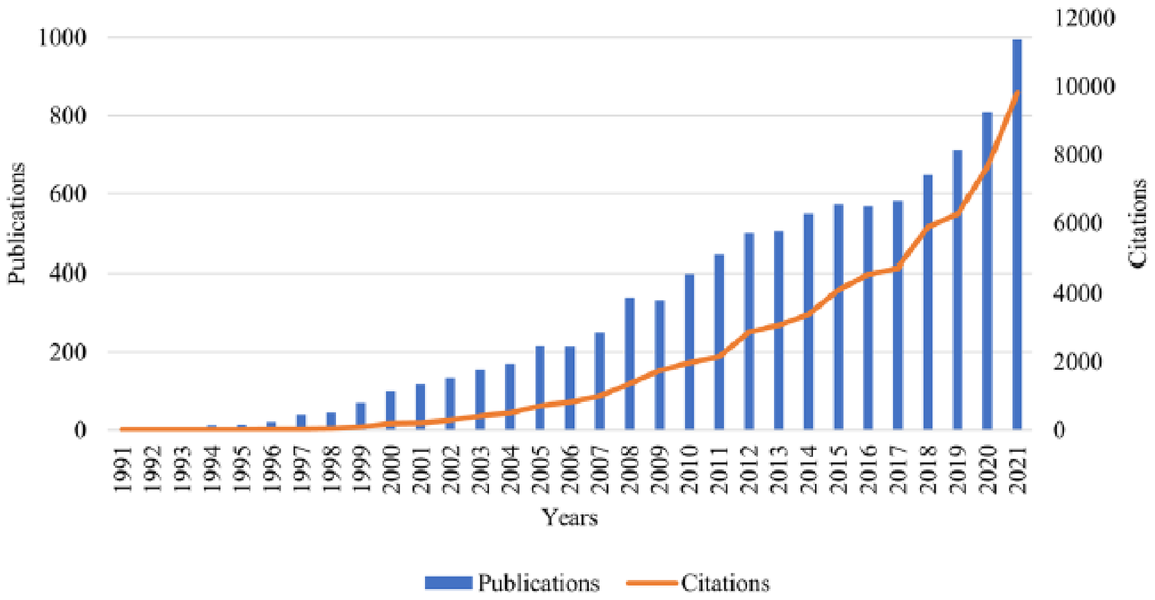
Search Strategy

Timeframe

This review focused on peer-reviewed journal articles published between 2002 and 2022. This timeframe was selected to explore the evolution of alternative laboratory experiences over the last two decades and because very little research was performed/cited on the use of non-traditional laboratory experiences prior to 2002 (Raman et al., 2022; Figure 2.1).

Figure 2.1

Number of Papers Published per Year on Alternative Laboratory Experiences: 1991-2021



Note. The bar graph illustrates the number of publications on alternative laboratory experiences per year (number of publications on the left); the line illustrates the number of times this work was cited per year (number of citations on the right). From “Virtual Laboratories – A Historical Review and Bibliometric Analysis of the Past Three Decades” by R. Raman, K. Achuthan, V. Nair, and P. Nedungadi, 2022, *Education and Information Technologies*, 27(8), p. 11062. https://doi.org/10.1007/s10639-022-11058-9

Keyword Searches

The studies included in this review (n=95) were identified through keyword searches of the EBSCO databases. A Boolean search strategy using the key terms “science laboratories” AND “online learning or e-learning or distance learning or remote learning or virtual learning” yielded 1559 records. From these records, nine (9) separate searches were performed for each of the following variables/constructs: AND “effectiveness or efficacy or effective” (n=414); AND “engagement or motivation” (n=175); AND “self-efficacy” (n=34); AND “ease of use or usefulness or usability” (n=53); AND “student perceptions or student attitudes” (n=250); AND “at-home or hands-on or lab kits” (n=293); AND “virtual lab or simulations or virtual manipulatives” (n=416); and, AND “adult learning or adult education or andragogy or adult learners” (n=10). This yielded a total of 1661 records of which 247 were selected on the basis of format (peer-reviewed journal articles, excluding meta-analyses/bibliometric analyses/literature reviews), title, and non-duplication.

Inclusion and Exclusion Criteria

The inclusion criteria narrowed the reports to studies that: (a) evaluated the efficacy of the non-traditional laboratory experience for a single science course; and (b) studies that focused on students’ performance in and/or students’ perceptions of the alternative laboratory experience. Publications were excluded if the study: (a) focused on a STEM discipline, course, topic, or skillset that lies outside of the scope of the natural sciences and engineering (math, nursing, psychology, pharmacology, pharmacy, animal behavior, computer science, information technology, networking, etc.) (n=64); (b) focused on teachers’ perceptions, teachers’ experiences, and/or teaching strategies (n=30); (c) focused on “how to” design an online science course and/or laboratory (n=30); (d) focused on findings from emergency remote learning (i.e.

alternative experiences quickly developed to “get through” the pandemic)(n=13); (e) focused on the use of a specific technology and/or mobile application (n=28); (f) focused on incorporating lecture videos as laboratory supplements (n=16); (g) focused on incorporating quizzes as pre-laboratory preparation (n=3); and (h) focused on more than one course (n=3). This screening produced 60 eligible articles. An additional 35 articles were added through a manual reference-list search of the selected articles and peer-reviewed meta-analyses, bibliometric analyses, and literature reviews resulting in a total of 95 studies included in this review.

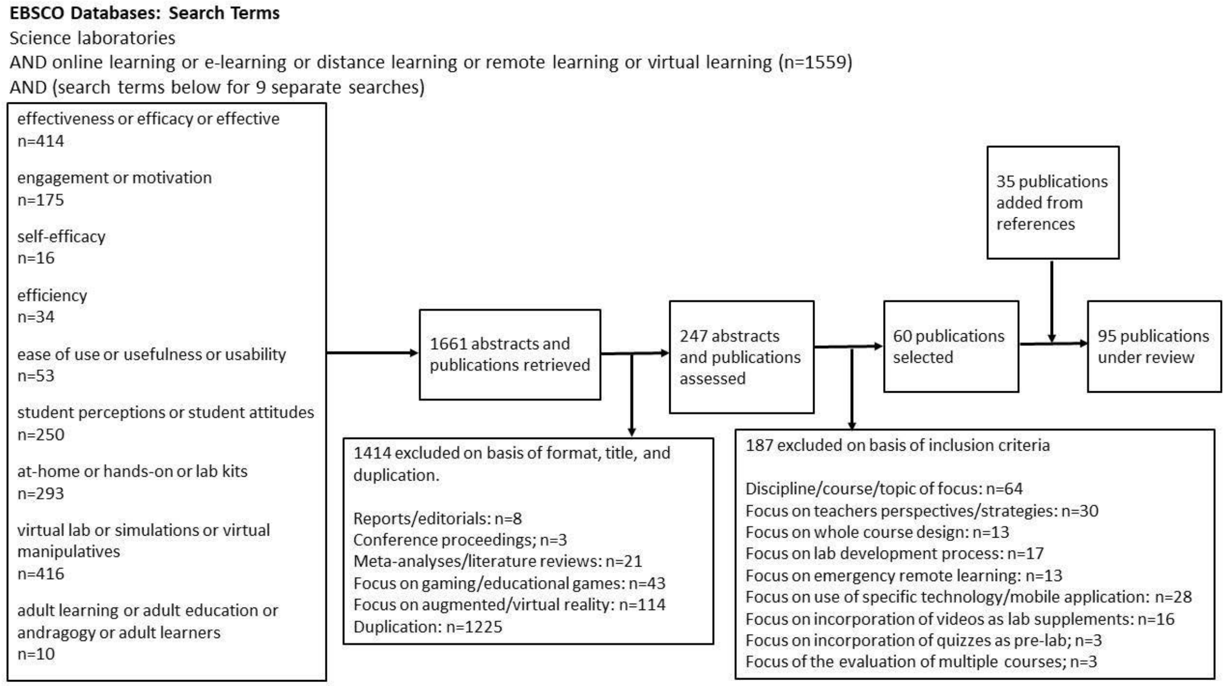
Secondary Search: Google Scholar

A secondary search was performed using Google Scholar. Adding the terms “online science laboratories”, “at-home science laboratories”, and “virtual laboratories” to the search bar yielded more than 900,000 reports; limiting the search to studies published in the same timeframe above (2002-2022) reduced the number of reports to approximately 20,000. However, after applying the same inclusion and exclusion criteria outlined above, no additional articles were added to the review. All publications that met the inclusion criteria were duplicates of those already identified through the EBSCO search.

Figure 2.2 illustrates the strategy utilized to search the EBSCO databases, filter, and select the studies (n=95) for this review. The Google Scholar search is not depicted in this figure as it did not yield any new studies for inclusion.

Figure 2.2

Flow Chart of Search Strategy: EBSCO Databases



Limitations of the Available Literature

At first glance (n=95 publications under review), it may appear that the selection criteria were too broad. However, the literature became extremely limited when the search criteria were narrowed. This issue was also noted in “A Review to Weigh the Pros and Cons of Online, Remote, and Distance Science Laboratory Experiences” by Faulconer and Gruss (2018). They determined that “the literature was neither robust enough nor was it homogenous enough” (p.157) to support a systematic review for a targeted population (Faulconer & Gruss, 2018). To illustrate this problem, when the articles selected for this review were filtered to include only those that address the specific population of interest for this study (i.e. adult learners in an online, nonmajors science class), only 8 of the 95 articles would meet these criteria (Table 2.2). While this supports the need for more research in this area, a pool of 8 is not sufficient for a meaningful review of the literature.

Table 2.2

Selected Articles that Included (or may have included) Adult Learners

Authors	Adult Learners (age >25)	Course M or NM	Meet Both Criteria (AL + NM)
Stuckey-Mickell & Stuckey-Danner (2007)	Age range: 18-55	Human Biology NM	Yes
Reuter (2009)	Mean age TL = 25 Mean age AHHO = 41	Soils: Sustainable Ecosystems NM	Yes
Mawn et al. (2011)	Mean age = 36	Integrated Science NM	Yes
Vogt et al. (2013)	TL (all <20) Blended (25-30)	Astronomy NM	Yes
Uribe et al. (2016)	70% participants, 26-40	Thermoelectricity M	No
Davies (2019)	Age ranges not reported; Findings mention AL	Chemistry M + NM	Maybe
Nguyen & Keuseman (2020)	≈ 30% participants > 25	Chemistry in the Kitchen NM	Yes
Schultz et al. (2020)	3.6% participants > 25	General Chemistry M + NM	Yes
Çivril and Özkul (2021)	Age range: 23-52	Circuit Analysis NM	Yes
Gutruf et al. (2021)	Age ranges not reported; Andragogy included in the title	Biomedical Engineering M	No
Youngblood et al. (2022)	Age ranges not reported 26% in-person 74% online (AL?)	Vertebrate Zoology M	No

Note. AL = adult learner; TL = traditional laboratory; AHHO = at-home, hands-on laboratory; M = majors-level course; NM = nonmajors-level course

Furthermore, it did not seem reasonable to exclude studies performed in middle and high schools as students in grades 7-12 are afforded the opportunity to take college-level laboratory science courses through dual enrollment programs (Xu et al., 2021). The inclusion and analysis of a broader range of studies (regardless of learners' age and level of course) gave rise to several themes that were used to develop the conceptual framework for this study. The following sections provide a general overview of the 95 selected studies and the themes that emerged from their summary analysis.

Overview of Selected Studies

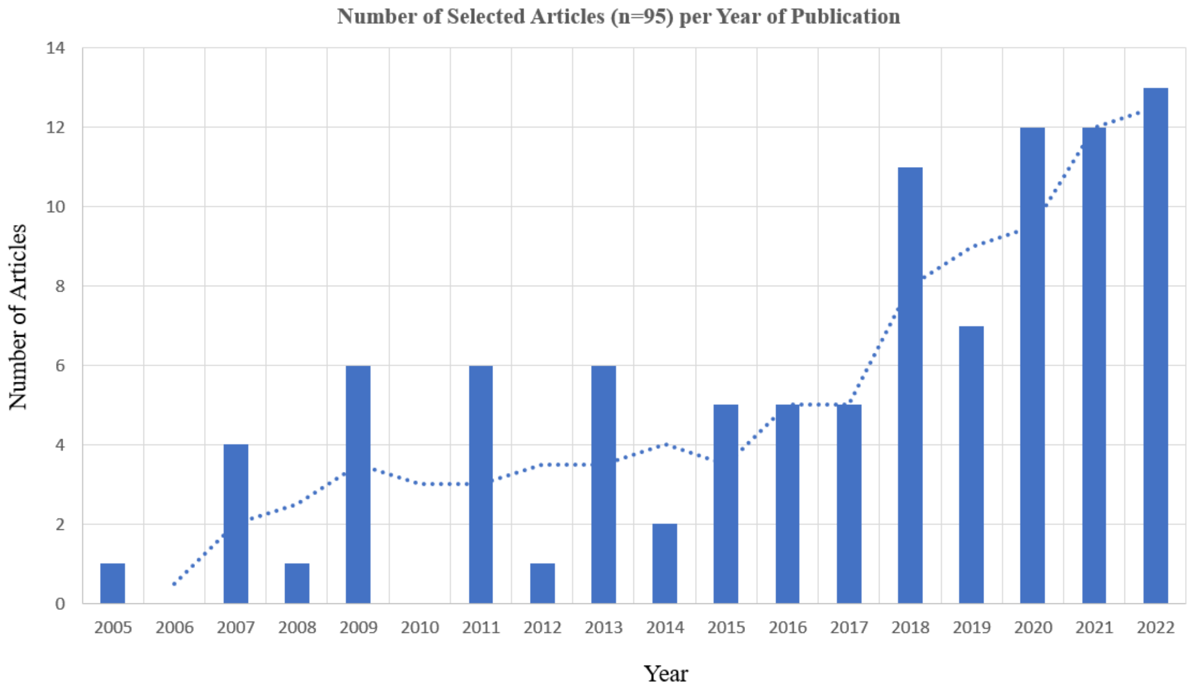
Number of Articles per Year of Publication

Extensive research has shown that hands-on laboratory activities play an essential role in science education (Clough, 2002; Deboer, 1991; Hofstein & Lunetta, 2003; Magin & Kanapathipallai, 2000). While there were some early adopters who embraced the use of new technologies to enhance or replace the traditional laboratory experience (Corter et al., 2007; Jara et al., 2009; Lindsay & Good, 2005; Koretsky et al., 2008; Koretsky et al., 2011; Zacharia, 2007; Zacharia et al., 2008), many science educators worried that online and/or computer-based laboratories deprived students of the hands-on experiences needed to construct deep learning and develop the practical skills necessary for their field (Clough, 2002; Dewhurst et al., 2000; DiBiase, 2000; Magin & Kanapathipallai, 2000).

However, in 2020, even those instructors that were hesitant to 'take the laboratory out of the lab' were forced to find viable alternatives due to the world-wide campus closures during the COVID-19 global pandemic. This created an opportunity for both students and faculty to reassess their perceptions of these alternative, at-home laboratory experiences (Sung et al., 2021; Unger & Merian, 2020; Visiliadou, 2020). Following their forced exposure to online learning, the number of students enrolled in distance education classes has nearly tripled since 2019 (NCES, 2021a). Faculty responded with a new wave of publications evaluating their students' performance in and perceptions of the alternative laboratory experiences (Achuthan et al., 2021; Barthet et al., 2021; Chen, 2022; Finne et al., 2022; Gnesdilow & Putambekar, 2022; Heng et al., 2022; Johnson & Barr, 2021; Rayment et al., 2022; Sithole et al., 2022; Sotelo et al., 2022; Stokes & Silverton, 2021; Youngblood et al., 2022). Figure 2.3 illustrates the number of selected articles per year of publication.

Figure 2.3

Summary of Selected Articles: Number of Articles per Year of Publication



Note. The bar graph illustrates the number of articles that met the inclusion criteria per year of publication; the dotted trendline represents the moving average in the number of selected publications per year. The trend in this sample of selected studies is similar to the trend reported in Raman et al., 2022 (Figure 2.1), including the sharp rise in the number of publications in 2020.

Types of Studies: Comparison, Non-Comparison, and Blended

The 95 selected publications were first organized by type of study based upon its research design. Each article was assigned to one of the following categories: (a) comparison studies (n=50) in which factors such as learning outcomes achievement, student satisfaction, and student perception of effectiveness, engagement, and/or ease of use of the alternative laboratory experience were compared with these same factors in the traditional face-to-face laboratories; (b) non-comparison studies (n=27) in which these factors were evaluated for the alternative laboratory experience (RL, VL, or AHHO) without the side-by-side comparison to the traditional laboratory experience (in studies where comparisons were made, they were drawn from student performance/student perception data from previous terms; these studies were not included with the comparison studies as the exams, laboratory reports, and/or surveys utilized in previous terms were not the same as those used for the study); and, (c) non-comparison studies (n=18) in which these factors were evaluated for an experience that “blended” the hands-on experience (most often performed in the campus laboratory) with elements of the virtual experience (simulations, manipulation of virtual objects, live-streamed or pre-recorded videos, and analysis images or pre-collected datasets). The blended studies were not included with the other non-comparison studies as the alternative experiences were not utilized for the laboratory as a whole. Rather, they were utilized as preparation for the campus laboratory (safety training, skills development, or practice with virtual manipulatives/laboratory equipment prior to hands-on campus experience), supplemental instruction for laboratory objectives (review the underlying principles), or for real-time collaboration (from different locations) while performing laboratory activities. Table 2.3 provides a summary of the types of studies and their definitions.

Table 2.3

Summary of Selected Articles: Types of Studies - Number, and Definition

Type of Study	Number	Definition
		Student performance and/or student perception data is:
Comparison	50	Compared with the same factors in the traditional face-to-face laboratories.
Non-comparison VL, RL, or AHHO	27	Evaluated for the alternative laboratory experience (RL, VL, or AHHO) without the side-by-side comparison to the traditional laboratory experience
Blended	18	Evaluated for an experience that "blended" the traditional laboratory experience with elements of the virtual laboratory experience

Note. VL = virtual laboratory experience; RL = remote laboratory experience; and AHHO = at-home, hands-on laboratory experience.

Summary Tables

Once the selected studies were classified by type, each was summarized and coded for further analysis in a set of source tables. The source tables provide a chronological view of the authors that have contributed to this field of study, the alternative lab experience(s) evaluated for each study, discipline/course/topic of study, level of education of learner, type of institution (and in which country) the study was performed, variables/constructs analyzed, quantitative and/or qualitative findings, the implications of the study, and whether or not adult learners were included (if age demographics were reported). The complete set of source tables is located in Appendix A (Source Table A-1, comparison studies; Source Table A-II, non-comparison studies [VL, RL, and AHHO], and Source Table A-III, non-comparison studies [blended]).

This information was organized into a series of five summary tables: Table 2.4 provides an overview of the STEM disciplines, educational levels, types of institution, and the countries represented in the (n=95) studies selected based upon the criteria discussed in the previous section; Table 2.5 provides an overview of the comparison study findings for effectiveness, engagement, and preferred laboratory experience; Table 2.6 provides an overview of the alternative laboratory experience(s) studied/evaluated and the variables/constructs used for the evaluation; Table 2.7 provides an overview of the metrics and methodologies used for evaluation; and, Table 2.8 provides an overview of the learners' age group (<25 or >25; i.e. whether or not the study included adult learners).

A critical analysis of the information presented in the summary tables revealed five major themes: (a) questions surrounding the efficacy of alternative laboratory experiences is prevalent in STEM education (Table 2.4); (b) it is difficult to ascertain students' preferred laboratory experience due mixed findings (Table 2.5) and because of a lack of generalizability (Table 2.6); (c) the primary variables measured in the selected studies closely align with those of usability (Table 2.6); (d) there is a lack of continuity in the metrics used to evaluate students' performance and perceptions of the alternative laboratory experiences (Table 2.7); and, (e) there is very little focus on adult learners' performance in and perception of alternative laboratory experiences (Table 2.8). The sections that follow will discuss each of the themes and how specific elements within each theme were used to inform the conceptual framework for this dissertation study.

Prevalence of the Research Problem in STEM Education

Summary Table 2.4 demonstrates that the question of which laboratory experience is the most effective, most engaging, and/or preferred by learners has impacted science educators from multiple STEM disciplines (biology, chemistry, physics, engineering, earth/natural sciences, and others), from nearly educational levels (middle school through medical school; introductory courses through advanced courses), from various types of institutions (K-12 and colleges/universities; public and private schools), and from all over the world (25+ countries).

Table 2.4

Summary of Selected Studies (n=95): Discipline, Type of Institution, Countries

Discipline	Type of Institution	Country
Biological Sciences (n=27)	K-12 (n=20)	United States (n=46)
General/Introductory Biology (n=7)	High School (n=12)	
Cellular/Molecular Biology (n=7)	Middle School (n=8)	International (n=49)
Anatomy & Physiology (n=8)		Canada (n=8)
Human Biology (n=1)	Higher Education (n=75)	Turkey (n=7)
Zoology/Animal Biology (n=3)	Medical School (n=3)	Australia (n=5)
Plant Biology (n=1)	T4Y (n=67)	Cyprus (n=3)
Physics (n=21)	Public RU (n=54)	Spain (n=3)
General/Introductory Physics (n=19)	Private (n=10)	China (n=2)
Upper-Level/Advanced Physics (n=2)	Private Ivy (n=1)	India (n=2)
Chemistry (n=22)	Branch (n=1)	Indonesia (n=2)
General/Introductory Chemistry (n=15)	Poly-Tech (n=1)	Taiwan (n=2)
Biochemistry (n=2)	4Y (n=3)	Other countries
Analytical/Physical (n=5)	Public (n=2)	(each @ n=1)
Integrated/General Science (n=4)	Private (n=1)	Denmark, Greece,
Earth/Natural Sciences (n=3)	2Y (n=1)	Malaysia, Manila,
Astronomy (n=1)	CC (n=1)	Mexico, Oman,
Geology (n=1)		Palestine, Russia,
Soils & Sustainable Ecosystems (n=1)		Scotland, Serbia,
Engineering (n=14)		Singapore,
Engineering/Design (n=12)		Slovenia,
Capstone (n=2)		Trinidad &
		Tobago,
Other (n=4)		United Kingdom
Forensic Science (n=2)		Germany/Croatia/
Food Science (n=2)		Austria

Note. T4Y = traditional 4-year university; RU = Research University; Poly-Tech = Polytechnical Institute; 4Y = 4-year college or university (large online presence); 2Y = 2-year college; CC= community college

While the problem is ubiquitous, the impetus to incorporate alternative laboratory experiences somewhat varies based upon the type of institution. Middle school and high school educators most often cited a lack of time, space, and resources to support a traditional laboratory experience (Ambusaidi et al., 2018; Chen et al, 2014; Gnesdilow & Putambekar, 2022; Kapici et al., 2019; Kelley, 2020; Klahr et al., 2007; Oser & Fraser, 2015; Pyatt & Sims, 2011; Tatli & Ayas, 2011). Although these same issues are present in higher education, post-secondary educators tended to incorporate alternative laboratory experiences to increase access to a broader population of students (Attardi et al., 2016; Barbeau et al., 2013; Mawn et al., 2011; Meintzer et al., 2017; Reuters, 2009; Rosen & Kelley, 2020; Stokes & Silverthorn, 2021; Stuckey-Mickell & Study-Danner, 2007), increase access to equipment (Achuthan et al., 2021; Cossovich et al., 2020; Corter et al, 2007; Paxinou et al., 2020), and to alleviate issues of safety and over-crowding (Reece & Butler, 2017).

Mixed Findings: Effectiveness, Engagement, and Preferred Experience

Research has shown that hands-on laboratory activities play a critical role in STEM fields of study (Clough, 2002; Hofstein & Lunetta, 2003; Hofstein & Mamlok-Naaman, 2007; Ma & Nickerson, 2006). As such, student performance and student perception of effectiveness, engagement, and/or ease of use in the alternative laboratory experiences have been extensively studied and carefully compared with these same factors in the traditional face-to-face laboratories (Appendix A-I). Although educators have conducted these comparison studies for more than two decades (Bhute et al., 2021; Raman et al., 2022; Reeves & Crippen 2021; Wei et al., 2019), there are still no clear answers as to which laboratory experience is the most effective, most engaging, or preferred by learners (Tsihouridis et al., 2018). These questions are difficult to answer from the available literature, in part, due to their mixed results. While the empirical data overwhelmingly showed that alternative laboratory experiences were at least as effective as their traditional counterparts (Table 2.5a), engagement was consistently greater in modalities that incorporated the opportunity to interface with or physically manipulate laboratory equipment (Table 2.5b).

Effectiveness

Summary Table 2.5a illustrates that students were able to learn key concepts in both laboratory formats. Ninety-three percent (93%) of the selected comparison studies reported that scores on assignments, pre-/post-tests, and/or grade distributions in the non-traditional laboratory that were greater than or equal to those in the traditional laboratory; of these, 80% showed there was no significant difference in student performance. Oser and Fraser (2015) posited that the finding of no significant difference was as important as the finding of a significant difference, as this affirmed that incorporating alternative laboratory experiences was not detrimental to outcomes achievement.

While the performance data overwhelmingly indicated the alternative experiences are as effective as the traditional laboratory experience, the student perception data was not as clear. Achuthan et al. (2021), Klahr et al. (2007), Koretsky et. al. (2011) reported that students perceived the alternative experience to be more effective because they were able to perform the experiment multiple times (without the time limitations of the traditional experience); this allowed students to engage in the iterative experimental design approach inherent to real-life engineering projects. On the other hand, Ambusaidi et al. (2018), Chen et al. (2014), Finne et al. (2022), and Johnson and Barr (2021) found that students perceived greater gains in laboratory skills and conceptual understanding in the traditional section. These findings are consistent with those from Meintzer et.al (2017) in which students indicated they learned more from the hands-on laboratory experience; the investigators postulated the technology may been a distraction. Similarly, Chen et al. (2014) reported that students exhibited an unrealistic trust in technology (i.e. they did not question the results generated by the simulation) and that the virtual manipulation of objects was not as memorable as the physical manipulation of objects.

Table 2.5a

Summary of Selected Articles: Comparison Studies Findings - Effectiveness

Findings	Authors
AL ≥ TL	
(n=32)	Studies that reported no significant difference in student performance Corter et al., 2007; Klahr et al., 2007; **Stuckey-Mickell & Stuckey-Danner, 2007; Reuter, 2009**; Koretsky et al., 2011; Olympiou & Zacharia, 2011; Pyatt & Sims, 2011; Barbeau et al., 2013; Brewer et al., 2013; Tatli & Ayas, 2013; Chen et al., 2014; Attardi & Rogers, 2015; Crandall et al., 2015; Oser & Fraser, 2015; Tekbiyuk & Ercan, 2015; Son & Narguizian, 2016; Meintzer et al., 2017; Reece & Butler, 2017; Ambusaidi et al., 2018; Javier & Lomuntad, 2018; Miller et al., 2018; Špernjak & Šorgo, 2017; Gunawan et al., 2019; Moosvi et al., 2019; Reck et al., 2019; Cossovich et al., 2020; Rosen & Kelley, 2020; Johnson & Barr, 2021; Sindelar & Witkowski, 2021; Chen, 2022; Gnesdilow & Putambekar, 2022; Rayment et al., 2022
(n=8)	Studies that reported learning gains in AL Zacharia, 2007; Zacharia et al., 2008; **Vogt et al., 2013**; Sari Ay & Yilmaz, 2015; Nolen & Koretsky, 2018; Kapici et al., 2019; Achuthan et al., 2021; DeChenne-Peters et al., 2022
AL < TL	Studies that reported significant differences in student performance
(n=3)	Kelley, 2020; **Schultz et al., 2020**; Stokes & Silverthorn, 2021

Note. AL = alternative laboratory experience; TL = traditional laboratory experience; **bold** = study included adult learners. Effectiveness was determined by comparing scores on assignments (lab reports, quizzes, or exams), grade distributions, gains pre-/post-test scores, student feedback (such as, “I learned more in the simulation” or “I was better able to answer questions in the traditional lab”) and/or survey responses for the AL and TL.

Engagement and Preferred Laboratory Experience

Summary Table 2.5b shows that engagement tended to be greater when all (TL) or part (blended) of the laboratory activities were performed in the campus laboratory.

Table 2.5b

Summary of Selected Articles: Comparison Studies Findings – Engagement and Preferred Laboratory Experience

Findings	Authors
Engagement	
AL ≥ TL	
(n=7)	Koretsky et al., 2011; Pyatt & Sims, 2011; Reece & Butler, 2017; Nolen & Koretsky, 2018; Cossovich et al., 2020; **Schultz et al., 2020**; Achuthan et al., 2021
AL < TL	
(n=15)	Corter et al., 2007; **Stuckey-Mickell & Stuckey-Danner, 2007**; Brewer et al., 2013; **Vogt et al., 2013**; Chen et al., 2014; Crandall et al., 2015; Javier & Lomuntad, 2018; Moosvi et al., 2019; Kelley, 2020; Rosen & Kelley, 2020; Johnson & Barr, 2021; Sindelar & Witkowski, 2021; Stokes & Silverthorn, 2021; Finne et al., 2022; Rayment et al., 2022
Preference	
TL	
(n=10)	Lindsay & Good, 2005; Corter et al., 2007; **Stuckey-Mickell & Stuckey-Danner, 2007**; West, 2012; Attardi et al., 2016; Meintzer et al., 2017; Johnson & Barr, 2021; Sindelar & Witkowski, 2021; Stokes & Silverthorn, 2021; Finne et al., 2022
AL	
(n=2)	Reece & Butler, 2017; Cossovich et al., 2020
Blended	
(n=13)	West, 2012; Barbeau et al., 2013; Chen et al., 2014; Attardi & Rogers, 2015; Crandall et al., 2015; Sari Ay & Yilmaz, 2015; Tekbiyuk & Ercan, 2015; Son & Narguizian, 2016; Ambusaidi et al., 2018; Miller et al., 2018; Špernjak & Šorgo, 2017; Kapici et al., 2019; Gnesdilow & Putambekar, 2022

Note. AL = alternative laboratory experience; TL = traditional laboratory experience; **bold** = study included adult learners. Engagement was determined by comparing student feedback (such as, "the TL was more fun"), survey responses, completion rates, and persistence. Preference was determined from student feedback and/or survey/questionnaire responses.

As with effectiveness, the student perception data was mixed. Pyatt and Sims (2011) found students did not prefer one experience over the other; Barbeau et. al. (2013) reported the laboratory experience did not affect student satisfaction; Reck et al. (2019) reported that students in both groups enjoyed their laboratory experiences; Sithole et al. (2022) reported perceived learning in all laboratory experiences; and, Reece and Butler (2017) found a similar decline in motivation to study biology in both the traditional and alternative laboratories. While students enjoyed the flexibility and convenience of the alternative laboratories (Attardi et al., 2016; Corter et al., 2007; Kelley, 2020; Rayment et al., 2022; Sindelar & Witkowski, 2021; West, 2012), many expressed that they felt they were “missing out”. The two most common weaknesses identified in the alternative laboratory experiences were their lack of physicality (Ambusaidi et al., 2018; Attardi et al., 2016; Chen et al., 2014; Corter et al., 2007; Crandall et al., 2015; Johnson & Barr, 2021; Kelley, 2020; Lindsay & Good, 2005; Stuckey-Mickell & Stuckey-Danner, 2007; Miller et al., 2018; West, 2012) and their lack of student-student and student-instructor interaction (Attardi et al., 2016; Corter et al., 2007; Crandall et al., 2015; Finne et al., 2022; Javier & Lomuntad, 2018; Johnson & Barr, 2021; Kelley, 2020; Meintzer et. al. 2017; Miller at al., 2018; Rosen & Kelley, 2020; Sindelar & Witkowski, 2021; Stokes & Silverthorn, 2021; Stucky-Mickell & Stuckey-Danner, 2007).

At-home, hands-on laboratory experiences can address both identified weaknesses. When coupled with synchronous class meetings, these ‘blended’ laboratories strike the perfect balance between students’ ‘touch’ preferences (Cossovich et al., 2020; Moosvi et al., 2019; Rayment et al., 2022; Sithole et al., 2022; Vogt et al., 2013) and their needs for social-interaction (Chen, 2022; Jara et al., 2009; Saitta et al., 2011). These at-home, hands-on laboratory experiences have proven to be an effective and engaging in K-12 (Kelley, 2020; Zidny et al., 2019) and in traditional, undergraduate (Al-Soufi et al., 2020; Andrews et al., 2020; Barthet, 2021; Cossovich et al., 2020; Doughan & Shahmuradyan, 2021; Easdon, 2020; Gutruf et al., 2021; Howard & Meier, 2021; Mawn et al., 2011; Moosvi et al., 2019; Nguyen & Keuseman, 2020; Rayment et al., 2022; Rosen & Kelley, 2020; Sindelar & Witkowski, 2021; Sithole et al., 2022; Vogt et al., 2013) student populations. Unfortunately, there is a paucity of research in their use with adult learners.

Lack of Generalizability

Summary Table 2.6 further demonstrates why it is difficult to discern which laboratory experience is the most effective, most engaging, or preferred by learners from the literature. The studies suffer from a problem that is common in educational research; each was designed to meet the very specific needs for a very specific student population at a very specific institution with a very specific alternative laboratory experience (Gopalan et al., 2020). Table 2.6 shows the wide range of alternative laboratory experiences investigated in the 95 selected studies. There are seven variations of the TL versus VL comparison studies alone!

While the most common instance of the TL versus VL is a comparison of the traditional laboratory experience with a simulation (n=19), differences in quality of the simulation, purpose of the simulation, learning outcomes for the laboratory, student demographics, class size, and location make it nearly impossible to generalize these findings to students at another institution (Gopalan et al., 2020). Furthermore, the use of convenience sampling, lack of continuity of metrics (discussed in another section), and lack of replicability makes it difficult to assess like-variables across studies and establish parallelism (Gopalan et al., 2020). As such, educators are able to glean some ‘lessons learned’ from the literature, but are not likely to find direct answers for which experience would be most effective for or preferred by their students.

Table 2.6

Summary of Selected Studies: Type of Study, Alternative Laboratory Experience(s) Studied/Evaluated, and the Variables/Constructs Evaluated

Type of Study Alternative Laboratory Experience(s) Studied/Evaluated		Variables/Constructs Evaluated
Comparison Studies (n=50)		
TL vs VL (simulation)	n=19	Effectiveness (n=93)
TL vs VL (virtual manipulatives)	n=1	Engagement (n=50)
TL vs VL (videos + dataset)	n=3	Efficiency (n=19)
TL vs VL (virtual manipulative + videos)	n=1	Self-efficacy (n=23)
TL vs VL (simulation + virtual manipulative)	n=1	Error tolerance (n=2)
TL vs VL (simulation + virtual manipulative + videos)	n=2	Ease of Use/ (n=24) Usability/Usefulness
TL vs VL (home supplies + datasets + videos)	n=1	
TL vs RL	n=2	
TL vs AHHO (kits)	n=5	Attitudes (n=11)
TL vs AHHO (home supplies)	n=2	Interest (n=1)
TL vs HO (kits, utilized in the traditional lab)	n=1	Motivation (n=6)
TL vs B (dataset + images + AHHO, home supplies)	n=1	Satisfaction (n=2)
RL vs VL (simulation)	n=1	
TL vs VL (simulation) vs B (TL + VL)	n=3	
TL vs VL (simulation) vs RL	n=2	
TL vs VL (virtual manipulative) vs B (TL + VL)	n=2	
TL vs VL (videos) vs AHHO (home supplies)	n=1	
TL vs AHHO (kits) vs AHHO (home supplies)	n=1	
VL (simulations) vs AHHO (kit) vs B (VL + AHHO)	n=1	
Non-comparison Studies - VL, RL, or AHHO (n=27)		
VL (simulation)	n=4	
VL (dataset + video)	n=1	
VL (simulation + dataset + video)	n=1	
VL (simulation + virtual manipulative)	n=1	
RL	n=4	
AHHO (kit)	n=9	
AHHO (home supplies)	n=6	
AHHO (kit) + VL (simulation)	n=1	
Non-comparison Studies - Blended (n=18)		
VL as Pre-Laboratory Preparation	n=11	
VL as Supplemental Instruction	n=4	
VL with real-time collaboration	n=1	
AHHO with real-time collaboration	n=1	
AHHO + VL (order of experience)	n=1	

Note. TL = Traditional Laboratory (hands-on, in-person, on-campus), VL = Virtual Laboratory (as indicated), RL = Remote Laboratory (virtual manipulation of real equipment), AHHO = At-Home Hands-On (as indicated), HO = Hands-On (as indicated); B = blended (as indicated)

Variables from Selected Studies Align with Perceived Usability

Summary Table 2.6 also illustrates the variables that were measured in the selected studies: effectiveness (n=93), engagement (n=50), ease of use (n=24), self-efficacy (n=23), efficiency (n=19), attitudes towards science (n=11), motivation (n=6), student satisfaction (n=2), and interest in science (n=1). These variables largely align with those of ‘perceived usability’.

What is Perceived Usability?

The term ‘usability’ was coined in the early 1980s to replace the term ‘user friendly’ (Bevan, 1995; Lacerda & von Wangenheim, 2018; Quesenbery, 2001). Although this term has been in use for over 40 years, there still is no single, accepted definition that fully encompasses all aspects of usability (Bevan, 1995; Lewis, 2018b). As usability can be evaluated from multiple perspectives (product, user, and performance), there most likely will never be absolute consensus in the literature (Bevan, 1995; Lacerda & von Wangenheim, 2018; Lewis, 2018a). This section will explore two definitions that focus on the user experience or ‘perceived usability’.

International Organization for Standardization

The current definition of perceived usability, as maintained by the International Organization for Standardization, is found in ISO 9241-11:2018 section 3.1.1:

> “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use”

This definition contains several important elements that align with concepts in teaching and learning (Pal & Vanijja, 2020; Stuckey-Mickell & Stuckey-Danner, 2007; Vlachogianni & Tselios, 2021). First, the phrase ‘product can be used by specified users’ denotes that the use of the product should be aligned with the specific traits of the user (Bevan, 1995; Brooke, 1996;

International Organization for Standardization, 2018; Lewis, 2018a). Whether the product is a calculator, piece of educational software, learning management system, or laboratory protocol, the product should be developed to meet the specific teaching and learning needs for the specific student population (Quiñones et al., 2018; Shackel, 2009; Vlachogianni & Tselios, 2021, 2022).

Next, the terms ‘effectiveness’, ‘efficiency’, and ‘satisfaction’ denotes that the quality of the product is directly related to the quality of the user experience (Bevan, 1995; Iwarsson & Ståhl, 2003; Quiñones et al., 2018; Shackel, 2009). ISO 9241-11:2018 specifically defines the terms in the following way: (a) effectiveness is the extent to which the product can be used by specified users to achieve specified goals (3.1.12); (b) efficiency is the extent to which specified users can complete assigned tasks within the allotted timeframe (3.1.13); and (c) satisfaction is the extent to which the product meets the user’s needs and expectations (3.1.14). The definitions for these terms are consistent with how they are used in educational contexts if the term ‘user’ is replaced with student (Pal & Vanijja, 2020; Vlachogianni & Tselios, 2021, 2022). In this case, user experience data can be viewed as equivalent to student experience (or student perception) data.

Quesenbery’s (2001) 5 E’s of Perceived Usability

In 2001, Quesenbery argued that the ISO 9241-11:2018 section 3.1.14 definition for the variable ‘satisfaction’ was too broad. To gain a deeper understanding of how the product meets the user’s needs and expectations, Quesenbery (2001) broke this variable down into engagement, error tolerance, and ease of use. In this context, ‘engagement’ is defined as the extent to which the product was pleasant and satisfying to use (i.e. specified users enjoyed the experience); ‘error tolerance’ is the extent to which the product helps users recover from errors; and ‘ease of use’ is the extent to which specified users are able to use the product without intervention (Iwarsson &

Ståhl, 2003; Quesenbery, 2001; Shackel, 2009). Again, when the term 'user' is replaced with 'student', these variables are consistent with student perception data measured in the selected studies. As such, perceived usability is a more wholistic measurement of student perception of effectiveness and engagement. Figure 2.4 illustrates the alignment of the variables from the selected studies with the definitions for perceived usability.

Figure 2.4

Alignment of Variables Measured in Selected Studies with Usability Definitions

Usability ISO 9241-11:2018	Variables Measured in Selected Studies		Usability Quesenbery (2001)
Effectiveness	Effectiveness	(98%)	Effectiveness
Efficiency	Efficiency	(20%)	Efficiency
Satisfaction	Satisfaction	(2%)	
	Engagement	(53%)	Engagement
	Error Tolerance	(2%)	Error Tolerance
	Ease of use	(25%)	Ease of Use
	Self-efficacy	(24%)	
	Attitudes	(12%)	
	Motivation	(6%)	
	Interest	(1%)	

Note. The variables listed for the selected studies are those that were identified by a critical analysis of their findings; the % represents the percentage of studies in which each variable was measured. The variables listed for ISO 9241-11:2018 and Quesenbury (2001) were taken directly from their definitions for perceived usability. The variables for usability are closely aligned with the variables evaluated in the selected studies.

Lack of Continuity in Metrics

As science educators added new technologies to enhance and/or replace the traditional laboratory experience, factors such as learning outcomes achievement, student satisfaction, and student perception of effectiveness, engagement, and/or ease of use of the alternative laboratory experiences have been studied extensively (Appendix A); 95 studies were selected for this literature review. While 93% of the selected studies claimed that the alternative laboratory experience was at least as effective as its traditional counterpart (Table 2.5a), this result is not as conclusive as it may seem.

Table 2.7 demonstrates the wide variability (and lack of continuity) in the metrics used to evaluate students' performance in and perception of their alternative laboratory experiences.

Table 2.7

Summary of Selected Studies: Metrics and Methodology for Evaluation

Metrics		Methodology
Student Performance	Student Perception	
Pre-/Post-Knowledge Test (n=29)	Self-Developed	Quantitative (n=50)
Pre-/Post-Skills Test (n=3)	Survey/Questionnaire (n=32)	Qualitative (n=2)
Scores	Pre-/Post-Survey (n=15)	Mixed-Method (n=42)
Lab Reports/Projects (n=35)	Normed/Validated	
Quizzes/Exams (n=20)	Survey/Questionnaire (n=5)	
Class Assignments (n=9)	Pre-/Post-Survey (n=2)	
Lab Notebook (n=1)	Interviews (n=7)	
Standardized Exams (n=3)	Focus Group (n=3)	
Final Grade in Course (n=16)	Student Reflections (n=16)	
Rubric (n=2)	Student Feedback	
Algorithm (n=1)	Course Evaluation (n=26)	
Persistence/Completion (n=6)	Assignments* (n=8)	
# Discussion Board Posts (n=2)		
Quality of Answers (n=4)		

Note. *Indicates that the assignments included open-ended questions added for research purposes

Only 10 of the selected studies utilized standardized, normed, or validated metrics; most utilized home-grown assignments (quizzes, exams, laboratory reports, projects, pre-/post-tests, etc.), final grades, or self-developed surveys/questionnaires/interview questions to assess their students. As previously mentioned, the pervasive use metrics that were developed by a specific instructor for a specific section of a specific class to meet the specific learning objectives at a specific institution make it nearly impossible to generalize these findings to students taking the same course at another institution (Gopalan et al., 2020). This highlights the need for a normed and validated metric that can be used to evaluate student perception data across a myriad of laboratory experiences. The System Usability Scale (SUS) can address this need.

What is the System Usability Scale (SUS)?

As discussed in the previous section, perceived usability is a more wholistic measurement of student perception of effectiveness and engagement. Perceived usability can be measured using Brooke's (1996) System Usability Scale (SUS). The SUS is one of the most frequently used instruments in usability studies (Bevan, 1995; Brooke, 2013; Lewis, 2018a; Lacerda & von Wangenheim, 2018) and is consistent with both the ISO definition for usability (Lewis, 2018b; Lewis & Sauro, 2019) and Quesenbery's (2001) 5 E's of perceived usability.

The popularity of the SUS can be attributed to several factors, including: (a) it is free to use (with acknowledgement to Brooke) and readily available to the public; (b) it is simple to use and can be completed in just a few minutes; (c) it has been standardized, validated, normalized, and is extremely reliable; (d) minor modifications to the questions does not affect validity or reliability; and (e) it has been determined to be reliable, even in small sample sizes (Bangor et al., 2008; Bevan, 1995; Brooke, 1996; Brooke, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016). The SUS has successfully been used to measure the perceived usability of a wide range of products, from thermometers to mobile devices (Lacerda & von Wangenheim, 2018). Vlachogianni & Tselios (2021, 2022) used the SUS to evaluate students' perceptions of educational technologies; they concluded that the SUS is a useful tool that should be utilized by more instructors.

Lack of Focus on Adult Learners

Summary Table 2.8 demonstrates the lack of focus on adult learners' perceptions of alternative laboratory experiences. Given that adults over the age of 25 account for 25% of total student enrollment in higher education in the United States (NCES, 2021b), this represents a significant gap in the literature.

Table 2.8

Summary of Selected Studies (n=95): Learner's Ages (<25, >25, or not reported)

Learners' Ages		Type of Institution
No Adult Learners (all participants <25 years old)	n=25	K-12 (n=20) T4Y Public, Research University (n=3) T4Y Private (n=1) Medical School (n=1)
Included Adult Learners (at least 1 participant >25 years old)	n=8	Community College (n=1) 2-Year College (n=1) 4-Year Commuter Campus (n=1) T4Y Public, Branch Campus (n=1) 4-Year Private, with online presence (n=1) T4Y Public, Research University (n=3)
Age Ranges Not Reported	n=62	T4Y Public, Research University (n=48) T4Y Private, Ivy (n=1) T4Y Private (n=9) 4Y Public, with online presence (n=1) Poly Technical Institute (n=1) Medical School (n=2)

Note. T4Y = Traditional 4-Year; Based upon the types of institutions that reported the inclusion of adult learners (AL), it is not likely that the studies that did not report ages (n=62) included adult learners.

What makes Adult Learners Different?

Malcolm Knowles defined andragogy as "the art and science of helping adults learn" (Knowles, 1980, p. 43). He posited that adult learners tend to be more independent, self-directed, and self-motivated than traditional students (Knowles, 1980, 1984). Furthermore, adult learners bring a wealth of experiences to the classroom and learn best when they are involved in the planning and evaluation of their learning (Remenick & Goralnik, 2019; Taylor & Hamdy, 2013). Finally, adult learners place a high value on relevant, active, real-world learning activities that are directly tied to learning objectives (Arghode et al., 2017; Calabrese & Capraro, 2021; Chen, 2014; Knowles, 1980; Loeng, 2018; Pratt, 1993; Taylor & Hamdy, 2013). In short, adult learners' time is valuable; the laboratory experience should be well-designed (so they can be performed at-home with little to no extra support) and be meaningful (support the development of conceptual knowledge or practical skills).

Impact of Instructional Design

Studies have shown that teaching approaches that are considered to be highly effective in traditional student populations may be not be as effective (or even ineffective) with an adult student population (Arghode et al., 2017; Chen, 2014; Gutruf et al., 2021; Knowles, 1984; Lewis & Bryan, 2021; Loeng, 2018; Pratt, 1993; Remenick & Goralnik, 2019; Taylor & Hamdy, 2013). Although the enrollment for students over the age of 25 has outpaced that of traditionally aged students for the last decade (Chen, 2014; NCES, 2021b), most courses are still designed for traditional students (Chen, 2014; Greene & Larsen, 2018; Lewis & Bryan, 2021; Remenick & Goralnik, 2019). As such, the user experiences reported in 92% of the selected studies (Table 2.8) may not hold true for adult learners.

Nolen and Koretsky (2018) found that student interest and engagement in the laboratory experience were more closely tied to the instructional design than to the mode of delivery. This point is well-illustrated by the student perception data reported by Gutruf et al. (2021) and Moosvi et al. (2019). Both of these studies utilized at-home, hands-on laboratory kits (i.e. incorporated the same type of laboratory experience), however, students' perceptions of their experiences were vastly different.

In the study, "Moving from Pedagogy to Andragogy in Biomedical Engineering Design: Strategies for Lab-at-Home and Distance Learning", Gutruf et al. (2021) intentionally designed their alternative laboratory experience to meet the needs of adult learners. They afforded each team of students the opportunity to be involved in the planning and evaluation of their learning; each team created their own learning contract that outlined the roles and responsibilities for each member, when and how often they would meet, designated the milestones for each stage of the design, the operational plan, and the assessment/evaluation plan for the overall project and each team members' contributions (Gutruf et al., 2021). Student feedback indicated that the laboratory experience was widely successful and helped them to alleviate the stress of forced distance learning during the initial phase of the COVID-19 pandemic (Gutruf et al., 2021).

On the other hand, Moosvi et al. (2019) were not as intentional in their laboratory design. While the students enjoyed the flexibility of the at-home, hands-on laboratory experience; they reported that the lab activities felt a bit childish and did not consider them to be real labs (Moosvi et al., 2019). Although neither of these studies reported the inclusion of adult learners, they clearly demonstrate that an andragogical approach to laboratory design can positively impact students' perception of their laboratory experiences (Knowles, 1980; Nolen & Koretsky, 2018).

Cost Factors

Financial Costs. Cost is another design factor that can impact students' perception of their alternative laboratory experience (Bhute et al. 2021; Faulconer & Gruss, 2018). The costs to students vary by institution (whether or not lab fees are assessed), by laboratory experience (Table 2.9), by discipline (purchase of required software, safety equipment [splash-proof googles, laboratory coat/apron, and gloves], or other materials [scientific calculators, laboratory notebook, etc.]), and by the purchase price for the laboratory manual (Faulconer & Gruss, 2018). While at-home laboratory experiences can alleviate the high costs to maintain and staff a traditional laboratory for the institution (Brinson, 2015; Hofstein & Lunetta, 2003; Ma & Nickerson, 2006), these cost savings are often realized at the expense of the student. Table 2.9 provides cost comparisons for at-home virtual simulations and at-home, hands-on laboratory kits.

As these costs tend to be higher than the typical laboratory fee (ranges from $150 - $300; NCES, 2021c), the financial burdens of the at-home experience can be a dissatisfier for students. While students appreciated the flexibility and convenience of the at-home laboratories (Corter et al., 2007; Rayment et al., 2022; Reece & Butler, 2017; West, 2012), they were frustrated with a perceived lack of alignment with the course objectives (Javier & Lomuntad, 2018; Stuckey-Mickell & Stuckey-Danner, 2007) and the lack of just-in-time feedback/guidance from their instructor (Crandall et al., 2015; Finne et al., 2022; Johnson & Barr, 2021; Kelley, 2020; Miller at al., 2018; Rosen & Kelley, 2020; Stokes & Silverthorn, 2021; Stucky-Mickell & Stuckey-Danner, 2007). In other words, the students were asked to pay more to 'get less' than students in the traditional laboratory experience (Finne et al., 2022; Kelley, 2020).

Furthermore, the added financial burdens of the at-home laboratory experience may reduce the possibility of college to the adult learner (Osam et al., 2017). Maintaining a home, paying bills, shouldering the costs to care for children and/or aging parents can severely limit the available resources for college tuition and fees (Bowers & Bergman, 2016; Osam et al., 2017).

Table 2.9

Costs for At-Home Laboratories: Virtual Simulations and Laboratory Kits

At-Home Lab Experience	Cost to Student	Definition
Simulations		
High quality	$80-$100 per simulation/student Examples: Reece & Butler (2017); Senapati (2022).	Students complete a laboratory experiment in 3-D virtual environment; tend to include built in supplemental materials, auto-graded quizzes, and just-in-time help/feedback (via a chat bot or virtual teaching assistant); can be purchased individually by students or as a subscription by the institution.
Lower quality	$50-$100 per 8-10 simulations/student Examples: Senapati (2022); Stuckey-Mickell & Stuckey-Danner (2007)	Students complete laboratory activities; often purchased by individual students as supplement or as part of the textbook package; very little guidance for activities.
Laboratory Kit		
Specialized kit	$800 - $1200 per kit/student Example: Londino-Smolar & Hansel (2021)	Students order kit; contains specialized equipment for one or more laboratory activities/experiments.
Standard kit	$200-$600 (all labs for term) per kit/student Example: Orozco (2017)	Students order kit; contains all materials/equipment for all labs for the term; may or may not include the laboratory manual.

Time Costs. To make laboratory activities accessible to distance students without increasing the cost to students, some institutions utilized free or low-cost simulations/virtual manipulatives (Klahr et al., 2007; Zacharia, 2007) or developed laboratory protocols that could be completed with supplies that students may have in their home or could be purchased from a local store (Al-Soufi et al., 2020; Andrews et al., 2020; Doughan & Shahmuradyan, 2021; Easdon, 2020; Kelley, 2020; Moosvi et al., 2019; Nguyen & Keuseman, 2020; Schultz et al., 2020; Vogt et al., 2013; Youssef et al., 2021). While these lower-cost alternatives did not have the same financial burdens for the students, they cost students another resource that is precious to adult learners - time (Al-Soufi et al., 2020; Andrews et al., 2020; Doughan & Shahmuradyan, 2021; Easdon, 2020; Moosvi et al., 2019).

Al-Soufi et al. (2020), Andrews et al. (2020), and Doughan & Shahmuradyan (2021) developed laboratory activities that incorporated the use of home supplies; students prepared samples and analyzed the solutions with an app of their smartphones. These studies are exciting because they have extended even those experiments that require specialized equipment (a photometer) to the at-home experience. However, these experiments that are simple to complete in the traditional laboratory were fraught with difficulties for the students: they required learners to build their own equipment, do a large amount of troubleshooting, and perform the experiment multiple times (Al-Soufi et al., 2020; Andrews et al., 2020; Doughan & Shahmuradyan, 2021). Students reported that balancing the time-investment for the laboratory experience with their other coursework to be challenging (Andrews et al., 2020). This is consistent with Easdon (2020) and Moosvi et al. (2019); students experienced difficulties performing the experiments at home due to lack of counter-space, difficulties obtaining supplies, and large time amount of time spent troubleshooting issues. These added time burdens can be especially troublesome for adult learners as they have a significant number of time commitments (Bowers & Bergman, 2016; Osam et al., 2017). Given the rise in enrollment of adult learners (NCES, 2021b), it would behoove distance instructors to incorporate design strategies that appeal to the adult learner.

Chapter Summary

This chapter provided a systematic review of the available literature published on student performance in and student perception of the effectiveness, engagement, and/or ease of use in the alternative laboratory experiences. After an extensive search of the EBSCO databases, 95 articles were selected for review. The inclusion and analysis of a broad range of studies (regardless of learners' age and level of course) gave rise to several themes that were used to develop the conceptual framework for this study.

While the literature demonstrated that the question of which laboratory experience is the most effective, most engaging, and/or preferred by students is prevalent in STEM education, there was very little focus on the perceptions of adult learners. To capture adult learners' perceptions of at-home laboratory experiences, and to avoid the issues identified with educational research (lack of generalizability), this study utilized a modified version of Brooke's (1996) System Usability Scale (SUS) to evaluate the perceived usability of at-home, hands-on and at-home, virtual laboratory experiments. The SUS is a standardized metric that generates results that are independent of the product (Brooke, 1996; Brooke, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016); and Quesenbery's (2001) 5E's of perceived usability was shown to provide a more wholistic view of the adult learners' laboratory experiences (as it encompassed the variables effectiveness, efficiency, engagement, error tolerance, and ease of use). The next chapter will provide an in-depth discussion of the research design and methodology.

Chapter 3 – Research Methodology

Despite a wealth of research examining students' perceptions of alternative laboratory experiences (Brinson, 2015; Faulconer & Gruss, 2018; Heradio et al., 2016; Ma & Nickerson, 2006; Sypsas & Kalles, 2018; Tsihouridis et al., 2018), very little focus has been aimed at the perceptions of the adult learner. As students' perceptions affect how and what they learn (Greene & Larsen, 2018; Ramsden, 1979; Tinto, 2017), laboratory experiences will be more effective and engaging if they are aligned with the specific needs and preferences of the targeted student population (Arghode et al., 2017; Chen, 2013; Faulconer & Gruss, 2018; Iwarsson & Ståhl, 2003; Lewis & Bryan, 2021; Rowe et al., 2017). Hence, the primary purpose of this study was to determine adult learners preferred at-home laboratory experience in an online, undergraduate science course for non-majors in order to make informed instructional design decisions that are both student-centered and cost-effective.

The study also explored the applicability of Brooke's (1996) System Usability Scale (SUS), a normed and validated user experience tool, in the context of perceived usability evaluation of laboratory experiments; and, which (if any) of Quesenbery's (2001) 5E's of perceived usability (effectiveness, efficiency, engagement, error tolerance, and/or ease of use) and Bandura's (1997) self-efficacy played a significant role in adult learners' user experiences. This chapter will provide an in-depth discussion of the study's research design and methodology.

Research Design and Rationale

Lindsay et al. (2009) posited that a students' first interaction with equipment builds a mental model of the laboratory experience and establishes the reality of laboratory experimentation. Future experiences are built upon that initial mental model which may positively or negatively influence students' perceptions of subsequent laboratory experiences

(Lindsay et al., 2009). Consequently, the student perception data may not be reliable if all groups evaluated the hands-on and virtual experiments in the same order. While there is disagreement in the literature regarding whether or not order matters (Crandall et al., 2015; Gnesdilow & Putambekar, 2022; Lindsay et al., 2009; Meintzer et al., 2017; Pyatt & Sims, 2012; Toth et al., 2009), this effect was taken into consideration as it could represent a serious threat to validity

To control for potential order-bias, this study utilized a quasi-experimental crossover design similar to that of Pyatt and Sims (2012). An experimental crossover design is a two-trial laboratory investigation in which participants in the control group and experimental group switch (or crossover to the other group) for the second trial (Redelmeier & Tibshirani, 1997). In lieu of control/experimental groups, participants in this study were randomly assigned to either Group A or Group B by the learning management system. As there was no designated control group (or pre-testing), this study did not meet the strict definition of an experimental study (Pajo, 2017). However, it did aim to establish a cause-and-effect relationship of sorts (i.e. if perceived usability effects students' preferred laboratory experience); therefore, it did meet the definition of a quasi-experimental study (Pajo, 2017).

In the first trial of the crossover phase, participants in Group A performed an AHHO experiment while those in Group B performed a virtual version of the same experiment; participants performed the opposite versions for Laboratory 2 in the second trial. Students were asked to complete a survey after each of these laboratory experiences. The surveys for trial-1 and trial-2 included a section for: (a) Brooke's (1996) SUS questionnaire (10 questions, 5-point Likert scale), the questions were modified to replace the term "system" with "laboratory experiment"; (b) open-ended reflection questions, slightly modified from those used by Reck et al., 2019; and (c) demographic data. The standard version of the SUS and the complete list of survey questions for Lab 1 and Lab 2 are located in Appendix B.

The rationale for utilizing a quasi-experimental crossover design for this study included: (a) this design allowed for a comparison of student perception data for each type of at-home laboratory experience (hands-on and virtual) for each trial and between trials; (b) each participant experienced both a hands-on and virtual laboratory experiment; (c) each group experienced the hands-on and virtual experiments in a different order (controls for order-bias); and (d) the laboratory protocols tested in each trial were identical, except that the hands-on laboratory experience utilized physical manipulatives while the virtual laboratory experience utilized virtual simulators (controls for content-bias). The laboratory protocols for trial-1 (*Probability and Statistics*; hands-on used real coins, virtual used a coin-flip simulator) and trial 2 *(Measurements and Significant Figures*; hands-on used a real ruler, virtual used an online ruler) are located in Appendix C.

Following the crossover phase, this study included a self-selection phase. In this phase, the students had the opportunity to self-select between an AHHO experience and a VL experience. While the hands-on and virtual laboratory protocols satisfied the same learning outcomes (and were both health-related in content), the procedures and pre- and post-lab questions were not the same for each laboratory. As this phase was not as tightly controlled as the crossover phase, the SUS questions were omitted from this survey. The final survey consisted of demographic and open-ended questions asking the participant to reflect on what factors lead to their laboratory selection. While the survey questions for the crossover phase mainly focused on the factors measured by the SUS (perceived usability, effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy), the open-ended questions in this phase were not only used to seek complementarity with crossover results, but also to allow for the inclusion of other factors that may have played a role in students' perception of their laboratory experience. The laboratory protocols for Laboratory 4 are located in Appendix D; the Lab 4 survey questions are located in Appendix E.

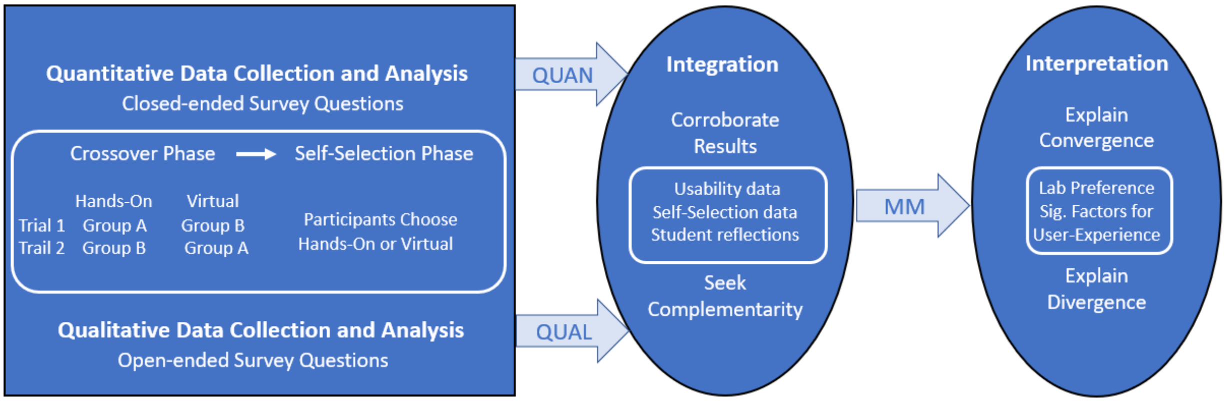
As the surveys contained both quantitative and qualitative data (and the data were collected concurrently), this study also met the definition of a convergent mixed methods design (Creswell, 2022). Mixed methods research is a procedure for collecting, analyzing, and integrating (or mixing) both quantitative and qualitative data (Creswell & Creswell, 2018; Creswell & Plano Clark, 2011; Pajo, 2017); a convergent mixed methods study occurs when the quantitative data and qualitative data are collected and analyzed in the same phase of the research study (Almeida, 2018; Creswell & Plano Clark, 2011; Schoonenboom et al., 2018).

Fetters and Freshwater (2015) used the equation "1+1 = 3" (p.116) to illustrate the rationale for mixed methods research; adding the strengths of the qualitative and quantitative research approaches leads to a depth of understanding that is greater than the sum of its parts. The integration of the data strands provided the opportunity to corroborate results of the SUS, seek complementarity, and identify areas of dissonance between the student perception and self-selection data sets (Almalki, 2016; Bryman, 2007; Creswell & Plano Clark, 2011; Fetters, 2020; McKim, 2017). This enabled the investigator to gain deeper insights than could be gleaned from either quantitative or qualitative methodologies alone (Schoonenboom et al., 2018).

Although the convergent mixed methods design is an efficient way to capture data from a student population (as it is difficult to remain in contact with students after they have completed the course), this design is also prone to certain challenges (Creswell, 2022). Some of the largest obstacles include difficulty merging the data strands, overcoming sample size issues, and determining how to deal with discrepancies between the quantitative and qualitative responses (Creswell, 2022). While each of these posed a significant challenge, the benefits of the convergent mixed methods design far outweighed these obstacles. Figure 3.1 provides a visual representation of this quasi-experimental crossover, convergent mixed methods study.

Figure 3.1

Quasi-Experimental Crossover, Convergent Mixed Methods Design



Note. The figure was adapted from Creswell (2022) to illustrate the convergent mixed methods design utilized for in this study. The first box shows the instruments used for data collection in both the crossover phase and the self-selection phase. Following data analysis, the two ovals demonstrate that the quantitative and qualitative results were merged and findings were then interpreted.

Methodology

Participants, Sampling Plan, and Data Sources

The target population for this research study was adult learners enrolled in online, undergraduate, science courses for non-majors. Data was collected from a convenience sample of students registered for a 12-week online, undergraduate, science course for non-majors (SCIE 211: Introduction to Scientific Analysis and Reasoning) at a private, professionally-focused university in the Midwestern United States. The research site offered a total of 11 online sections of SCIE 211 during the Fall 2022 term: 5 sections in the U-term (with a start date in August) and 6 sections in the Q-term (with a start date in September). The staggered start dates allowed for replication of the study within the same term. All students enrolled (n_U=122; n_Q=148; n_{total}=270) were given the opportunity to participate.

Students were asked to voluntarily complete a total of three anonymous surveys during the course of the term; one after the completion of each lab in the crossover phase (Lab 1 and Lab 2) and another after the completion of the lab in the self-selection phase (Lab 4). The links for each of the surveys (developed in Microsoft Forms) were distributed as part of the instructional materials for its corresponding laboratory exercise. The first question on each survey asked respondents to identify which version of the laboratory that they completed. This was followed by perceived usability questions from the SUS (for Lab 1 and Lab 2 only), open-ended reflection questions, and demographic data.

Operationalization of Variables

To operationalize adult learners preferred at-home laboratory experience, the researcher chose the construct ‘perceived usability’ from the field of software engineering (International Organization for Standardization, 2018). Perceived usability measures the quality of the product by the quality of the user experience (Iwarsson & Ståhl, 2003; Shackel, 2009). The current definition is found in ISO 9241-11:2018 section 3.1.1:

> “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use”

ISO 9241-11:2018 specifically defines the variables “effectiveness” (3.1.12), “efficiency (3.1.13), and “satisfaction” (3.1.14) as follows: (a) effectiveness is the extent to which the product can be used by specified users to achieve specified goals; (b) efficiency is the extent to which specified users can complete assigned tasks within the allotted timeframe; and (c) satisfaction is the extent to which the product meets the user’s needs and expectations (International Organization for Standardization, 2018). The definitions for these variables are consistent with how they are used in educational contexts (Pal & Vanijja, 2020; Vlachogianni & Tselios, 2021, 2022).

Dependent Variables

Quesenbery (2001) proposed that the ISO 9241-11:2018 definition for the variable “satisfaction” was too broad. To gain a deeper understanding of how the product meets a user’s needs and expectations, Quesenbery (2001) broke this variable down into engagement, error tolerance, and ease of use. Furthermore, Lewis and Sauro (2019) interpreted (and benchmarked) the individual items on the SUS. For example, Item 9 relates to ‘confidence in ability to use’ the product; in educational terms, this item would be associated with self-efficacy, or the user’s belief in their capability to perform specified tasks (Bandura, 1997). Using Quesenbery’s (2001) 5E’s of perceived usability and Bandura’s (1997) self-efficacy, the dependent variables in this study were: effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy. Perceived usability is a composite score calculated from these variables (Sauro & Lewis, 2016).

In terms of students’ perception of the laboratory experience, the dependent variables are defined as follows: ‘effectiveness’ is the extent to which specified users are able to achieve specified laboratory goals; ‘efficiency’ is the extent to which specified users are able to complete the laboratory procedures within the allotted timeframe; ‘engagement’ is the extent to which the laboratory protocol was able to pique the interest of specified users and/or specified users enjoyed the experience; 'error tolerance' is the extent to which the specified users are able to self-correct/recover from errors without having to start over; 'ease of use' is the extent to which specified users are able to understand and follow the laboratory protocol without intervention/support (Iwarsson & Ståhl, 2003; Quesenbery, 2001; Shackel, 2009); and 'self-efficacy' is the extent to which the specified users' believe in their ability to perform the tasks required to complete the laboratory experiment (Bandura, 1997).

Quesenbery's (2001) 5E's of perceived usability closely align with adult learners' expectations for educational activities (Table 3.1). Adults learners tend to be independent, self-directed, self-motivated students that value relevant, real-world learning activities that are directly tied to learning outcomes (Arghode et al., 2017; Calabrese & Capraro, 2021; Chen, 2014; Knowles, 1980, 1984). Furthermore, adult learners do not appreciate "busy work" as this is viewed as a waste of their valuable time (Osam et al., 2016).

Table 3.1

Alignment of Variables: Perceived Usability with Adult Learner Characteristics/Values

Perceived Usability	Definition	Adult Learner
Effectiveness	*ability of specified users to achieve specified laboratory goals*	Activity directly tied to learning outcomes
Efficiency	*ability of specified users to complete laboratory procedures in allotted time*	Time not wasted
Engagement	*ability to pique interest of users; users find the experience enjoyable*	Relevant, real-world experiences
Error tolerance	*ability of specified users to self-correct/recover from mistakes*	Self-directed Independent
Ease of Use	*ability of specified users to understand and follow the laboratory protocol without intervention/support*	Self-directed Self-motivated Independent

Independent Variable

The independent variable identified in this study is ‘preferred laboratory experience’. Preferred laboratory experience will be evaluated from both perceived usability data and self-selection data. These data were cross-referenced to determine if the preferred mode of laboratory instruction based upon user experience data was consistent with the preferred mode of laboratory instruction identified by self-selection data. Furthermore, the dual streams of data were collected, analyzed, and corroborated in an effort to confirm the applicability of the SUS in the context of perceived usability evaluation of laboratory experiments.

Controls for Confounding Variables

In addition to controlling for order-bias with the crossover design, this study also attempted to control for the confounding variables ‘laboratory setting’ and ‘laboratory materials’. Nearly all of the previous studies compared the traditional, hands-on laboratory experience, that was performed in an actual science laboratory with easy access to an instructor and included the social experience of working with other students, with a non-traditional laboratory experience, that was performed at home, alone, with limited access to instructors and peers (Brinson, 2015; Faulconer & Gruss, 2018; Ma & Nickerson, 2006; Tsihouridis et al., 2018). While previous studies revealed that students largely preferred hands-on laboratory experiences (Johnson & Barr, 2021; Kelley, 2020; Meintzer et al., 2017; Stuckey-Mickell & Study-Danner, 2007; Tsihouridis et al., 2018), it is difficult to discern if students preferred the ability to physically manipulate objects or if they preferred the ‘laboratory setting’ with easy access to an instructor and the social aspects of working with peers.

Furthermore, the laboratory protocols for the traditional and non-traditional laboratory experiences were not always the same in previous studies (Faulconer & Gruss, 2018; Kelley, 2020; Stuckey-Mickell & Study-Danner, 2007). While the laboratories may have tested the same general concepts, the hands-on laboratory utilized different equipment, different procedures, different background materials, and different pre- and post-lab questions than the alternative laboratory (Attardi et al., 2015; Koretsky et al., 2011; Meintzer et al., 2017; Reece & Butler, 2017; Stuckey-Mickell & Study-Danner, 2007). Again, this makes it difficult to discern if students' preference for 'hands-on' experiences was due to the ability to physically manipulate objects or because the hands-on 'laboratory materials' were more interesting, better aligned with learning outcomes, or easier to complete.

For this study, it was assumed that the at-home laboratory settings were the same for both the hands-on and virtual experiments. However, the design of this study cannot guarantee that this was the case. The laboratories were completed in any space the student chose for their "at-home" location. This was a fair assumption given that students' "at-home" laboratory settings were not likely to include access to a brick-and-mortar science laboratory where they performed the experiment with lab partners under the direct supervision of an instructor.

Furthermore, the laboratory materials (background materials, procedures, goals/objectives and pre-/post-lab questions) were identical for each set of laboratories in the crossover phase of this study. The only difference between the laboratories was that the hands-on experiment utilized real objects while the virtual experiment utilized virtual objects. As such, it was assumed that student perception data would reflect their preference for manipulating physical equipment versus virtual equipment. Again, the design of this study cannot guarantee that this was the case. Students' lack of access to the physical equipment necessary to complete the hands-on experiment or technical issues with the simulators for the virtual experiment may impact their perceptions of the experience. However, this was a fair assumption given that the only physical equipment needed were coins, a ruler, and a balloon (common household items) and that the students were registered in an online course and are required to have the technical capabilities to operate these simple simulators, per university policy.

Instrumentation and Measurements

The instruments used in this study included three sets of SCIE 211 laboratory protocols (at-home hands-on and at-home virtual versions for Lab 1, Lab 2, and Lab 4) and three anonymous surveys (one designated for each of the laboratory experiments). The following sections will provide a description of each of the instruments utilized in this study.

Laboratory Protocols: Crossover Phase

Laboratory protocols for trial-1 and trial-2 are located in Appendix C. The laboratory background materials, procedures, goals/objectives and pre-/post-lab questions were identical for each set of laboratories, the only difference was that the hands-on laboratory utilized physical manipulatives while the virtual laboratory utilized virtual manipulatives. Table 3.2 provides a brief summary of the purpose of each laboratory and the key differences between the hands-on and virtual versions of the experiments.

Table 3.2

Summary of Laboratory 1 and Laboratory 2

Laboratory	Purpose	Hands-On	Virtual
Laboratory 1: Probability and Statistics	Explore the effect of sample size on the magnitude of the percent deviation from expected. Classic coin-flip experiment.	Real coins	Coin-Flip simulator
Laboratory 2: Measurements and Significant Figures	Explore how to accurately record measurements and report calculated values (area, volume) with the correct number of significant figures. Classic measurement experiment.	Real Ruler Real Objects	Virtual Ruler Virtual Objects

Laboratory Protocols: Self-Selection Phase

In this phase of the study, students were given the opportunity to self-select between an AHHO and VL laboratory experience. While the hands-on and virtual laboratory protocols were both health-related in content and satisfied the same purpose (explore correlation and causation), the procedures and pre- and post-lab questions were not the same for each laboratory.

In the hands-on version of Laboratory 4, students used balloons to measure their tidal volume (volume of air moved with a normal respiratory cycle) and vital capacity (maximum volume of air moved following deep inhalation and forced exhalation). Students then compared their respiratory volumes to the standard values, identified any confounding factors that may have affected their measured values (respiratory illness, smoker, athlete, etc.), and discussed whether these factors correlated with abnormal respiratory volumes. Finally, they were asked to determine if there was enough evidence to establish causation (i.e. that the identified factors/conditions were the cause of the differences of their respiratory volumes from standard values).

In the virtual version of Laboratory 4, students were provided with a graph depicting a positive correlation between organic food sales and the prevalence of autism. They were asked to perform background research on ‘organic food’, ‘autism’, and to identify any confounding or spurious factors that may have contributed to this relationship. The students were then asked to outline the steps of an experiment that could be performed to test this relationship and discuss what types of evidence would be needed to determine if increased sales of organic food caused the increased prevalence of autism.

Students’ choice for their final laboratory experience was used as one of the means to evaluate students’ preferred laboratory experience. The proportion of students that self-selected the hands-on lab was compared with the proportion of students that self-selected the virtual lab. This self-selection data was cross-referenced with the perceived usability data to determine if the preferred mode of laboratory instruction based upon user experience data was consistent with the preferred mode of laboratory instruction identified by self-selection data. If these data arrived at the same conclusion, this would serve as evidence that the SUS is an appropriate tool to measure perceived usability of a laboratory experiment.

Anonymous Surveys

Students were asked to voluntarily complete a total of three anonymous surveys during the course of the term. The surveys for trial-1 and trial-2 of the crossover phase included a section for: (a) identification of experiment (hands-on or virtual); (b) modified version of Brooke's (1996) SUS questionnaire; (c) open-ended reflection questions modified from Reck et al., 2019; and (d) demographic data. The survey for self-selection phase was similar to that of the crossover phase, except that it did not include the SUS questionnaire.

Identification of Hands-On or Virtual Experiment. The first question for each survey asked the respondents to identify which version of the laboratory they completed. This was accomplished by a simple description of the materials used to complete the laboratory. For example, the first question on the trial-1 survey was: "For Lab 1, I completed the experiment that (a) utilized real coins; or (b) utilized a coin-flip simulator". The response to this question was used to identify which perception data was "hands-on" and "virtual"; the mean perception data was compared for each trial, between trials, and overall.

Modified Version of Brooke's (1996) SUS questionnaire. The SUS is one of the most frequently used instruments in usability studies (Bevan, 1995; Brooke, 2013; Lewis, 2018a; Lacerda & von Wangenheim, 2018). The popularity of the SUS can be attributed to several factors, including: (a) it is free to use (with acknowledgement to Brooke, Appendix F); (b) it is readily available to the public; (c) it is simple to use; (d) it can be completed in just a few minutes; (e) it has been normed, validated, and shown to be very reliable; (f) minor modifications to the questions does not affect validity or reliability; and (g) the results have been shown to be reliable with a sample size as small as 8 participants (Bangor et al., 2008; Bevan, 1995; Brooke, 1996, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016; Shackel, 2009).

While the SUS is a quick and easy questionnaire for participants to use, calculating the perceived usability score is somewhat complicated due to the alternating positive/negative tone of the questions (Bevan, 1995; Brooke, 1996; Brooke, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016). Brooke (1996) purposefully designed the instrument with the alternating positive and negative statements to avoid response bias. However, this means that the raw scores had to be adjusted when calculating the composite SUS score and when aggregating positive/negative item responses for like-variables (Lewis & Sauro, 2019; Sauro & Lewis, 2016).

To norm the instrument, Sauro and Lewis (2016) developed a curved grading scale in which the median SUS score (68, based upon n=241 responses) was assigned a letter grade of C (designated as an average user experience). Moreover, the highest and lowest percentiles were assigned letter grades of A and F, respectively; and percentile ranges were further broken down to assign the full set of ‘grades’ on the plus/minus scale. Table 3.3 illustrates the complete curved grading scale, with the range of SUS scores and their corresponding percentile range (Sauro & Lewis, 2016).

Table 3.3

Curved Grading Scale for the SUS

Grade	SUS	Percentile range
A+	84.1 – 100	96 – 100
A	80.8 – 84	90 – 95
A-	78.9 – 80.7	85 – 89
B+	77.2 – 78.8	80 – 84
B	74.1 – 77.1	70 – 79
B-	72.6 – 74.0	65 – 69
C+	71.1 – 72.5	60 – 64
C	65 – 71.0	41 – 59
C-	62.7 – 64.9	35 – 40
D	51.7 – 62.6	15 – 34
F	0 – 51.6	0 – 14

Note: The information in Table 3.3 was adapted from Sauro & Lewis, 2016.

While norming the SUS score gives an individual score meaning (to evaluate if a product is good or poor), this requires further manipulation of raw item scores (Brooke, 2013; Sauro & Lewis, 2016). The following steps outline how to calculate the normed SUS score (Sauro & Lewis, 2016):

i. Assign all non-response items a raw score of 3 (neutral, on a scale in which 1=strongly disagree and 5=strongly agree)

ii. Subtract 1 from the raw score for each odd-numbered item (positive tone);

iii. Subtract the raw score of each even-numbered item (negative tone) from 5;

iv. Add all of the adjusted raw item scores together (yields max. sum of 40)

v. Multiply this sum by 2.5 (yields max. score of 100)

The SUS scores (= perceived usability) were calculated for each lab for each participant in both trial-1 and trial-2 of the crossover phase of the study. The mean perceived usability score for the hands-on experiment was compared to the mean perceived usability score for the virtual experiment in each trial, between trials, and overall. Additionally, the mean perception data for each of the subfactors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy) were compared in each trial, between trials, and overall.

Table 3.4.

Summary of Item Codes and Benchmarks for SUS=68 and SUS=80

	Question	Code	Mean Raw Item Score	
			SUS=68	SUS=80
1	I think I would like to do more lab experiments like this one.	Engagement	≥ 3.39	≥ 3.80
2	I found this lab experiment unnecessarily complex.	Ease of Use	≤ 2.44	≤ 1.85
3	I thought this lab experiment was easy to do.	Ease of Use	≥ 3.67	≥ 4.24
4	I think I would need more support to be able to do more lab experiments like this one.	Error Tolerance	≤ 1.85	≤ 1.51
5	I found the various parts of this lab experiment to be well integrated.	Effectiveness	≥ 3.55	≥ 3.96
6	I thought there was too much inconsistency in this lab experiment.	Effectiveness	≤ 2.20	≤ 1.77
7	I would imagine that most people would be able to do lab experiments like this one very quickly.	Efficiency	≥ 3.71	≥ 4.19
8	I found this lab experiment very awkward to do.	Engagement	≤ 2.25	≤ 1.66
9	I felt very confident doing this lab experiment.	Self-efficacy	≥ 3.72	≥ 4.25
10	I needed to learn a lot of things before I could do this lab experiment.	Efficiency	≤ 2.09	≤ 1.64

Note. Benchmark values are based upon Lewis and Sauro (2019). SUS=68 is defined as average user experience; SUS=80 is defined as above average user experience.

Lewis and Sauro (2019) benchmarked the mean raw item scores based upon a SUS score of 68 (average user experience) and 80 (above average user experience). Table 3.4 shows the alignment of each item on the SUS with Quesenbery's (2001) 5E's of perceived usability (effectiveness, efficiency, engagement, error tolerance, and/or ease of use) and Bandura's (1997) self-efficacy and their benchmark values as determined by Lewis and Sauro (2019). Each factor corresponded to both a positive statement and negative statement, with the exception of error tolerance and self-efficacy.

Open-ended Reflection Questions. Open-ended reflection questions for the crossover phase were used (with permission, Appendix G) from Reck et al., 2019; the open-ended reflection questions for the self-selection phase were self-developed based upon an extensive review of the literature (Attardi et al., 2015; Folmer & Bosch, 2004; Greene & Larsen, 2018; Meintzer et al., 2017; Nolen & Koretsky, 2018; Pal & Vanijja, 2020; Reck et al., 2019; Stuckey-Mickell & Stuckey-Danner, 2007). The reflection question responses were used to corroborate results of the SUS, seek complementarity, and identify areas of dissonance between the student perception and self-selection data sets.

Demographic Data. Data collected regarding gender, race/ethnicity, and age of participants was compared to the University's student profile dashboard to determine if the study sample is representative of this student population. The survey choices for gender and race/ethnicity were selected to be compatible with University data. Table 3.5 depicts student demographic data for the research site (internal communication, University Dashboard). Additionally, participants were asked to identify their marital status, student status (full-time, >12 credit hours; or part-time, <12 credit hours), employment status (full-time, part-time, or not employed), and whether or not they have dependents to determine if the sample population is consistent with the definition of an adult learner (NCES, n.d.; Osam et al., 2017).

Table 3.5

Student Profile Data

Demographic Data								
Gender	Female 55.78%	Male 41.51%	Other 2.7%					
Race/ Ethnicity	American Indian	Asian	Black/ African American	Hawaiian/ Pacific Islander	Hispanic	White	2+	Unknown
	0.56%	2.96%	18.24%	0%	2.54%	56.78%	2.99%	16.19%
Age	<25 16.2%	25-29 22.3%	30-34 19.1%	35-39 14.7%	40-44 10.7%	45-49 7.5%	50-54 4.8%	55+ 3.1%

Note. The percentages were generated from admission data and reported in the University Dashboard (internal communication, July 2022).

Furthermore, a small negative correlation between SUS rating and participants' age has been identified in previous studies (Bangor et al., 2008; Granić & Ćukušić, 2011; Harrati et al., 2016; Orfanou et al., 2015; Vlachogianni & Tselios, 2021). To determine if this same correlation exists in this study population, participants' ages were collected in 5-year increments (<25, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55+) rather than simply <25 or 25+. Instead of complicating the study, this reported relationship may be used to corroborate that perceived usability and the 5E's of perceived usability (effectiveness, engagement, efficiency, error tolerance, and ease of use) and self-efficacy are important metrics for adult learners.

Reliability and Validity

This study leveraged instruments that have been utilized (at least in part) in previous terms of SCIE 211 (laboratory protocols) or in in previous studies (SUS and open-ended questions used in

Reck et al., 2019). The three hands-on laboratory experiments have been reviewed and approved by subject matter experts (content validity) and have performed well with students (as indicated by positive student feedback and consistent score distributions) over the last two academic years (2020-2021, 2021-2022). The virtual laboratory experiments for trial-1 and trial-2 are identical to the hands-on experiments, except for the utilization of virtual tools (a coin-flip simulator for lab 1; an online ruler for lab 2). The virtual laboratory experiment in the self-selection phase was derived from laboratory protocol that was utilized in SCIE 211 prior to academic year 2020-2021. As such, the virtual laboratories were expected to demonstrate the same level of reliability as their hands-on counterparts.

As previously mentioned, the SUS is one of the most frequently used instruments in usability studies (Bevan, 1995; Brooke, 2013; Lewis, 2018a; Lacerda & von Wangenheim, 2018) and has been used in more than 13000 studies (J. Brooke, personal communication, April 27, 2022). The popularity of the SUS can be attributed to several factors, including: (a) it is free to use and readily available to the public; (b) it has been normed, validated, and has been shown to be very reliable (coefficient alpha ≈ 0.90); (c) minor modifications to the questions does not affect validity or reliability; (d) it is sensitive to a wide variety of independent variables; and (e) results have been shown to be reliable with small sample sizes (Bangor et al., 2008; Bevan, 1995; Brooke, 1996; Brooke, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016; Shackel, 2009).

The open-ended survey questions were modified from a previous study (crossover phase) or self-developed (self-selection phase) based upon an extensive review of the literature (Attardi et al., 2015; Folmer & Bosch, 2004; Greene & Larsen, 2018; Meintzer et al., 2017; Nolen & Koretsky, 2018; Pal & Vanijja, 2020; Reck et al., 2019; Stuckey-Mickell & Stuckey-Danner, 2007; and others). The reflection question responses were cross-referenced with SUS results to seek complementarity and to identify areas of dissonance between the student perception and self-selection data sets. These efforts to corroborate results between data streams (and the replication of the study) increased the trustworthiness of the previously untested reflection questions.

While the instruments utilized in this study are solid, there were several potential threats to validity. First, there is the issue of volunteer bias. All students enrolled in SCIE 211 (n =270) were given the opportunity to participate. As data was only analyzed for those students that consented to participate in the study, the perceptions of the volunteers may not be representative of the target population. Often, those that volunteer to participate in studies have attitudes or opinions on the extreme ends of the spectrum (Creswell, 2022).

Instructor contamination was another potential threat. For ethical reasons, the researcher did not teach the course during data collection; the 11-sections of SCIE 211 were taught by seven (7) different adjunct faculty: three of the instructors taught a section in both the U- and Q-terms, one instructor taught two sections in the Q-term, two instructors taught one section in the U-term, and one instructor taught one section in the Q-term. The instructors' attitudes and/or behaviors may have influenced student participation and students' perception of the experience. For example, if an instructor graded one type of laboratory report more harshly than another, it may have impacted students' perception of the experience and/or influenced student self-selection for the last laboratory experience.

Finally, this study utilized a modified version of the SUS. In order to align the SUS items with language that reflects the "system" being tested is a laboratory experiment, the questions were slightly altered. Some questions were easily adapted to fit the context of a laboratory experiment. For example, question 2 on the SUS, “I found the system unnecessarily complex” (Brooke, 1996) was easily modified to, “I found this laboratory experiment unnecessarily complex”. However, question 1 on the SUS, “I think I would like to use this system frequently” (Brooke, 1996) was not as easily modified to, “I think I would like to do more lab experiments like this one”. Although the reliability and validity of the SUS is not affected by slight modifications (Bangor et al., 2008; Bevan, 1995; Brooke, 1996; Brooke, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016; Shackel, 2009), these alternations may be considered to be more than ‘slight’ modifications. Cronbach's alpha was calculated for both the Lab 1 Survey and Lab 2 Survey to determine if the modified SUS was as reliable as the classic SUS.

Data Collection Plan

At the beginning of the term, all students registered for SCIE 211 received recruitment materials via a course announcement and a short video that was played during the Week 1 Meet Session. All students (whether they consented to participate in the study or not) were randomly assigned to either Group A or Group B by the learning management system. In the first trial, participants in Group A performed an AHHO experiment while those in Group B performed a virtual version of the same experiment; participants performed the opposite versions for Laboratory 2 in the second trial. Students were given the opportunity to complete a survey to reflect on their experience at the end of each laboratory. Following this crossover phase, all students were asked to self-select between an at-home, hands-on laboratory and an at-home, virtual laboratory. Again, students were given the opportunity to complete a survey to reflect upon why they chose to complete the lab they selected. The links for each of the surveys (developed in Microsoft Forms) were distributed as part of the instructional materials for its corresponding laboratory exercise.

Data was collected for one semester. U-term sections ran for 12-weeks beginning 8/15/2022; Q-term sections ran for 12-weeks beginning 9/26/2022. Data was collected in both Q- and U-terms for replication.

Data Analysis Plan

Both the quantitative and qualitative strands of this study were focused on students' perception of their experience with at least one AHHO experiment and at least one VL experiment. Students were asked to voluntarily complete a total of three anonymous surveys during the course of the term; survey questions included: (a) a modified version of Brooke's (1996) SUS questionnaire (for Lab 1 and Lab 2 only); (b) open-ended reflection questions; and (c) demographic data.

Survey Results – Quantitative Strand

The survey results were downloaded as Microsoft Excel worksheets. The SUS score was calculated for each participant for each lab in the crossover phase (hands-on Lab 1, virtual Lab 1, hands-on Lab 2, and virtual Lab 2) using the equation functions in Excel. Prior to this calculation, data was cleaned by assigning a score of 3 (neutral option on a 5-point Likert-scale) for all non-response SUS items (Sauro & Lewis, 2016).

Mean SUS Scores. Descriptive statistics were generated by and all statistical tests were performed in SAS OnDemand for Academics. The mean SUS score (and standard deviation) was calculated the following data sets: hands-on Lab 1; virtual Lab 1; hands-on Lab 2; virtual Lab 2; aggregated hands on (Lab 1 and Lab 2); and aggregated virtual (Lab 1 and Lab 2). The mean SUS score for each data set was assigned a letter grade as described by Sauro and Lewis (2016).

Furthermore, the mean SUS scores were compared: hands-on Lab 1/virtual Lab 1; hands-on lab 2/virtual lab 2; hands-on Lab 1/hands-on Lab2; virtual Lab 1/virtual Lab 2; and aggregate hands-on/aggregate virtual. Two-sample, independent t-tests were used to determine if any of the differences in means is significant. Finally, a comparison of the mean SUS scores (=perceived usability) for the aggregate hands-on and aggregate virtual was used to determine participants' preferred mode of laboratory instruction, based on student perception.

Mean Raw Item Scores. The mean raw values for each item on the SUS calculated for the following data sets: hands-on Lab 1; virtual Lab 1; hands-on Lab 2; virtual Lab 2; aggregated hands on (Lab 1 and Lab 2); and aggregated virtual (Lab 1 and Lab 2). The mean raw item score for each data set was compared to benchmark data (Lewis & Sauro, 2019) to rank the quality of the user experience.

Each of the variables (with the exception of error tolerance and self-efficacy) were assigned to a positive and negative statement. In order to aggregate these data for comparison analysis, the adjusted raw item scores were calculated according to Sauro and Lewis (2016) and combined for like-variables. The mean adjusted raw item scores for each variable were compared (using the one-way AVOVA test) to determine if the differences in means were significant.

Self-Selection Data. The proportion of students that self-selected the hand-on and virtual versions of Lab 4 was calculated in Microsoft Excel. A comparison of the proportion of students that self-selected hands-on and virtual; a two-proportion Z-test was used to determine if the proportions are statistically the same or different. Moreover, the proportions were used to determine participants' preferred mode of laboratory instruction, based on self-selection. This self-selection data was cross-referenced with perceived usability data to determine if the preferred mode of laboratory instruction based upon user experience data was consistent with the preferred mode of laboratory instruction identified by self-selection data. If these data arrived at the same conclusion, this would serve as evidence that the SUS is an appropriate tool to measure perceived usability of a laboratory experiment.

Qualitative Strand. Responses for the open-ended reflection questions were initially coded using hypothesis coding. The pre-determined set of codes (Creswell & Creswell, 2018) utilized were in alignment with the dependent variables (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy). In the second round of coding, the researcher looked for emergent codes to determine if respondents identified other factors that impacted their laboratory experience.

Ethical Considerations

IRB approval (#IRB-2022-47) was obtained prior to data collection. While the study was designed to mitigate ethical issues, there were some items worth noting. First, the primary investigator for this study is the lead faculty for the SCIE 211 course. Lead faculty at the research site are responsible for developing course content, monitoring academic content for relevance and rigor, performing teaching observations, staffing courses, and supervising adjunct faculty. Although all adjunct faculty readily agreed to facilitate this study, they may have done so out of a sense of obligation.

Another issue was associated with the research design. All students (whether they consented to participate in the study or not) were randomly assigned to either Group A or Group B by the learning management system. In the first trial, all students in Group A performed an AHHO experiment while those in Group B performed a virtual version of the same lab; all students performed the opposite versions for Laboratory 2 in the second trial. Following this crossover phase, all students had the opportunity to self-select between an AHHO and VL laboratory experience.

While instructors typically provide students with their assigned laboratory materials (via email and a course announcement), all students in the same section typically complete the same laboratory exercises. As previously discussed, the laboratory materials for the hands-on and virtual experiments in the crossover phase were nearly identical (by design). Therefore, the small differences should not have created a significant imposition or barrier to successful completion of these required assignments.

Finally, all students registered in SCIE 211 were given the opportunity to participate in the study. The links for the surveys were distributed as part of the instructional sheet for its corresponding laboratory exercise. As the web-based surveys were completely anonymous, it was not required to collect written consent documents (as this would violate the anonymity of participants). The informed consent language was located on the initial landing page of the survey. Participants were advised that proceeding to the next page (to begin the survey) implied consent. However, students may not have fully appreciated the significance of this statement.

Chapter Summary

This chapter discussed the research design and methodology for this study. The quasi-experimental crossover phase of the study utilized a modified version of Brooke’s (1996) System Usability Scale (SUS) to evaluate if adult learners in an online, undergraduate science course for non-majors at a private, professionally-focused university in the Midwestern United States prefer hands-on or virtual laboratory experiences. The results of the SUS were corroborated by the self-selection phase and the responses to the open-ended survey questions. The next chapter will provide an extensive discussion of the Lab 1, Lab 2, and Lab 4 survey results and analysis.

Chapter 4 – Data Collection and Analysis

This convergent mixed methods comparison study utilized Brooke's (1996) System Usability Scale (SUS) as a framework to evaluate the preferred at-home, laboratory experience for adult learners in an online, undergraduate science course for non-majors at a private, professionally-focused university in the Midwestern United States. Students were asked to voluntarily submit a total of three anonymous surveys; one after the completion of each lab in the crossover phase (Lab 1 and Lab 2) and another after the completion of the lab in the self-selection phase (Lab 4). The surveys for trial-1 and trial-2 of the crossover phase included a section for: (a) identification of laboratory experience (hands-on or virtual); (b) modified SUS questionnaire; (c) open-ended reflection questions modified from Reck et al. (2019); and (d) demographic data. The survey for self-selection phase was similar to that of the crossover phase, except that it did not include the modified SUS questionnaire. Data were collected for the U-term (with a start date in August) and the Q-term (with a start date in September) during Fall 2022. The staggered start dates allowed for replication of the study within the same trimester.

This chapter will discuss the findings from the statistical analysis of the quantitative data and the themes that emerged from the analysis of the qualitative data. These data were merged to corroborate the results of the SUS, seek complementarity, and identify areas of dissonance between the crossover and self-selection datasets. The sections that follow include participant demographics, a review of the research questions and hypotheses, and the data analysis from the quantitative and qualitative strands. The chapter will conclude with the merged findings and how they were used to determine which laboratory experience participants preferred and which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) played a significant role in respondents' user experiences.

Participant Data

Data was collected from a convenience sample of students registered for a 12-week online, undergraduate, science course for non-majors (SCIE 211: Introduction to Scientific Analysis and Reasoning) at a private, professionally-focused university in the Midwestern United States. The research site offered a total of 11 online sections of SCIE 211 during the Fall 2022 term: 5 sections in the U-term (with a start date in August) and 6 sections in the Q-term (with a start date in September). The staggered start dates allowed for replication of the study within the same trimester. All students registered in SCIE 211 (n_U=121; n_Q=149; n_{total}=270) were given the opportunity to complete three anonymous laboratory surveys.

To foster participants' trust in the anonymity of their responses, participants in all 5 U-term sections and all 6 Q-term sections utilized the same link for the Lab 1 survey, the same link for the Lab 2 survey, and the same link for the Lab 4 survey. Furthermore, participants were not asked to identify their class section or their instructor. This limited the researcher's ability to conduct complex group analysis for the various participant sub-populations (i.e. by individual class section or by instructor). However, respondents could be identified as a U-term or Q-term student by the date of survey completion. U-term/Q-term identification assumed that the date of the Lab X Survey completion (where X=1, 2, or 4) occurred within ± 20 days of the due date of its corresponding laboratory assignment. As such, the U-term, Q-term, and Overall (or combined U-term/Q-term) participant data for each survey was analyzed (when appropriate).

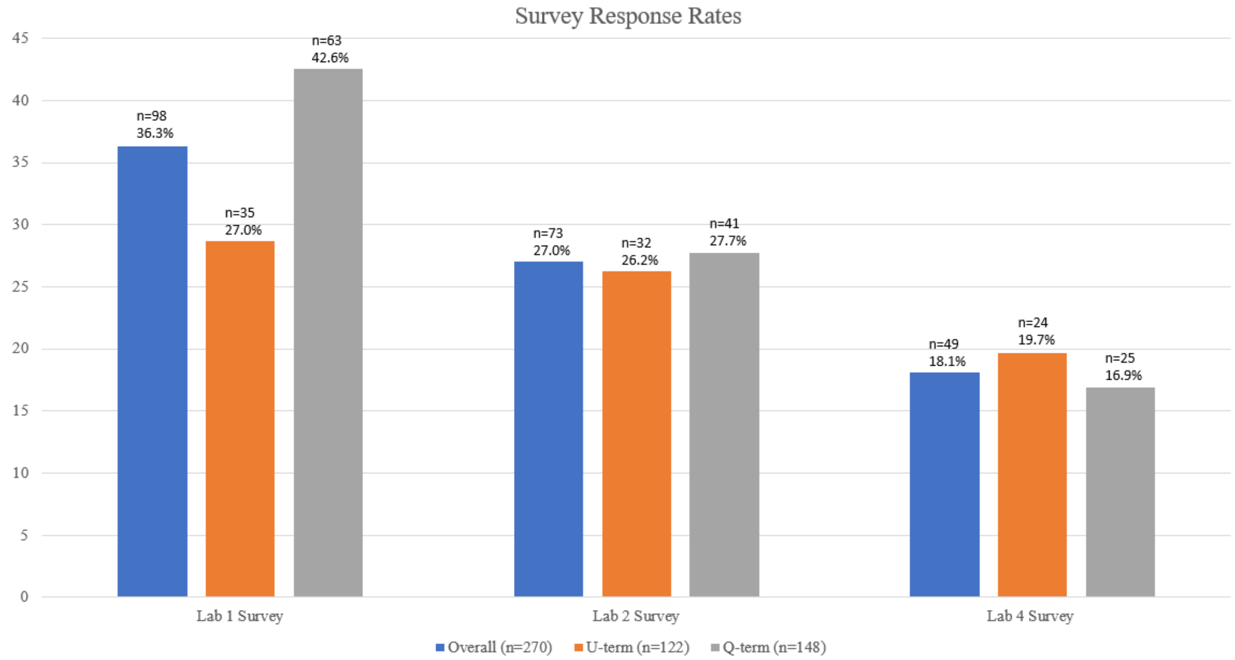
Response Rates

A total of 98 students (n_{U1}=35; n_{Q1}=63) completed the Lab 1 survey with an overall response rate (RR) of 36.3% (RR_{U1}=28.7%; RR_{Q1} =42.6%); 73 students (n_{U2}=32; n_{Q2}=41) completed the Lab 2 survey ($RR_{overall}$=27.0%; RR_{U2}=26.2%; RR_{Q2} =27.7%); and 49 students

(n_{U4}=24; n_{Q4}=25) completed the Lab 4 survey ($RR_{overall}$=18.1%; RR_{U4}=19.7%; RR_{Q4} =16.9%). Figure 4.1 shows that the U-term, Q-term, and Overall RR decreased with each survey ($RR_{Lab\ 1} > RR_{Lab\ 2} > RR_{Lab\ 4}$). This was not surprising as attrition is common in research studies that involve the completion of two or more surveys by the same group of participants at different points in time (de Leeuw & Lugtig, 2015).

Figure 4.1

Response Rates for Anonymous Lab Surveys: Overall, U-term, and Q-term



Note. The bar graph illustrates the Overall, U-term, and Q-term response rates (RR) for each of the anonymous lab surveys; the graph shows that $RR_{Lab\ 1} > RR_{Lab\ 2} > RR_{Lab\ 4}$ for the Overall, U-term, and Q-term data.

Figure 4.1 also shows what appears to be a significant difference between RR_U and RR_Q for the Lab 1 survey. A Chi-squared goodness-of-fit test was performed to determine if the response rates for the U- and Q-terms were within expectations. The test statistic was calculated using 36.3% ($RR_{overall\ Lab\ 1}$) to determine the expected number (E_i) of U-term and Q-term responses for Lab Survey 1. Although the test statistic was large ($\chi^2 = 3.549$), it was less than the critical value (3.841, $\alpha = 0.05$, df =1) so we failed to reject the null hypothesis (H_0: $\pi_i = \pi_{i0}$; H_a: at least one of the proportions differs from the hypothesized value). This means that the difference between the observed RR_{U1} and RR_{Q1} and their expected values was not significant. The Chi-squared analysis yielded the same results for the Lab 2 and Lab 4 surveys (failed to reject H_0 because $\chi^2_{\text{Lab2 (Ei based on RR=27\%)}}=0.054$, $\chi^2_{\text{Lab 4 (Ei based on RR=18.1\%)}}=0.285 < 3.841$, $\alpha = 0.05$, df =1); therefore, all observed RR_U and RR_Q were consistent with expected values. This indicates that the assumptions (cut-off dates) used to categorize respondents as U-term or Q-term were good. Furthermore, any mis-labeled respondents should not be considered a significant source of error.

Demographic Data

Demographic data was collected for all three lab surveys. The survey choices for gender and race/ethnicity were selected to be compatible with the data available on the University's student profile dashboard. Participants' ages were collected in 5- year increments (<25, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55+). Table 4.1 and Figure 4.2 show the comparisons of gender, race/ethnicity, and age data for each lab survey with the University's student profile data.

A Chi-squared goodness-of-fit test was performed for each of the lab surveys to determine if the sample population was representative of the University population. The test indicated that the sample population was consistent with the University gender and race/ethnicity profiles (with the exception of race/ethnicity for the Lab 1 Survey). However, the age-data was not consistent with the University student profiles. The proportion of adult learners (25+ years old) in the sample population ($85.7\%_{\text{Lab 1}}$, $84.9\%_{\text{Lab 2}}$, and $89.8\%_{\text{Lab 4}}$) was greater than the student population as a whole ($83.8\%_{\text{university}}$).

It should be noted that the Chi-squared analysis was only performed on the Overall dataset. The combination of small percentages (< 5% for several of the race/ethnicity and age categories) and smaller U-term and Q-term sample sizes decreased the accuracy of the approximation. Cochran (1954) determined that the results of the Chi-squared goodness-of-fit test are not valid if more than 20% of the expected values (E_is) are less than 5. As more than 20% of the U-term and Q-term E_is for race/ethnicity and age were less than 5, it was not appropriate to report U-term and Q-term results.

Table 4.1

Demographic Data: Gender, Race/Ethnicity, and Age

(a) Gender (in %)

Gender	Female	Male	Other
University	55.78	41.51	2.70
Lab 1	56.12	41.84	2.04
Lab 2	58.90	39.73	1.37
Lab 4	57.14	42.86	0

(b) Race/Ethnicity (in %)

Race/ Ethnicity	American Indian	Asian	Black/ African American	Hawaiian/ Pacific Islander	Hispanic	White	2+	Unknown
University	0.56	2.96	18.24	0	2.54	56.78	2.99	16.19
Lab 1*	1.02	2.04	14.29	0	9.18	59.18	7.14	7.14
Lab 2	1.37	0	16.44	0	6.85	57.53	5.48	12.34
Lab 4	2.04	0	14.29	0	4.08	63.26	8.16	8.16

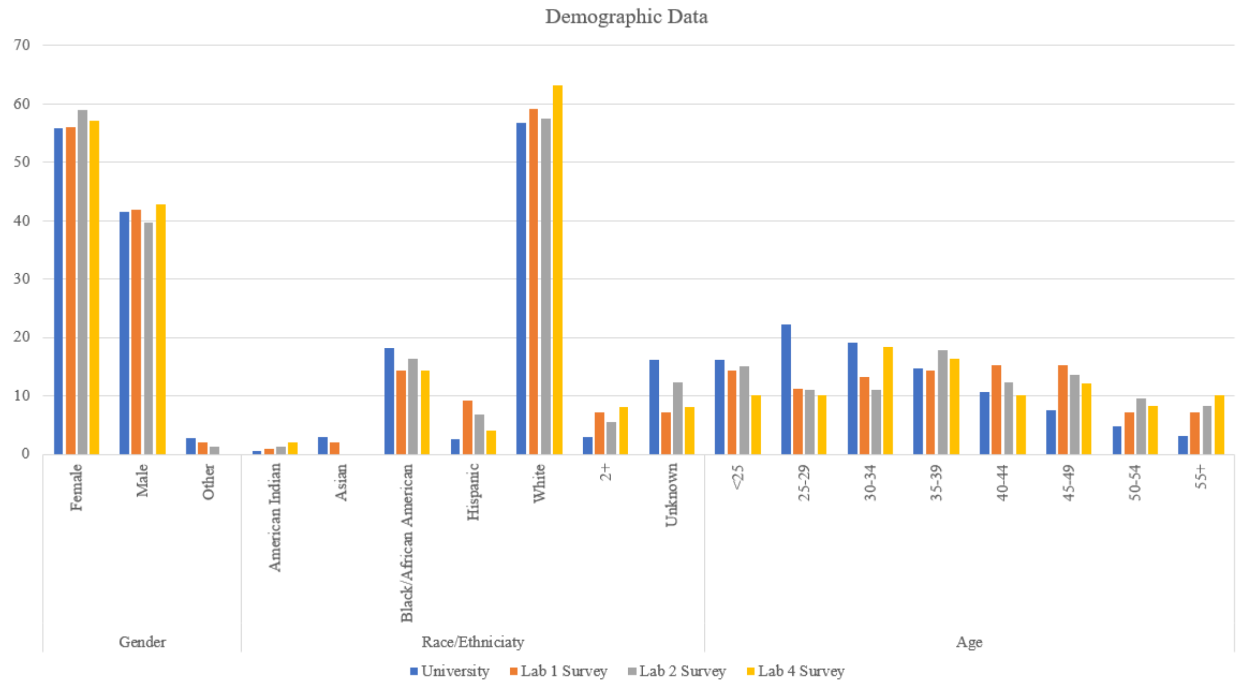
(c) Age (in %)

Age	<25	25-29	30-34	35-39	40-44	45-49	50-54	55+
University	16.2	22.3	19.1	14.7	10.7	7.5%	4.8	3.1
Lab 1*	14.3	11.2	13.3	14.3	15.3	15.3%	7.1	7.1
Lab 2*	15.1	11.0	11.0	17.8	12.3	13.7%	9.6	8.2
Lab 4*	10.2	10.2	18.4	16.3	10.2	12.2%	8.2	10.2

Note. Based upon the Chi-squared goodness-of-fit ($\chi^2_{\text{test statistic}} > 14.07$, $\alpha=0.05$, 7df), the sample populations indicated with (*) were not representative of the University population.

Figure 4.2

Demographic Data: Gender, Race/Ethnicity, and Age

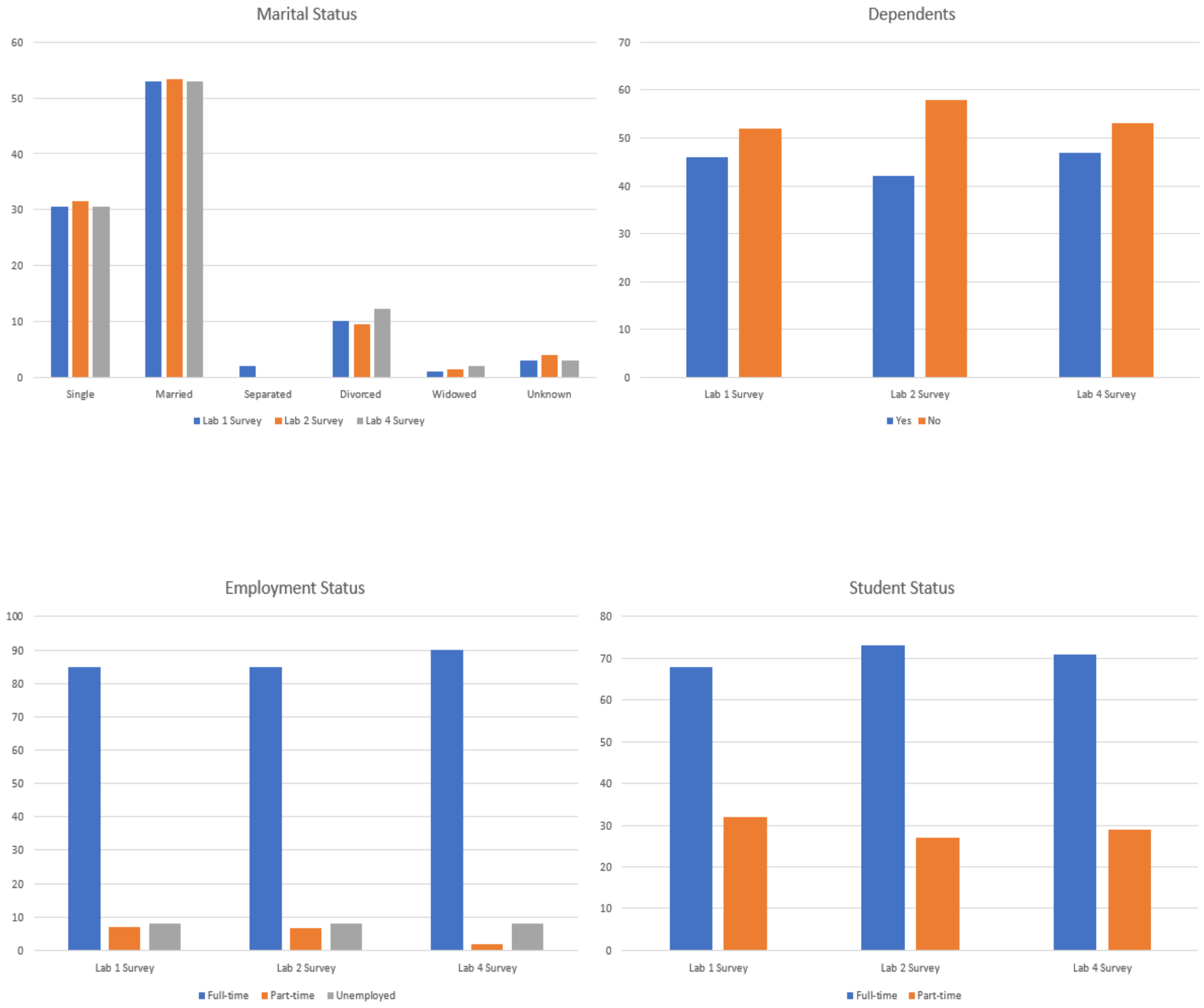


Note. The bar graph shows the comparison of the gender, race/ethnicity, and age data (in percentages) for each lab survey with the University's student profile data.

Participants were also asked to identify their marital status, student status, employment status, and whether or not they have dependents to determine if the sample population is consistent with other characteristics of adult learners. Adults learners are students over the age of 25 that have one or more of the following characteristics: attend school part-time, work full-time, have dependents other than a spouse, and are single parents (NCES, n.d.; Osam et al., 2016). The bar charts in Figure 4.3 show that approximately 70% of the sample population is currently married or was married; 45% have dependents; 70% are full-time students; and 86% work 36+ hours per week (i.e. large majority of respondents meet the definition of an adult learner).

Figure 4.3

Participant Demographics: Marital Status, Dependents, Employment Status, and Student Status



Note. The bar charts show the participants' marital status (in %), dependent status (% yes = have children under the age of 18; % no = do not have children under the age of 18), employment status (% full-time = work 36+ hours per week; % part-time = work < 36 hours per week; or not employed, and student status (% full-time = >12 credit hours; % part-time = <12 credit hours). The majority of the sample population meets the definition of an adult learner: approximately 86% are working full-time while attending school, 70% are (or were) married, and 45% have dependents.

Review of Research Questions and Hypotheses

Adult learning theories emphasize the importance of active, hands-on learning for this student population (Chen, 2013; Knowles, 1980, 1984; Lewis & Bryan, 2021). However, adult learners' perceptions are largely missing from the available literature. To determine if adult learners preferred at-home, hands-on laboratory activities or at-home, virtual laboratory activities (and why), the overarching mixed-methods research questions for this study were:

(i) Will adult learners in an online, undergraduate, science course for non-majors at a private, professionally-focused university in the Midwest prefer at-home, hands-on laboratory experiences or at-home, virtual laboratory experiences?

(ii) Which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) play a significant role in participants' user experiences?

This mixed-methods study occurred in two phases. In Phase I (quasi-experimental cross-over phase), students completed the Lab 1 and Lab 2 Surveys after performing an at-home hands-on (AHHO) and after performing an at-home virtual experience (VL). In Phase II (self-selection phase), students were given the opportunity to self-select between AHHO or VL experience for their last laboratory assignment; participants then completed the Lab 4 survey in which they identified which laboratory experience they selected and why. Table 4.2a summarizes the research questions and hypotheses for the quantitative strand; Table 4.2b summarizes the research questions for the qualitative strand.

Table 4.2a

Research Questions and Hypotheses for Quantitative Strand

Quantitative Research Questions (RQ)			
RQ1: Will adult learners in an online, undergraduate, science course for non-majors at a private, professionally-focused university in the Midwestern United States prefer at-home, hands-on laboratory experiences or at-home, virtual laboratory experiences?			
Hypotheses	Test	Preference based on:	Source(s)
Perceived Usability: Phase I $H_{01\text{-}1}: \mu_{AHHO} - \mu_{VL} = 0$ $H_{1\text{-}1}: \mu_{AHHO} - \mu_{VL} \neq 0$	T test (α=0.05)	Perceived usability (SUS)	Lab 1 Survey Lab 2 Survey Q 1-2
Sub-factors: Phase I $H_{01\text{-}1'}$: means are equal $H_{1\text{-}1'}$: at least one of the means is different	One-way ANOVA (α=0.05)	Effectiveness (5,6) Efficiency (7,10) Engagement (1,8) Error tolerance (4) Ease of use (2,3) Self-efficacy (9)	Lab 1 Survey Lab 2 Survey Q 1-2
Self-selection: Phase II $H_{01\text{-}2}: \hat{p}_{AHHO} = \hat{p}_{VL}$ $H_{1\text{-}2}: \hat{p}_{AHHO} \neq \hat{p}_{VL}$	Two proportion Z-test (α=0.05)	Proportions of self-selected experiences	Lab 4 Survey Q1
RQ2: Is there evidence to support that one or more of the factors (effectiveness, efficiency, engagement, error tolerance, ease of use, self-efficacy) plays a significant role in this population of students' user experience?			
Hypotheses	Test	Experience	Source(s)
Sub-factors: Phase I $H_{02a}: \beta_i = 0$ $H_{2a}: \beta_i \neq 0$	Multiple regression (α=0.05)	AHHO	Lab 1 Survey Lab 2 Survey Q 1-2
Sub-factors: Phase I $H_{02b}: \beta_i = 0$ $H_{2b}: \beta_i \neq 0$	Multiple regression (α=0.05)	VL	Lab 1 Survey Lab 2 Survey Q 1-2

Note. μ_{AHHO} = mean value for at-home, hands-on experience (AHHO); μ_{VL} = mean value for at-home, virtual experience (VL); Q = Question; ANOVA = analysis of variance; (number) = item on modified SUS questionnaire (Brooke, 1996); $\hat{p}_{AHHO}$ = proportion of students that self-selected the AHHO experience; $\hat{p}_{VL}$ = proportion of students that self-selected the VL experience; β_i = slope for sub-factor$_i$ in the multiple regression analysis.

Table 4.2b

Research Questions for Qualitative Strand

Qualitative Research Questions (RQ)	Source(s)	Questions
RQ3: Which at-home laboratory experience do participants say they prefer (hands-on or virtual)? Why do they say they prefer this laboratory experience?	Lab 4 Survey	1-3
RQ4: Which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, self-efficacy, or other) played an important role in students' user experiences?	Lab 1/Lab 2 Survey Lab 4 Survey	1, 3-9 1, 4-6
RQ5: Do the same factors identified by user experience responses play a role in the students' self-selection process?	Lab 1/Lab 2 Survey Lab 4 Survey	1, 3-9 1-6

Quantitative Findings

In Phase I (quasi-experimental crossover phase), the first two questions on the Lab 1 Survey (trial-1) and Lab 2 Survey (trial-2) were used to gather data on perceived usability (SUS score) and its sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy) for the AHHO and VL laboratory experiences. During Phase II (self-selection phase), the first question on the Lab 4 Survey was used to calculate the proportion of respondents that chose to complete AHHO and VL experiences. These data were compared to determine students preferred at-home laboratory experience.

The results of the Phase I analysis (Mean Raw Item Scores, SUS Scores, and Mean Adjusted Raw Item Scores for Sub-factor analysis) are presented in the following sections. The section will begin with a brief description of the crossover trials and the primary data sources (question 1 and 2 on the Lab Surveys).

Phase I: Quasi-experimental Crossover Phase

Crossover Experiences: Trial-1 and Trial-2

A total of 270 students completed the trial-1laboratory experience (*Probability and Statistics*). Students were randomly assigned into Group A (n_U=62; n_Q=78; $n_{overall}$=140) and Group B (n_U=59; n_Q=71; $n_{overall}$=130) by the learning management system. The laboratory protocols were identical, except that students in Group A utilized real coins (AHHO) and students in Group B utilized a coin-flip simulator (VL). All students were given the opportunity to complete the Lab 1 Survey; 98 students (n_{U1}=35; n_{Q1}=63) responded.

Likewise, a total of 270 students completed the trial-2 laboratory experience (*Measurements and Significant Figures*). Students were randomly assigned into Group A (n_U=62; n_Q=78; $n_{overall}$=140) and Group B (n_U=59; n_Q=71; $n_{overall}$=130). The laboratory protocols were identical, except that students in Group A utilized a virtual metric ruler (VL) and students in Group B utilized a real metric ruler (AHHO). All students were given the opportunity to complete the Lab 2 Survey; 73 students (n_{U2}=34; n_{Q2}=39) responded. Table 4.3 provides a summary of the crossover laboratory experiences.

Table 4.3

Summary of the Crossover Phase

	Lab Protocol	Group A	Group B	Respondents (n)
Trial-1	Probability and Statistics	AHHO	VL	n=98 total U-term (n=35) Q-term (n=63)
Trial-2	Measurements and Significant Figures	VL	AHHO	n=73 total U-term (n=34) Q-term (n=39)

Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience. The study was run in both the U- and Q-terms in Fall 2022; the number of respondents for the Lab 1 and Lab 2 Surveys (U-, Q-, and total) are reported in the table.

Identification of the Laboratory Experience

The first question on the Lab 1 Survey and on the Lab 2 Survey asked participants to identify which version of the laboratory they completed. This was accomplished by a simple description of the materials used for the laboratory. For example, the first question on the trial-1 survey was: “For Lab 1, I completed the experiment that (a) utilized real coins; or (b) utilized a coin-flip simulator”. Their response to this question was used to identify which perception data should be assigned to the AHHO (utilized physical manipulatives) and VL (utilized virtual manipulatives).

SUS Questionnaire

The second question on the Lab 1 survey and on the Lab 2 survey was a modified version of Brooke’s (1996) SUS questionnaire. The SUS questionnaire has been used in more than 13000 studies (J. Brooke, personal communication, April 27, 2022). The popularity of the SUS can be attributed to several factors, including: (a) it is free to use and readily available to the public; (b) it has been normed, validated, and has proven to be a reliable instrument; (c) minor modifications to the items does not affect the reliability of the scale; (d) it is sensitive to a wide variety of independent variables; and (e) results have been shown to be valid, even with small sample sizes (Bangor et al., 2008; Brooke, 1996; Brooke, 2013; Lewis, 2018a; Lewis, 2018b; Lewis & Sauro, 2019; Sauro & Lewis, 2016). Moreover, the SUS has proven to be a particularly effective instrument when comparing the user-experiences for two versions of the same ‘application’ (Brooke, 2013).

Finally, the SUS questionnaire is simple to use and can be completed in just a few minutes. It consists of 10 items with alternating positive- and negative-toned statements (Brooke, 1996). Each statement is rated on a 5-point Likert scale where *5*=strongly agree, *4*=somewhat agree, *3*=neutral, *2*=somewhat disagree, and *1*=strongly disagree. Table 4.4 lists the items for Brooke's (1996) SUS questionnaire and the modified items utilized for the Lab 1 and Lab 2 Surveys.

Table 4.4

SUS Questionnaire: Brooke's (1996) classic SUS and the Modified SUS Items for this Study

Item	Brooke's (1996) SUS	Modified SUS
1	I think that I would like to use this system frequently.	I think I would like to do more lab experiments like this one.
2	I found the system unnecessarily complex.	I found this lab experiment unnecessarily complex.
3	I thought the system was easy to use.	I thought this lab experiment was easy to do.
4	I think that I would need the support of a technical person to be able to use this system.	I think I would need more support to be able to do more lab experiments like this one.
5	I found the various functions in this system were well integrated.	I found the various parts of this lab experiment to be well integrated.
6	I thought there was too much inconsistency in this system.	I thought there was too much inconsistency in this lab experiment.
7	I would imagine that most people would learn to use this system very quickly.	I would imagine that most people would be able to do lab experiments like this one very quickly.
8	I found the system very awkward to use.	I found this lab experiment very awkward to do.
9	I felt very confident using the system.	I felt very confident doing this lab.
10	I needed to learn a lot of things before I could get going with this system.	I needed to learn a lot of things before I could do this lab experiment.

Note. SUS = system usability scale. The items on Brooke's (1996) SUS questionnaire were modified to replace the term 'system' with 'lab experiment'. Some items were easily adapted to fit the context of a laboratory experiment (e.g. items 2 and 8); others required more extensive modification (e.g. items 1 and 7).

To determine if the item modifications affected the internal consistency of the SUS, Cronbach's alpha coefficient (α) was calculated for the Lab 1 Survey and Lab 2 Survey using SAS OnDemand for Academics. The results of this analysis ($\alpha_{\text{Lab 1 Survey}} = 0.88$; $\alpha_{\text{Lab 2 Survey}} = 0.87$) indicated that the item modifications did not adversely affect the reliability of the scale; the calculated values fell within the range of values (α=0.85-0.91) reported within the literature (Brooke, 1996; Brooke, 2013; Lewis & Sauro, 2019; Orfanou et al., 2015; Sauro & Lewis, 2016). This is also consistent with the assertion that minor modifications to the items does not affect the reliability of the scale (Brooke, 2013; Lewis & Sauro, 2019; Sauro & Lewis, 2016).

SUS Questionnaire: Mean Raw Item Scores. In their study, "Item Benchmarks for the System Usability Scale", Lewis and Sauro (2019) utilized a large database (n=11,855) of completed SUS questionnaires to benchmark the mean raw item scores (MRIS) based upon a SUS score of 68 (=average user experience) and 80 (=above average user experience). Furthermore, they identified specific attributes for each of the SUS items. For example, item 2 "I found the system unnecessarily complex" was associated with 'perceived complexity'; item 3 "I thought the system was easy to use" was associated with 'perceived ease of use'; and, item 9 "I felt very confident using the system" was associated with 'confidence in use'. Based upon their descriptions, each of the SUS items were coded as a measure of effectiveness, efficiency, engagement, error tolerance, ease of use, or self-efficacy. The MRIS (and the standard deviation) were calculated for each item for each trial using SAS OnDemand for Academics. The MRISs were used to determine which items/variables met the SUS=68 (average user experience) and SUS=80 (above average user experience) benchmarks.

Mean Raw Item Scores: Summary Statistics – Trial 1. Table 4.5 shows the MRIS results for trial-1.

Table 4.5

Mean Raw Item Score Analysis – Trial 1

Lab 1	SUS=68 SUS=80	U-term (n=35; 33 analyzed)		Q-term (n=63; 62 analyzed)		Overall (n=98; 95 analyzed)	
SUS item	MRIS	Mean $\frac{A}{B}$	SD $\frac{A}{B}$	Mean $\frac{A}{B}$	SD $\frac{A}{B}$	Mean $\frac{A}{B}$	SD $\frac{A}{B}$
1 (N)	≥ 3.39	3.294	±1.532	**3.583**	±1.297	**3.491**	±1.368
	≥ 3.80	**4.167***	±0.786	**3.889***	±1.013	**4.000***	±0.929
2 (EU)	≤ 2.44	2.662	±1.409	**2.417**	±1.228	2.566	±1.294
	≤ 1.85	**1.889**	±0.758	**2.111**	±1.340	**2.022**	±1.138
3 (EU)	≥ 3.67	3.624	±1.334	**4.111**	±0.950	**4.019**	±1.083
	≥ 4.24	**4.167**	±0.985	**4.444***	±0.698	**4.333***	±0.826
4 (ET)	≤ 1.85	2.411	±1.278	2.306	±1.369	2.340	±1.329
	≤ 1.51	1.944	±0.938	2.333	±1.330	2.178	±1.193
5 (E)	≥ 3.55	**3.765**	±1.200	**3.722**	±1.137	**3.736**	±1.146
	≥ 3.96	**3.944**	±1.162	**4.000***	±0.832	**3.978***	±0.965
6 (E)	≤ 2.20	2.235	±1.393	**1.944**	±1.040	**2.038**	±1.160
	≤1.77	**1.889**	±1.132	**1.778**	±1.121	**1.822**	±1.114
7 (F)	≥ 3.71	3.353	±1.272	3.528	±1.106	3.472	±1.154
	≥ 4.19	**4.167**	±0.985	3.370	±1.043	3.689	±1.083
8 (N)	≤ 2.25	2.353	±1.579	2.306	±1.346	2.321	±1.411
	≤ 1.66	**1.667**	±0.840	**1.778**	±1.155	**1.733**	±1.031
9 (SE)	≥ 3.72	**4.000**	±1.225	**3.722**	±1.323	**3.811**	±1.287
	≥ 4.25	**4.056**	±0.996	**3.963**	±1.091	**4.000**	±1.044
10 (F)	≤ 2.09	2.236	±1.582	2.167	±1.159	2.189	±1.287
	≤ 1.64	**2.000**	±0.907	**1.889**	±1.013	**1.933**	±0.963

Note. MRIS = mean raw item score; SD = standard deviation; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; $\frac{A}{B} = \frac{Group\ A}{Group\ B}$; Group A = at-home, hands-on laboratory experience (AHHO); Group B = virtual laboratory experience (VL); **bold** = met benchmark for average user experience; * met benchmark for above average user experience. There were 2 U-term responses and 1 Q-term response that were removed from the dataset as all items were given the same, non-neutral rating.

The findings show that nearly all of the MRISs for Group B (VL_1) were greater than Group A ($AHHO_1$) for the positive statements (odd-numbered items) and that nearly all of the MRISs for Group B (VL_1) were less than Group A ($AHHO_1$) for the negative statements (even-numbered items). Participants in Group B indicated an above average user experience for item 1 (engagement; U-term, Q-term, and Overall), item 3 (ease of use; Q-term and Overall), and item 5 (effectiveness; Q-term and Overall); they indicated an average user experience for item 2 (ease of use; U-term, Q-term, and Overall), item 6 (effectiveness; U-term, Q-term, Overall), item 7 (efficiency; U-term), item 8 (engagement; U-term, Q-term, and Overall), item 9 (self-efficacy; U-term, Q-term, and Overall), and item 10 (efficiency; U-term, Q-term, and Overall). None of the participants in Group A indicated they had an above average user experience; however, they did indicate an average user experience for item 1 (engagement; Q-term and Overall), item 2 (ease of use; Q-term), item 3 (ease of use; Q-term and Overall), item 5 (effectiveness; U-term, Q-term, and Overall), item 6 (effectiveness; Q-term and Overall), and item 9 (self-efficacy; U-term, Q-term, and Overall).

Item 4 (error tolerance) did not meet the SUS=68 or SUS=80 benchmark for either group. It should be noted that the standard deviation for this item (and items 2, 6, 8, and 10) was relatively high in that the deviation was ≥ 50% of the mean. As these are all even-numbered items, the negative-tone of these statements may have impacted participants' responses. Saldivar et al. (2019) found that the wording of questions (positive versus negative tone) affected the strength of agreement or disagreement in satisfaction surveys.

However, Brooke (1996) purposefully designed the instrument with alternating positive and negative statements to avoid response bias, or the condition in which respondents provide inaccurate or false answers in a survey (Pajo, 2017). The researcher carefully reviewed the ratings for positive-negative response pairs as part of the data cleaning plan. Pairs with incongruent ratings were removed from the dataset. For example, a rating of 4 (=somewhat agree) for item 2 (“I found this lab experiment to be unnecessarily complex”) is not consistent with a rating of 4 for item 3 (“I thought this lab experiment was easy to do”). There were 2 U-term respondents (≈6% of n_{U1}) and 1 Q-term respondent (≈1.5% of n_{Q1}) removed from the dataset as all items were given the same, non-neutral rating.

Mean Raw Item Scores: Summary Statistics – Trial 2. Table 4.6 shows the MRIS results for trial-2. The findings for Lab 2 were not as positive as for Lab 1. The results from the Lab 1 Survey indicated that participants in both groups (VL_1 and $AHHO_1$) had a good user experience (with the overall user experience for $VL_1 > AHHO_1$). However, the Lab 2 Survey results largely indicated that the participants had a below average user experience. Participants in Group A (VL_2) only indicated an average user experience for item 5 (effectiveness, U-term, Q-term, and Overall); participants in Group B ($AHHO_2$) indicated an average user experience for item 2 (ease of use; Q-term only), item 5 (effectiveness, Q-term and Overall), item 6 (effectiveness, Q-term only), and item 10 (efficiency; Q-term only). There were no items (in either group) that met the SUS=80 benchmark (above average user experience). Furthermore, the standard deviation for nearly all of the items was relatively high (≥40% of the mean).

While the Group A/Group B comparison results were clear for Lab 1 ($VL_1 > AHHO_1$ for all items), the comparison results were mixed for Lab 2. The user experience for the $AHHO_2$ was greater than VL_2 for items 1 (engagement), 2 (ease of use), 5 (effectiveness), 6 (effectiveness), 7 (efficiency), 8 (engagement), and 9 (self-efficacy); on the other hand, the user experience for the VL_2 was greater than $AHHO_2$ for items 3 (ease of use), 4 (error tolerance), and 10 (efficiency).

Table 4.6

Mean Raw Item Score Analysis – Trial 2

Lab 2	SUS=68 SUS=80	U-term (n=34)		Q-term (n=39)		Overall (n=73)	
SUS item	MRIS	Mean $\frac{A}{B}$	SD $\frac{A}{B}$	Mean $\frac{A}{B}$	SD $\frac{A}{B}$	Mean $\frac{A}{B}$	SD $\frac{A}{B}$
1 (N)	≥ 3.39	3.000	±1.472	2.952	±1.596	2.971	±1.527
	≥ 3.80	2.789	±1.032	3.300	±1.129	3.051	±1.099
2 (EU)	≤ 2.44	2.538	±1.198	2.905	±1.411	2.765	±1.327
	≤ 1.85	2.789	±1.357	**2.350**	±1.226	2.564	±1.294
3 (EU)	≥ 3.67	3.308	±1.182	3.476	±1.470	3.412	±1.351
	≥ 4.24	3.062	±1.268	3.300	±1.218	3.179	±1.233
4 (ET)	≤ 1.85	2.000	±1.155	2.952	±1.283	2.589	±1.305
	≤ 1.51	3.158	±1.214	2.550	±1.234	2.846	±1.247
5 (E)	≥ 3.55	**3.692**	±0.947	**3.571**	±1.121	**3.617**	±1.045
	≥ 3.96	3.316	±0.946	**4.000**	±0.858	**3.667**	±0.955
6 (E)	≤ 2.20	2.615	±1.193	2.762	±1.336	2.706	±1.268
	≤1.77	2.421	±1.281	**2.100**	±0.852	2.256	±1.069
7 (F)	≥ 3.71	2.923	±1.188	2.905	±1.221	2.912	±1.190
	≥ 4.19	3.211	±1.084	3.150	±1.089	3.179	±1.073
8 (N)	≤ 2.25	3.461	±1.391	3.286	±1.347	3.352	±1.346
	≤ 1.66	2.526	±1.349	2.400	±1.231	2.461	±1.274
9 (SE)	≥ 3.72	3.000	±1.000	3.190	±1.327	3.118	±1.200
	≥ 4.25	2.947	±1.178	3.400	±1.095	3.179	±1.144
10 (F)	≤ 2.09	2.231	±1.166	2.667	±1.155	2.500	±1.161
	≤ 1.64	2.526	±1.073	**1.940**	±0.940	2.564	±0.995

Note. MRIS = mean raw item score; SD = standard deviation; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; $\frac{A}{B} = \frac{Group\ A}{Group\ B}$; Group A = virtual laboratory experience (VL); Group B = at-home, hands-on laboratory experience (AHHO); **bold** = meets SUS = 68 benchmark; * meets SUS=80 benchmark. There were no participants removed from this dataset for giving the same, non-neutral rating to all items.

Moreover, the differences in the Overall MRISs (Group A_2 and Group B_2) for items 1, 2, 5, 9, and 10 were very small. This is pertinent as the magnitude of the standard deviation has a substantial impact on the ability to "see" differences in populations. A large standard deviation (or variance) requires an even larger difference in means to rise above the noise in the dataset (Ott & Longnecker, 2016). The impact of the wide variation in raw item scores (for both Trial-1 and Trial-2) on statistical hypothesis testing is explored in the next two sections.

Mean Raw Item Scores: Comparison of Means – Trial 1 and Trial 2. A two-tailed, two-sample t-test ($\alpha = 0.05$) was conducted to determine if the differences in MRIS for trial-1 (Group A_1 versus Group B_1) and trial-2 (Group A_2 versus Group B_2) were significant (H_0: $MRIS_A - MRIS_B = 0$; H_a: $MRIS_A - MRIS_B \neq 0$). However, the results of the t-test may not be valid as the raw item score sample populations were not normally distributed. The Wilcoxon rank sum test (or Wilcoxon two-sample test), a nonparametric alternative to the two-sample t-test (Ott & Longnecker, 2016), was also performed to determine if the differences in populations were significant (H_0: $\Delta_{A/B}=0$; H_a: $\Delta_{A/B}\neq 0$).

Table 4.7 provides a summary of the MRIS analysis for trial-1 and trial-2 datasets. The results (p-values) for both the two-sample t-test and Wilcoxon rank sum test are displayed in the data table. In most cases, the hypothesis tests yielded the same conclusion. However, in the two cases that the tests yielded different results (item 1 and item 8 in trial-1), the p-value for the Wilcoxon rank sum test was used for the determination of significance. It is interesting that both item 1 and item 8 were coded as a measure of engagement.

Table 4.7

Summary of Trial 1 and Trial 2: Mean Raw Item Scores, User Experiences, and p-values

	Lab Survey 1 Results				Lab Survey 2 Results			
	Group			p-value	Group			p-value
SUS Item	A (AHHO)	B (VL)	UX	T W	A (VL)	B (AHHO)	UX	T W
	MRIS	MRIS			MRIS	MRIS		
1 (N)	**3.941**	**4.000**	VL > AHHO	*0.0368** *0.1077*	2.971	3.051	AHHO > VL	0.7945 0.7947
2 (EU)	2.566	**2.022**	VL > AHHO	0.0309* 0.0316*	2.765	2.564	AHHO > VL	0.5159 0.5216
3 (EU)	**4.091**	**4.333**	VL > AHHO	0.1143 0.1616	3.412	3.179	VL > AHHO	0.4451 0.3614
4 (ET)	2.340	2.178	VL > AHHO	0.5306 0.6297	2.589	2.846	VL > AHHO	0.3913 0.3551
5 (E)	**3.736**	**3.978**	VL > AHHO	0.2661 0.3188	**3.617**	**3.667**	AHHO > VL	0.8348 0.9672
6 (E)	**2.038**	**1.822**	VL > AHHO	0.3529 0.3078	2.706	2.256	AHHO > VL	0.1048 0.1202
7 (F)	3.472	3.689	VL > AHHO	0.3420 0.3962	2.912	3.179	AHHO > VL	0.3156 0.3461
8 (N)	2.231	**1.733**	VL > AHHO	*0.0227** *0.0530***	3.352	2.461	AHHO > VL	0.0049* 0.0057*
9 (SE)	**3.811**	**4.000**	VL > AHHO	0.4329 0.6394	3.118	3.179	AHHO > VL	0.8225 0.9226
10 (F)	2.189	**1.933**	VL > AHHO	0.2650 0.5160	2.500	2.564	VL > AHHO	0.8002 0.7834

Note. MRIS = mean raw item score; UX = user experience; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience; T = two-sample t-test; W = Wilcoxon rank sum test; **bold** = meets SUS = 68 benchmark; *italic* = p-values for T and W did not yield the same conclusion; * $p < 0.05$, reject the null = difference in means is significant; $^{**}p \approx 0.05$ = weak evidence to accept or reject the null = inconclusive.

Based upon the t-test (and corroborated by the Wilcoxon rank sum test), the results of the MRIS analysis for the Lab 1 Survey were not as conclusive as they appeared. Although the user experience for the VL_1 was greater than $AHHO_1$ for all variables, the differences in the MRISs were not significant, except for item 2 (ease of use; $p=0.0316 < 0.05$, therefore we reject H_0). This means that the MRIS analysis for trial-1 did not yield conclusive results for participants' preferred laboratory experience ($VL_1 \approx AHHO_1$). The variables that met the SUS=68 benchmark (for both $AHHO_1$ and VL_1) included effectiveness, engagement, ease of use, and self-efficacy. These variables may have played a role in the positive user experience for Lab 1.

For the Lab 2 Survey, there was only one variable (item 8, engagement) that showed a significant difference between the MRIS values for the $AHHO_2$ and VL_2 ($p < 0.05$, therefore we reject H_0). Again, this means that the MRIS analysis for trial-2 did not yield conclusive results for participants' preferred laboratory experience. As previously noted (and shown Tables 4.5 and 4.6), the standard deviations for many of the items in both trial-1 and trial-2 were quite large (≥40% of the mean). Therefore, it was not surprising that the differences in MRIS values were not significant (i.e. the difference in the means was not larger than the variance).

The only variable that met the SUS=68 benchmark (for both $AHHO_2$ and VL_2) was effectiveness. This implies that participants were able to achieve the specified goals of the laboratory; however, they found it to be difficult and did not enjoy the experience.

It is worth noting that the MRIS for self-efficacy for both VL_2 (3.118) and $AHHO_2$ (3.179) fell well below the SUS=68 benchmark (≥3.72). The objectives for this experiment were to collect and analyze measurement data (length, width, and height of 3 different objects in centimeters); to use the correct number of significant figures for their specific measurement tool (to avoid the fallacy of false precision); and to use their measurement data to investigate the strengths and weaknesses of the three measures of central tendency (mean, median, and mode). This researcher did not anticipate that the use of a metric ruler (real or virtual) or that simple calculations (mean and standard deviation) would negatively impact respondents' belief in their ability to perform the specified tasks required to complete the laboratory experiment.

To better understand this result, a review of the open-ended responses revealed that: 14% of the responses indicated participants did not understand the purpose for (and were frustrated by) measuring the length, width, and height of the same object multiple times (they believed that their measurements would/should be the same for all 3 trials); 31% (of responses for those assigned to $AHHO_2$) did not have access to a metric ruler at home; 29% (of responses for those assigned to VL_2) did not see the relevance of measuring virtual objects; and 30% of the responses indicated that they did not know how to (or had difficulty) reading the ruler (both real and virtual) to the 0.01 place.

However, the most telling statistic was that >70% of the responses indicated that the math (the number of calculations and the calculations themselves) were the main source of frustration (and the primary obstacle) to completion of Lab 2. Studies have demonstrated there is a strong correlation between math self-efficacy and STEM interest/perception (Blotnicky et al., 2018; Rozgonjuk et al., 2020; Simpkins et al., 2006). As such, the negative user experience data may have more to do with participants' perception of math than the actual laboratory experience (manipulation of real of virtual objects).

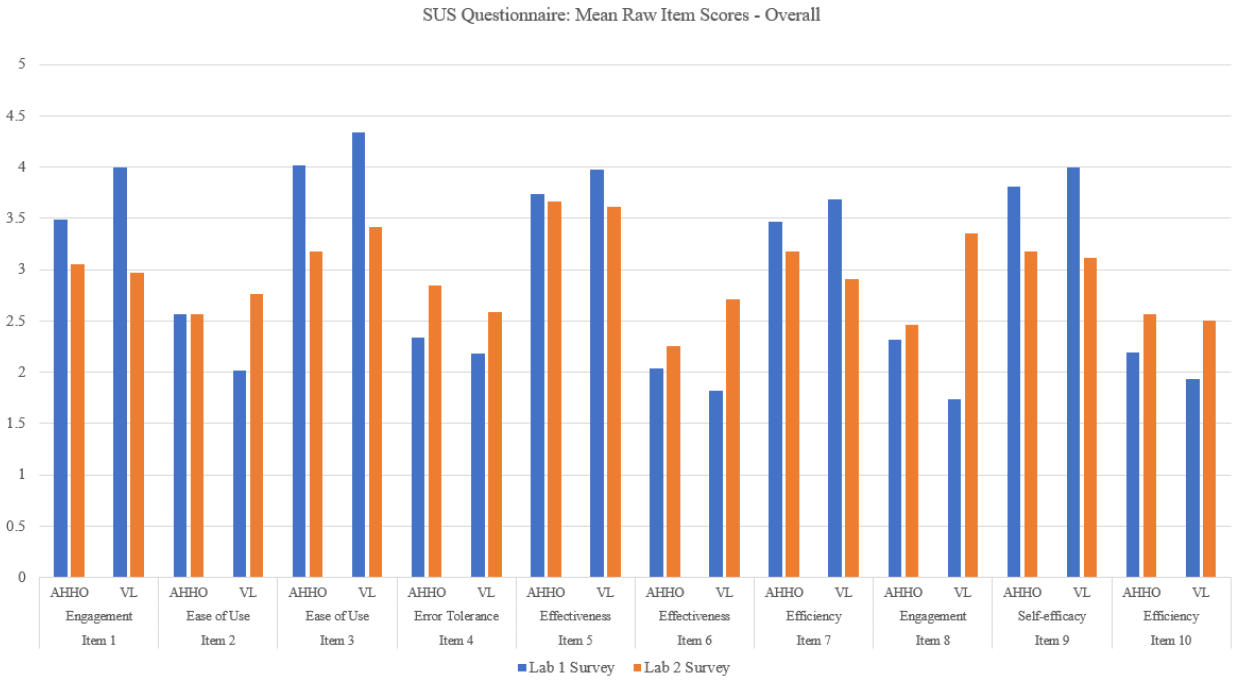
This is supported by Dalgarno et al. (2009) in "Effectiveness of a Virtual Laboratory as a Preparatory Resource for Distance Education Chemistry Students". They hypothesized that a lack of familiarity of the laboratory layout and equipment was the underlying cause of chemistry students' pre-laboratory anxiety. They developed and tested a 3-dimensional simulation that allowed students to explore the laboratory space/equipment prior to the intensive 3-day, on-campus laboratory experience. While the students that utilized the tool found it to be useful, the actual laboratory experience had very little to do with their pre-laboratory anxiety. The interview and questionnaire data revealed that math anxiety was the primary cause of students' stress and frustration with the on-campus laboratory experience.

It is also worth noting that the MRIS for efficiency did not meet the SUS=68 benchmark for either item 7 or item 10 on the Lab 2 Survey. This observation is salient as a large majority of respondents indicated that they are full-time students (70%) and work more than 36 hours/week (86%). Laboratory experiences that take more time (and effort) to complete than anticipated can create additional stress and decrease overall satisfaction (Andrews et al., 2020; Easdon, 2020; Moosvi et al., 2019). A correlation analysis revealed a moderate, positive relationship between 'efficiency' and 'self-efficacy' ($r = 0.5858$). As such, the perceived inefficiency of Lab 2 may have been influenced by participants' lack of self-confidence in measurement skills and mathematical abilities (i.e. participants spent additional time repeating/redoing work because they doubted the accuracy of their measurements and calculations).

Mean Raw Item Scores: Comparison of Means – Trial 1 versus Trial 2. In this section, the MRIS analysis will shift from the results from the individual trials to a comparison of results between trials. Figure 4.4 shows a comparison of the MRISs for the Lab 1 and Lab 2 Surveys. Recall that a large mean value for the odd-numbered items (positive statements) is desirable as it indicates a good user experience; conversely, a small mean value indicates a good user experience for the even-numbered items (negative statements). As such, the bar graph shows that participants in both Group A and Group B had a better user experience for Lab 1 than for Lab 2; participants in Group B_1 (VL) had a better user experience than participants in Group A_1 (AHHO); and, the results for Lab 2 were mixed.

Figure 4.4

Mean Raw Item Scores – Lab 1 Survey versus Lab 2 Survey



Note. The bar graph illustrates the mean raw item score (overall) for each SUS item. The results indicate the participants had a better user-experience for Lab 1 than for Lab 2 (larger mean values for odd-numbered items; smaller mean values for even-numbered items); the results also indicate that participants in Group B_1 (VL) had a better user experience than Group A_1 (AHHO). The results are mixed for Lab 2. The mean values indicate a better user experience for Group A_2 (VL) for items 3, 4, and 10; the mean values indicate a better user experience for Group B_2 (AHHO) for items 1, 2, 5, 6, 7, 8, and 9.

A one-way ANOVA (α = 0.05) was conducted using SAS OnDemand for Academics to determine if the differences between the mean raw items scores for Group A and Group B for each SUS item for each survey were significant (H_0: MRISs are all equal; H_a: at least one of the MRISs is different). However, the results of the ANOVA may not be valid as the raw item scores were not normally distributed. The Kruskal-Wallis test, a nonparametric alternative to the one-way ANOVA (Ott & Longnecker, 2016) was also performed to determine if the differences in populations were significant (H_0: population medians are identical; H_a: at least one of the population medians is different). The results of this analysis are found in Table 4.8.

Based upon the one-way ANOVA test (and corroborated by the Kruskal-Wallis test), there were significant differences in the MRIS values for Group A ($AHHO_1/VL_2$) for item 8 (engagement) and item 9 (self-efficacy); Group B ($AHHO_2/VL_1$) showed significant differences for item 1 (engagement), item 3 (ease of use), item 8 (engagement), item 9 (self-efficacy), and item 10 (efficiency). At first glance, it would appear that perception data may have been influenced by the order of the laboratory experiences.

Lindsay et al. (2009) posited that a students' first interaction with equipment builds a mental model of the laboratory experience and establishes the reality of what a 'laboratory experience' should be. As such, it could be argued that the Lab 2 user experience was not as good as the Lab 1 user experience (for both crossover group A and B) because the second lab experiences were not coherent with the participants' mental model of the first laboratory experiences. However, (based upon the review of students' open-ended responses presented in the previous section) it is more likely that the significant differences in MRISs for the crossover groups are due to the nature of the laboratory experiment (and participants' lack of self-confidence in measurement skills and mathematical abilities) rather than the nature of the laboratory experience (AHHO versus VL).

Table 4.8

Mean Raw Item Score Analysis: Lab 1 Survey versus Lab 2 Survey

p values for one-way ANOVA (top) and Kruskal-Wallis test (bottom)						
SUS Item	AHHO$_1$ AHHO$_2$	AHHO$_1$ VL$_1$	**AHHO$_1$ VL$_2$**	***AHHO$_2$ VL$_1$***	AHHO$_2$ VL$_2$	VL$_1$ VL$_2$
1 (N)	0.3402	0.1842	**0.2305**	***0.0035****	0.9926	0.0020*
	0.1601	0.3716	**0.4701**	***0.0004****	0.9934	0.0183*
2 (EU)	1.000	0.1488	**0.8903**	***0.2060***	0.9054	0.0506**
	1.000	0.1366	**0.8985**	***0.1922***	0.9168	0.0537**
3 (EU)	0.0027*	0.5093	**0.0684**	***<0.0001****	0.8127	0.0027*
	0.0046*	0.4971	**0.1386**	***<0.0001****	0.7950	0.0070*
4 (ET)	0.2367	0.9229	**0.8099**	***0.0803***	0.8299	0.4879
	0.2083	0.9623	**0.7773**	***0.0633***	0.7885	0.4846
5 (E)	0.9891	0.6593	**0.9546**	***0.5200***	0.9971	0.4239
	0.8676	0.7590	**0.8888**	***0.3243***	1.0000	0.3624
6 (E)	0.8043	0.7919	**0.0441***	***0.3139***	0.3454	0.0049*
	0.6039	0.7356	**0.0602**	***0.2934***	0.4019	0.0059*
7 (F)	0.6079	0.7765	**0.1104**	***0.1673***	0.7413	0.0145*
	0.5958	0.8294	**0.1615**	***0.1850***	0.7789	0.0296*
8 (N)	0.9534	0.1089	**0.0018***	***0.0482****	0.0174*	<0.0001*
	0.8836	0.2117	**0.0078***	***0.0223****	0.0286*	<0.0001*
9 (SE)	0.0569**	0.8586	**0.0398***	***0.0092****	0.9960	0.0065*
	0.0385*	0.9652	**0.0324***	***0.0062****	0.9996	0.0056*
10 (F)	0.3867	0.6740	**0.5855**	***0.0523*****	0.9949	0.1196
	0.1868	0.9144	**0.5049**	***0.0191****	0.9923	0.1293

Note. ANOVA = analysis of variance; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience; $_1$ = results from Lab 1 Survey; $_2$ = results from Lab 2 Survey; * $p < 0.05$, reject the null = difference in means is significant; $^{**}p \approx 0.05$, weak evidence to accept or reject the null = difference in means may be significant; **bold** = Group A crossover experience (AHHO then VL); ***bold*** = Group B crossover experience (VL then AHHO).

The results also showed that the differences in the MRIS values for Lab 1 ($AHHO_1/VL_1$) and Lab 2 ($AHHO_2/VL_2$) were not significant ($p > 0.05$ for all items except item 8 for Lab 2). There were also no major significant differences for $AHHO_1/AHHO_2$ ($p > 0.05$ for all items except item 3 and item 9). The most significant differences in MRIS values were found in VL_1/VL_2 ($p < 0.05$ for 6 of the 10 items). Again, this is most likely due to the differences in number and complexity of mathematical calculations between VL_1 and VL_2.

The nature of the simulator may have also contributed to the disparity. The simulator utilized for Lab 2 had more pop-up ads than the simulator for Lab 1. Pop-up ads are common with free online tools, but can detract from the user experience. Furthermore, there were more steps involved in data collection for the Lab 2 simulator than the Lab 1 simulator; participants had to manipulate the virtual ruler and three virtual images for VL_2, rather than simply clicking on a virtual coin for VL_1.

It should also be noted that a one-way ANOVA (α=0.05) and the Kruskal-Wallis test (α=0.05) was conducted to determine is the differences between MRISs values were significant for AHHOQ/AHHOU, AHHOQ/VLQ, AHHOQ/VLQ, AHHOU/VLQ, AHHOU/VLU, and VLU/VLQ for each trial and between trials. This data is not presented in Table 4.8 as the differences in MRIS values were not significant for any of these comparisons (all of the p-values were >0.05, fail to reject H_0). This means there were no significant differences in the perceptions of the U-term and Q-term participant sub-populations.

While the SUS score (an aggregate of the raw item scores) has proven to be a reliable measure of the user experience (Bangor et al., 2009; Brooke, 2013; Lewis, 2019), Lewis and Sauro (2018) proposed that the raw item score analysis could provide a more expansive understanding of the user experience. Their assertion was proven true as the trends identified in this section are reflected throughout (and helped to make sense of) the rest of the quantitative analyses. The next section will focus on the SUS score calculation and analysis.

SUS Questionnaire: Perceived Usability (Research Question 1-1). The SUS score was calculated for each participant based upon their responses to the ten items for the SUS questionnaire (Question 2 on the Lab 1 survey and the Lab 2 Survey). The equation functions in Excel were used to calculate the normed SUS score according to the steps outlined in Sauro and Lewis (2016):

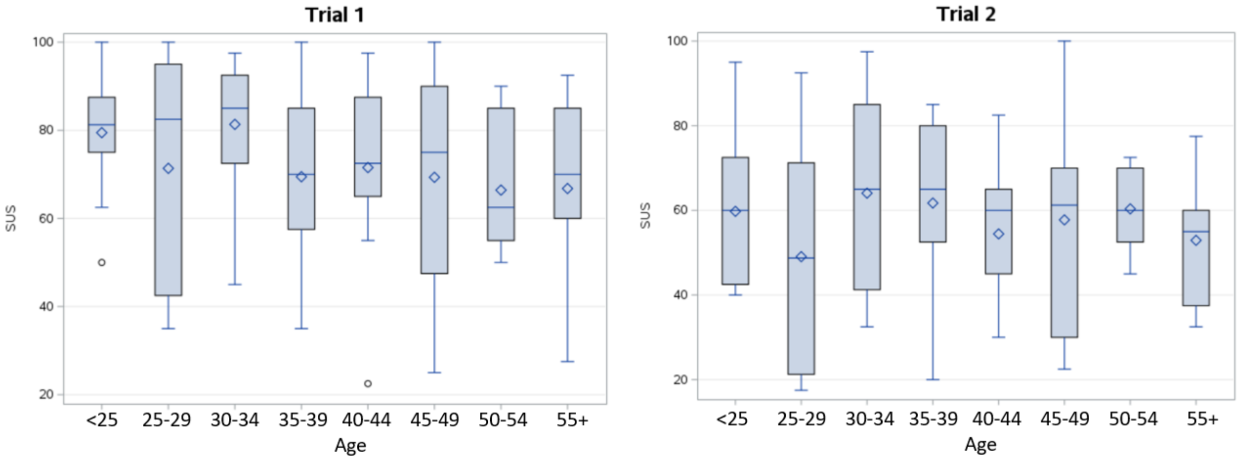
i. Assign all non-response items a raw score of 3 (neutral, on a scale in which 1=strongly disagree and 5=strongly agree)
ii. Subtract 1 from the raw score for each odd-numbered item (positive tone);
iii. Subtract the raw score of each even-numbered item (negative tone) from 5;
iv. Add all of the adjusted raw item scores together (yields max. sum of 40)
v. Multiply this sum by 2.5 (yields max. score of 100).

The SUS scores were organized by trial (Lab 1, trial-1; Lab 2, trial-2), by laboratory experience (AHHO/VL as determined by Question 1 on the Phase I Surveys), and by student population (U-term, Q-term, and Overall).

SUS Score and Age: Trial-1 and Trial-2. Participants' ages were collected in 5- year increments (<25, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55+). A test for correlation was performed using SAS OnDemand for Academics (H_0: r=0; H_a: $r \neq 0$, $\alpha = 0.05$) to determine if the slight negative correlation between the SUS rating and participants' age identified in previous studies also existed in this study population (Bangor et al., 2008; Granić & Ćukušić, 2011; Harrati et al., 2016; Vlachogianni & Tselios, 2021). Figure 4.5 shows the small non-significant negative correlation that was found between the SUS score and the age groups of the respondents for both trial-1 (r = -0.1982, p=0.0555) and trial-2 (r=-0.0194, p=0.8713). However, it should be noted that the results of the correlation analysis were based upon age groups (midpoint for each age range, age of 24 for the <25 group, and 56 for 55+ age group) rather than respondents' actual ages. As such, the results of the correlation analysis were most likely not valid

Figure 4.5

SUS versus Age: Trial-1 and Trial-2



Note. box plots show a slight (nonsignificant) negative correlation for Trial-1 and a very slight (nonsignificant) negative correlation for Trial-2. The plots also show that the largest difference in mean SUS score (between Trial-1 and Trial-2) were within the <25, 25-29, and 30-34 age groups.

Although the correlation analysis was most likely not valid (as respondents' actual ages were not collected), the boxplots did reveal some interesting trends. Figure 4.5 shows that the mean SUS scores were fairly consistent for the 35-39, 40-44, 45-49, 50-54, and 55+ age groups for both trials (SUS score ≈ 65 ±5). It also shows that the largest difference in mean SUS scores from trial-1 to trial-2 is in the <25, 25-29, and 30-34 age groups. This steep decline in user experience for these age groups was further explored in the qualitative findings.

Mean SUS Scores: Trial-1 and Trial-2. The mean SUS scores (perceived usability) were calculated and analyzed for each trial using SAS OnDemand for Academics. Table 4.9 shows the mean normed SUS score, the standard deviation, and the letter grade based upon Sauro and Lewis' (2016) curved grading scale (Table 3.3).

Table 4.9

Comparison of Mean SUS Score Analysis: Trial-1 and Trial-2

	U-term (n=32)			Q-term (n=62)			Overall (n=95)		
Trial 1	Mean	SD	Grade	Mean	SD	Grade	Mean	SD	Grade
Group A (AHHO$_1$)	69.375	±18.700	C	68.889	±22.776	C	69.039	±21.421	C
Group B (VL$_1$)	77.206	±14.276	B+	74.444	±19.701	B	75.778	±17.563	B
		p-value			p-value			p-value	
t-test		0.1845			0.3146			0.0967	
	U-term (n=34)			Q-term (n=39)			Overall (n=73)		
Trial 2	Mean	SD	Grade	Mean	SD	Grade	Mean	SD	Grade
Group A (VL$_2$)	57.692	±14.522	D	53.810	±24.143	D	55.294	±20.824	D
Group B (AHHO$_2$)	57.500	±20.256	D	62.875	±20.021	C-	60.329	±20.044	D
		p-value			p-value			p-value	
t-test		0.9769			0.1995			0.2997	

Note. SD = standard deviation; AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience (VL). The mean SUS$_{B1}$ (VL) > mean SUS$_{A1}$ (AHHO) for U-term, Q-term, and Overall datasets; the overall grade for VL$_1$ (B = average user experience) was higher than the overall grade for AHHO$_1$ (C = average user experience). The mean SUS$_{B2}$ (AHHO) > mean SUS$_{A2}$ (VL) for the Q-term and Overall datasets; the SUS$_{B2}$ (AHHO) ≈ mean SUS$_{A2}$ (VL) for the U-term dataset. Both the AHHO and VL for trial 2 received a D-grade (below average user experience). All of the p-values were > 0.05; therefore, the differences in mean perceived usability were not significant.

As the adjusted raw item scores were used to calculate the SUS score, it is not surprising that the findings for the mean SUS data are similar to those from the analysis of MRIS data. The mean SUS (and grade) for VL_1 was greater than $AHHO_1$ for the U-term, Q-term, and Overall datasets; and the Lab 2 user experience (for both $AHHO_2$ and VL_2) was not as positive as the Lab 1 user experience. As previously discussed, these differences may have more to do with the nature of the laboratory (number and complexity of mathematical calculations) than the nature of the laboratory experience (AHHO versus VL).

Table 4.9 also shows the results of the two-tailed, two-sample t-tests ($\alpha = 0.05$) conducted to determine if the differences in mean SUS scores for trial-1 (Group A_1 versus Group B_1) and trial-2 (Group A_2 versus Group B_2) were significant (H_0: $\mu_{SUSA} - \mu_{SUSB} = 0$; H_a: $\mu_{SUSA} - \mu_{SUSB} \neq 0$). Like the findings of the MRIS analysis, the differences in mean SUS scores were not significant (all p-values > 0.05, therefore we fail to reject H_0). As such, the mean SUS analysis for trial-1 and trial - did not yield conclusive results for participants' preferred laboratory experience.

It should be noted that the distribution of SUS scores for VL_1 (U-term, Q-term, and Overall) was heavily skewed to the left. As the two-sample t-test is based on the condition of normality for each group (Ott & Longnecker, 2016), the conclusions for the $AHHO_1$ versus VL_1 tests may not be valid. A Wilcoxon rank sum test (a nonparametric alternative to the two-sample t-test) was performed in order to determine if the conclusions for the t-test were indeed accurate. The results of the Wilcoxon rank sum test (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$) yielded the same conclusion as the two-sample t-tests; all p-values > 0.05 ($p_U = 0.2263$; $p_Q = 0.3661$; and $p_{Overall} = 0.1301$). This means that there were no significant differences in populations.

Mean SUS Scores: Trial-1 versus Trial-2. In this section, the mean SUS analysis will shift from the results from the individual trials to a comparison of results between trials. A two-tailed, two-sample t-test ($\alpha = 0.05$) was conducted to determine if the differences mean SUS score for trial-1 versus trial- were significant (H_0: $\mu_{SUS1} - \mu_{SUS2} = 0$; H_a: $\mu_{SUS1} - \mu_{SUS2} \neq 0$). Table 4.10 provides a summary of the mean perceived usability (SUS score) analysis.

Table 4.10 also shows the results of the Wilcoxon rank sum test (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$) where appropriate. Again, the distribution of SUS scores for VL_1 (U-term, Q-term, and Overall) was heavily skewed to the left and did not meet all of the conditions for the two-sample t-test. As such, a non-parametric Wilcoxon rank sum test was performed for comparisons that involved VL_1. As seen in the trial-1and trial-2 data, the results of the Wilcoxon rank sum yielded the same conclusions as the two-sample t-tests (with the exception of Q-term $AHHO_2/VL_1$ in which the Wilcoxon two-sample test clarified the conclusion from the t-test).

The findings show there were significant differences in the mean perceived usability for the Group A crossover experiences (Q-term and Overall) and Group B crossover experiences (U-term, Q-term, and Overall). As previously mentioned, these differences may have more to do with the nature of the laboratory than the nature of the laboratory experience (AHHO versus VL). The results also showed that the differences in the SUS scores for Lab 1 ($AHHO_1/VL_1$) and Lab 2 ($AHHO_2/VL_2$) were not significant ($p > 0.05$ for U-term, Q-term, and Overall). There were also no major significant differences for $AHHO_1/AHHO_2$ ($p > 0.05$ for U-term, Q-term; $p \approx 0.5$ for Overall).

Table 4.10

Comparison of Mean SUS scores: t-tests for Trial-1 versus Trial-2

	AHHO$_1$ AHHO$_2$	AHHO$_1$ VL$_1$	**AHHO$_1$** **VL$_2$**	***AHHO$_2$*** ***VL$_1$***	AHHO$_2$ VL$_2$	VL$_1$ VL$_2$
U-term						
SUS	69.375	69.375	**69.375**	***57.500***	57.500	77.206
	57.500	77.206	**57.692**	***77.206***	57.692	57.692
t-test	0.0865	0.1845	**0.0762**	***0.0014****	0.9769	0.0006*
W		0.2263		***0.0030****		0.0020*
Q-term						
SUS	68.889	68.889	**68.889**	***62.875***	62.875	74.444
	62.875	74.444	**53.810**	***74.444***	53.810	53.810
t-test	0.3280	0.3146	**0.0219***	***0.0542*****	0.1995	0.0021*
W		0.3661		***0.0488****		0.0036*
Overall						
SUS	69.039	69.039	**69.039**	***60.329***	60.329	75.778
	60.329	75.778	**55.294**	***75.778***	55.294	55.294
t-test	0.0535**	0.0967	**0.0042***	***0.0003****	0.2997	<0.0001*
W		0.1301		***0.0005****		<0.0001*

Note. AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience; $_1$ = results from Lab 1 Survey; $_2$ = results from Lab 2 Survey; W = Wilcoxon sum rank test (nonparametric test); * $p < 0.05$, reject the null = significant difference in means (t-test) or significant difference in populations (W); **$p \approx 0.05$, weak evidence to reject the null = differences may be significant; **bold** = Group A crossover experience (AHHO then VL); ***bold*** = Group B crossover experience (VL then AHHO). The trial 1 data (AHHO$_1$ vs. VL$_1$) and trial 2 data (AHHO$_2$ vs. VL$_2$) were included in this table for reference.

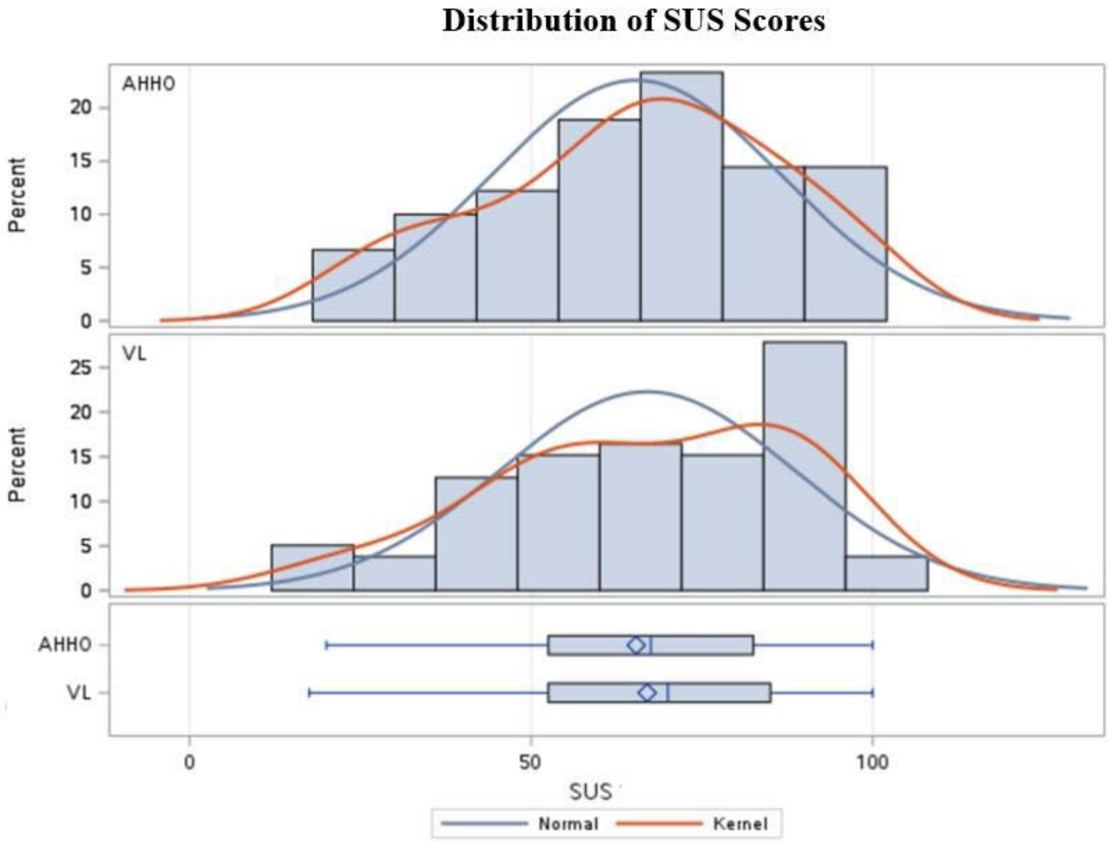
Like the MRIS findings, the most significant differences in mean perceived usability was found in VL_1/ VL_2 ($p << 0.05$ for U-term, Q-term, and Overall). As previously discussed, these results suggest that math self-efficacy, measurement self-efficacy, and the simulator utilized for Lab 2 may have negatively impacted the user experience. The Lab 2 simulator had more pop-up ads than the Lab 1 simulator and the Lab 2 simulator required more "work" than the Lab 1 simulator. Students had to upload virtual images, manipulate the virtual ruler to measure the length, width, and height of each object multiple times for Lab 2; while students only had to click on a virtual coin for Lab 1. This affirms that the SUS is an effective instrument that is sensitive to various dimensions of the user experience.

Finally, it should be noted that comparison tests were also run between trials for U-term and Q-term data (Q_1/U_1, Q_1/U_2, Q_2/U_2, and Q_2/U_1). For simplicity in reporting, these data were not shown in Table 4.10. as all of the differences in mean perceived usability were not significant (all p-values were > 0.05). For example, the p-value for $AHHO_{Q1}/AHHO_{U1}$ was 0.9408, for $AHHO_{Q2}/AHHO_{U2}$ was 0.4166, and for VL_{Q2}/VL_{U2} was 0.6049. This means there were not significant differences in the perceptions of the U-term and Q-term participant sub-populations within each trial.

Mean SUS Scores: AHHO versus VL. For the final comparison of mean perceived usability, a two-tailed, two-sample t-test ($\alpha=0.05$) was conducted to determine if the differences mean SUS score for AHHO (trial 1 + trial 2) and VL (trial 1 + trial 2) were significant (H_0: $\mu_{AHHO} - \mu_{VL}= 0$; H_a: $\mu_{AHHO} - \mu_{VL} \neq 0$). Figure 4.6 shows that the distribution of SUS scores for VL did not meet the condition of normality for the two-sample t-test (Ott & Longnecker, 2016). Therefore, the Wilcoxon rank sum test (a nonparametric alternative to the two-sample t-test) was also performed to determine if the differences in sample populations were significantly different (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$). The results are shown in Table 4.11.

Figure 4.6

Distribution of SUS scores for AHHO and VL



Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience; SUS = system usability scale. The distribution of SUS scores for AHHO was approximately normal; however, the VL distribution was not as close to normal.

Table 4.11

Comparison of Mean SUS Scores: T-test and Wilcoxon Rank Sum Test - AHHO vs VL

	t-test		Wilcoxon rank sum test	
	Mean	p-value	Mean Score	p-value
AHHO	65.3611*	0.6269	83.2278	0.6161
VL	66.9620*		87.0190	

Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience. The p-value > 0.05 for both the two-sample t-test and the Wilcoxon rank sum test; therefore, the difference in means (t-test) or populations (Wilcoxon rank sum test) is not significant.

*Based upon the curved grading scaled developed by Sauro and Lewis (2016), the mean SUS for both the AHHO and VL experiences fall at the lower end of the C-range (65.0 – 71.0).

The findings showed that the perceived usability was nearly the same for AHHO and VL. The difference is means (65.3611_{AHHO} and 66.9620_{VL}) was not significant ($p=0.6269 > 0.05$; fail to reject H_0); the shift between populations (87.2278_{AHHO} and 87.0190_{VL}) was not significant ($p=0.6161 > 0.05$; fail to reject H_0). Although the findings were inconclusive for preferred laboratory experience, the conclusion of 'nearly the same' is consistent with the literature. The next section will focus on the Mean Adjusted Raw Item Scores (MARIS) for each sub-factor.

SUS Questionnaire: Sub-factors (Research Question 1-1'). Based upon Lewis and Sauro's (2019) description of item attributes, each of the SUS items were coded as a measure of effectiveness, efficiency, engagement, error tolerance, ease of use, or self-efficacy. Each variable corresponded to both a positive statement and negative statement, with the exception of error tolerance (negative statement only) and self-efficacy (positive statement only). Table 4.12 provides a summary of the sub-factor assigned to each SUS item.

Table 4.12

Summary of Item Codes

Variable	SUS item	Tone
Effectiveness (E)	5	(+)
	6	(-)
Efficiency (F)	7	(+)
	10	(-)
Engagement (N)	1	(+)
	8	(-)
Error Tolerance (ET)	4	(-)
Ease of Use (EU)	2	(-)
	3	(+)
Self-efficacy (SE)	9	(+)

Note. SUS = system usability scale; (+) = positive-toned statement; (-) = negatively-toned statement.

As the raw item scores for the positive statements would be incongruent with the raw item scores for the negative statements, all of the raw item scores were adjusted prior to combining like-item scores using the equation function in Excel.

i. Subtract 1 from the raw score for each odd-numbered item (positive tone);

ii. Subtract the raw score of each even-numbered item (negative tone) from 5.

iii. Add each of the like- adjusted raw item scores together for each variable (according to Table 4.12).

The combined mean adjusted raw item scores (MARIS) were used to evaluate the SUS sub-factors in the following sections. It should be noted that the first two steps were completed as part of the SUS score calculation (Sauro & Lewis, 2016). The third step, adding the adjusted raw item scores for like-variables, is unique to this study.

Mean Adjusted Raw Item Score: Summary Statistics. The combined MARISs were calculated for each trial using SAS OnDemand for Academics. It should be noted that the adjusted raw item scores range from 0-4 where *0*=strongly disagree, *1*= somewhat disagree, *2*= neutral, *3*= somewhat agree, and *4*= strongly agree. Table 4.13a and 4.13b shows the MARIS summary statistics for each trial.

Table 4.13a

Summary Statistics: Mean Adjusted Raw Item Score – Trial 1

Trial-1 Lab 1	U-term (n=33)		Q-term (n=62)		Overall (n=95)	
Variable	MARIS $^{A}_{B}$	SD $^{A}_{B}$	MARIS $^{A}_{B}$	SD $^{A}_{B}$	MARIS $^{A}_{B}$	SD $^{A}_{B}$
E	2.9375	±1.1053	2.8889	±1.0949	2.9038	±1.0930
	3.0278	±1.1335	3.1111	±0.9842	3.0778	±1.0410
F	2.7189	±1.3010	2.6806	±1.1362	2.6923	±1.1830
	3.0833	±0.9373	2.6667	±1.0989	2.8333	±1.0520
N	2.6250	±1.4536	2.6389	±1.3141	2.6346	±1.3514
	3.250	±0.8062	3.0556	±1.0889	3.1333	±0.9853
ET	2.7500	±1.1255	2.6944	±1.3695	2.7115	±1.2885
	3.0556	±0.9376	2.6667	±1.3301	2.8222	±1.1957
EU	2.6250	±1.2889	2.8472	±1.1216	2.7788	±1.1738
	3.1389	±0.8670	3.16667	±1.0946	3.1556	±1.0046
SE	3.1875	±0.9811	2.7222	±1.3226	2.8654	±1.2372
	3.0556	±0.9984	2.9630	±1.0913	3.0000	±1.0444

Note. MARIS = mean adjusted raw item score; SD = standard deviation; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; $\frac{A}{B}$ = $\frac{Group\ A}{Group\ B}$; Group A = at-home, hands-on laboratory experience (AHHO); Group B = virtual laboratory experience (VL). There were 2 U-term responses and 1 Q-term response that were removed from the dataset as all items were given the same, non-neutral rating.

Table 4.13b

Summary Statistics: Mean Adjusted Raw Item Score – Trial 2

Trial-2 Lab 2	U-term (n=34)		Q-term (n=39)		Overall (n=73)	
Variable	MARIS $\frac{A}{B}$	SD $\frac{A}{B}$	MARIS $\frac{A}{B}$	SD $\frac{A}{B}$	MARIS $\frac{A}{B}$	SD $\frac{A}{B}$
E	2.5385	±1.0670	2.4048	±1.2309	2.4559	±1.1646
	2.4474	±1.1076	2.9500	±0.8458	2.7051	±1.0079
F	2.3462	±1.2310	2.1190	±1.1935	2.2059	±1.2040
	2.3421	±1.0724	2.2750	±1.0124	2.3077	±1.0358
N	1.7692	±1.4229	1.8333	±1.4637	1.8088	±1.4378
	2.1316	±1.2340	2.4500	±1.1756	2.2949	±1.2072
ET	3.0000	±1.1547	2.0476	±1.2836	2.4118	±1.3054
	1.8421	±1.2314	2.4500	±1.2344	2.1538	±1.2468
EU	2.3846	±1.1688	2.2857	±1.4362	2.3235	±1.3321
	2.1316	±1.2980	2.4750	±1.2192	2.3077	1.2619
SE	2.0000	±1.0000	2.1905	±1.3274	2.1176	±1.2001
	1.9474	±1.1772	2.4000	±1.0954	2.1795	±1.1441

Note. MARIS = mean adjusted raw item score; SD = standard deviation; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; $\frac{A}{B}$ = $\frac{Group\ A}{Group\ B}$; Group A = virtual laboratory experience (VL); Group B = at-home, hands-on laboratory experience (AHHO).

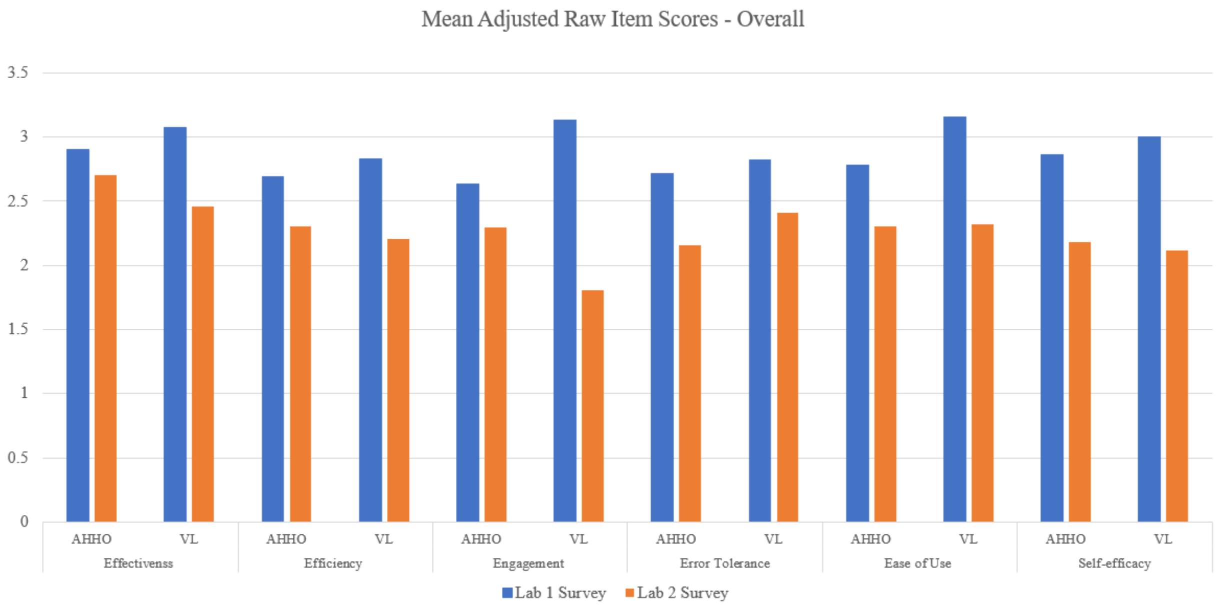
Like the MRIS results, the findings for Lab 2 were not as positive as for Lab 1. The results from the Lab 1 Survey indicated that participants in both groups (VL$_1$ and AHHO$_1$) had an average-good user experience (all MARIS values >2.5) and the overall user experience for VL$_1$ > AHHO$_1$. However, the Lab 2 Survey results largely indicated that the participants had a below average user experience (all MARIS values < 2.5, with the exception of effectiveness

$AHHO_2 = 2.7$). Again, this implies that participants were able to achieve the specified goals of the laboratory; however, they found it to be difficult and did not enjoy the experience.

Comparison of Mean Adjusted Raw Item Score: Trial-1 and Trial-2. Figure 4.7 provides a visual comparison of the MARIS results for the Phase I surveys. The bar graph shows all of the MARISs for $VL_1 > AHHO_1$; therefore, it would appear that respondents preferred the virtual laboratory experience for Lab 1. The graph also shows the overall MARIS value for effectiveness, efficiency, and engagement of $AHHO_2 > VL_2$; the MARIS for error tolerance of $VL_2 > AHHO_{2;}$ and the MARIS for ease of use and self-efficacy of $AHHO_2 \approx VL_2$. As such, the Lab 2 MARIS data was inconclusive for preferred laboratory experience and which variable(s) played a significant role in the user experience.

Figure 4.7

Summary: MARIS for Lab 1 Survey and Lab 2 Survey



Note. The bar graph illustrates the mean adjusted raw item score (overall) for each variable (sub-factor of the SUS score).

A two-tailed, two sample t-test ($\alpha = 0.05$) was conducted to determine if the differences in MARISs for trial-1 (Group A_1 versus Group B_1) and trial-2 (Group A_2 versus Group B_2) were significant (H_0: $MARIS_A - MARIS_B = 0$; H_a: $MARIS_A - MARIS_B \neq 0$). However, the results of the t-test may not be valid as the adjusted raw item scores were not normally distributed. To determine if the differences in populations were significant (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$) a Wilcoxon rank sum test (a nonparametric alternative to the two-sample t-test) was also performed (Ott & Longnecker, 2016). Table 4.14 provides a summary of the MARIS analysis for trial-1 and trial-2 datasets.

Table 4.14

Summary of Trial 1 and Trial 2: MARIS, User experiences, and p-values for Statistical Tests

	Lab Survey 1 Results				Lab Survey 2 Results			
Variable	Group			p-value	Group			p-value
	A (AHHO)	B (VL)	UX	T W	A (VL)	B (AHHO)	UX	T W
	MARIS	MARIS			MARIS	MARIS		
E	2.904	3.078	VL > AHHO	0.2599 0.2147	2.456	2.705	AHHO > VL	0.1678 0.2161
F	2.692	2.833	VL > AHHO	0.3847 0.5112	2.206	2.308	AHHO > VL	0.5837 0.6388
N	2.635	3.133	VL > AHHO	0.0042* 0.0179*	1.809	2.295	AHHO > VL	0.0280* 0.0264*
ET	2.712	2.822	VL > AHHO	0.6634 0.7430	2.412	2.154	VL > AHHO	0.3913 0.3551
EU	2.779	3.156	VL > AHHO	0.0182* 0.0203*	2.324	2.308	VL > AHHO	0.9143 0.8658
SE	2.865	3.000	VL > AHHO	0.5673 0.7605	2.118	2.180	AHHO > VL	0.8225 0.9226

Note. MARIS = mean adjusted raw item score; UX = user experience; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience; T= two-sample t-test; W = Wilcoxon rank sum test; * $p < 0.05$, reject the null = difference in means/populations is significant.

Based upon the t-tests (and corroborated by the Wilcoxon rank sum tests), the results of the Lab 1 Survey were not as conclusive as they appeared. Although the user experience for the VL_1 was greater than $AHHO_1$ for all variables, the differences in MARIS values were only significant was for engagement (p=0.0042) and ease of use (p=0.0182). This means that the students preferred the VL laboratory experience for trial-1 based upon engagement and ease of use; the preferred laboratory experience for the other sub-factors (effectiveness, efficiency, error tolerance, and self-efficacy) was inconclusive.

For the Lab 2 Survey, engagement (p=0.0280) was the only variable that showed a significant difference between the MARIS values ($AHHO_2 > VL_2$). This means that the students preferred the AHHO laboratory experience for trial-2 based upon engagement; the preferred laboratory experience for the other sub-factors (effectiveness, efficiency, error tolerance, ease of use, and self-efficacy) was inconclusive.

It should be noted that the standard deviations for many of the items in both trials were relatively high (≥50% of the mean) (Table 4.12). Therefore, it was not surprising that the difference in MARIS values for most of the variables were not significant (i.e. the difference in the means was not larger than the variance). As the differences in MARIS values for engagement in both trial-1 and trial-2 were significant, engagement most likely played a significant role in the users' preferred laboratory experience.

Comparison of Mean Adjusted Raw Item Scores: Trial-1 vs. Trial-2. A one-way ANOVA (α=0.05) and its non-parametric alternative (Kruskal-Wallis test) were conducted to determine of the differences between MARIS for trial-1 and trial-2 for each sub-factor were significant. The results of this analysis are found in Table 4.15.

Table 4.15

Mean Adjusted Raw Item Score Analysis: Lab 1 Survey versus Lab 2 Survey

p values for one-way ANOVA (top) and Kruskal-Wallis test (bottom)						
Variable	AHHO$_1$ AHHO$_2$	AHHO$_1$ VL$_1$	**AHHO$_1$** **VL$_2$**	***AHHO$_2$*** ***VL$_1$***	AHHO$_2$ VL$_2$	VL$_1$ VL$_2$
E	0.6059	0.6753	**0.0394***	***0.1147***	0.5021	0.0021*
	0.3838	0.6000	**0.0445***	***0.0268****	0.6020	0.0018*
F	0.1023	0.8185	**0.0290***	***0.0139****	0.9472	0.0031*
	0.0586	0.9126	**0.0434***	***0.0071****	0.9654	0.0064*
N	0.2684	0.0300*	**0.0002***	***0.0001****	0.0905	<0.0001*
	0.1629	0.0832	**0.0018***	***<0.0001****	0.1170	<0.0001*
ET	0.1595	0.9729	**0.7018**	***0.0756***	0.8183	0.4786
	0.1459	0.9874	**0.6876**	***0.0633***	0.7885	0.4846
EU	0.0417*	0.1240	**0.0682**	***<0.0001****	0.9998	0.0001*
	0.0492*	0.0931	**0.1164**	***<0.0001****	0.9982	0.0003*
SE	0.0296*	0.9409	**0.0204***	***0.0080****	0.9958	0.0055*
	0.0222*	0.9898	**0.0190***	***0.0062****	0.9996	0.0056*

Note. ANOVA = analysis of variance; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience; $_1$ = results from Lab 1 Survey; $_2$ = results from Lab 2 Survey; * p < 0.05, reject the null = difference in means is significant; **p ≈0.05, weak evidence to accept or reject the null = difference in means may be significant; **bold** = Group A crossover experience (AHHO then VL); ***bold*** = Group B crossover experience (VL then AHHO).

Based upon the one-way ANOVA test (and corroborated by the Kruskal-Wallis test), there were significant differences between combined MARIS values for Group A for all of the variables except error tolerance (p=0.7018) and ease of use (p=0.0682); Group B showed significant differences for all of the variables except for error tolerance (p=0.0756). The Lab 2 user experience (for both AHHO and VL) was shown to be a below average experience across all of the statistical tests. As previously discussed, these differences are most likely due to the nature of Lab 2 (and participants' lack of confidence in measurement skills and mathematics abilities) than the nature of the laboratory experience (AHHO versus VL).

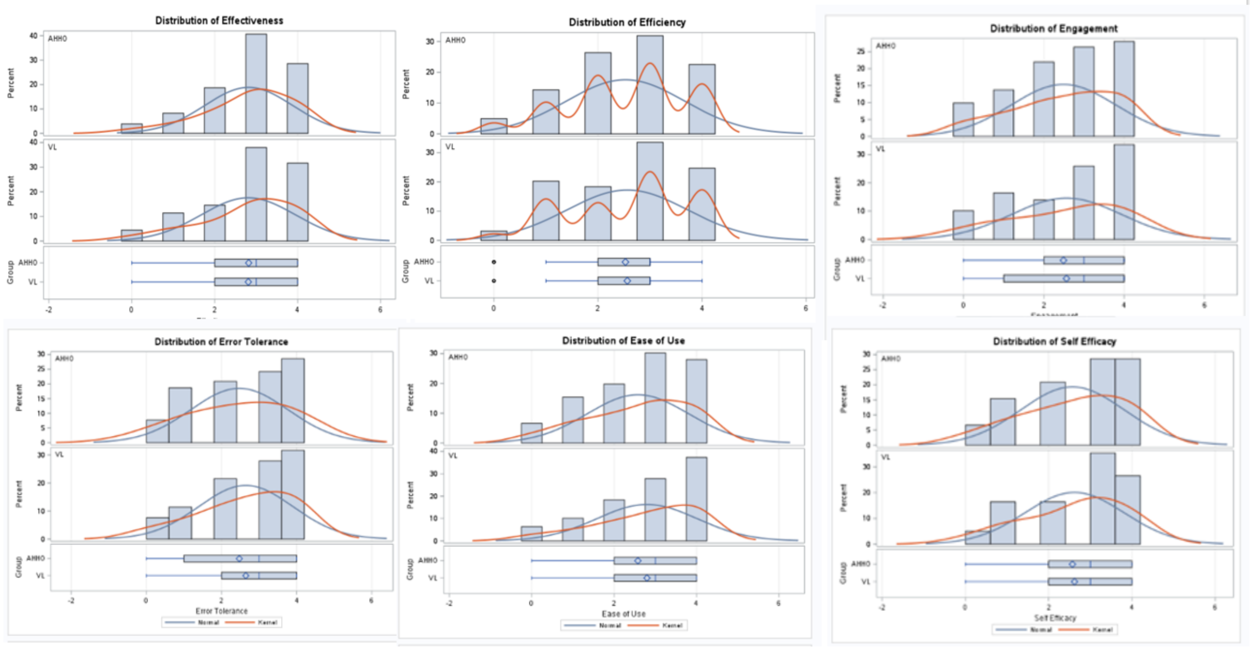
The results also showed that the differences in the MARIS values (sub-factors) for Lab 1 ($AHHO_1/VL_1$) and Lab 2 ($AHHO_2/VL_2$) were not significant ($p > 0.05$ for all variables). The $AHHO_1/AHHO_2$ data showed a significant difference in MARIS for self-efficacy (p=0.0222); the ease of use data was inconclusive as $p \approx 0.05$; and no significant differences in MARIS for the remaining variables ($p > 0.05$). The most significant differences in MRIS values were found in VL_1/ VL_2 ($p < 0.05$ for all but error tolerance). Again, this may have more to do with the nature of the simulator than the nature of the laboratory experience (crossover from $AHHO_1$ to VL_2).

It should also be noted that a one-way ANOVA ($\alpha=0.05$) and the Kruskal-Wallis test ($\alpha=0.05$) was conducted to determine is the differences between MARIS values were significant for AHHOQ/AHHOU, AHHOQ/VLQ, AHHOQ/VLQ, AHHOU/VLQ, AHHOU/VLU, and VLU/VLQ for each trial (H_0: MARIS values are equal; H_a: at least one MARIS is different) for each trial. This data was not presented in Table 4.15 as the differences in MARIS values were not significant for any of these comparisons (all of the p-values were >0.05, fail to reject H_0). This means there were not significant differences in the perceptions of the U-term and Q-term participant sub-populations within each trial.

Comparison of Mean Adjusted Raw Item Scores: AHHO versus VL. For the final comparison of sub-factors, a two-tailed, two sample t-test ($\alpha=0.05$) was conducted to determine if the differences combined MARIS values for AHHO (trial 1 + trial 2) and VL (trial 1 + trial 2) were significant (H_0: $\mu_{AHHO} - \mu_{VL} = 0$; H_a: $\mu_{AHHO} - \mu_{VL} \neq 0$). As the t-test diagnostics revealed the combined adjusted raw item scores (for the most part) were not normally distributed (Figure 4.8), the Wilcoxon rank sum test (a nonparametric alternative to the two-sample t-test) was also performed to determine if the differences in sample populations were significantly different (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$). The results are shown in Table 4.16.

Figure 4.8

Distribution of Mean Adjusted Raw Item Scores: AHHO versus VL for each Sub-Factor



Note. AHHO (top distribution in pair) = at-home, hands-on laboratory experience; VL (bottom distribution in pair) = at-home, virtual laboratory experience. Top row shows the MARIS distribution for Effectiveness, Efficiency, and Engagement; bottom row shows the MARIS distribution for Error Tolerance, Ease of Use and Self-Efficacy.

Table 4.16

Sub-Factor Analysis: T-tests and Wilcoxon Rank Sum Tests - AHHO vs VL

Variable	Mean $^{AHHO}_{VL}$	p-values	
		Two-sample t-test	Wilcoxon rank sum test
E	2.8187 2.8101	0.9428	0.8307
F	2.5275 2.5633	0.7740	0.7371
N	2.4090 2.5633	0.6080	0.4758
ET	2.4725 2.6456	0.3784	0.3795
EU	2.5769 2.7975	0.0996	0.0726
SE	2.5714 2.6203	0.7945	0.8375

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience. The p-value > 0.05 for both the two-sample t-test and the Wilcoxon rank sum test for all variables; therefore, the difference in means (two-sample t-test) or populations (Wilcoxon rank sum test) is not significant.

The findings showed that all of the sub-factors were nearly the same for AHHO and VL. The difference is means for effectiveness (2.8187_{AHHO} and 2.8101_{VL}), efficiency (2.5275_{AHHO} and 2.5633_{VL}), engagement (2.4090_{AHHO} and 2.5633_{VL}), error tolerance (2.4725_{AHHO} and 2.6456_{VL}), ease of use (2.5769_{AHHO} and 2.7975_{VL}), and self-efficacy (2.5714_{AHHO} and 2.6203_{VL}) were not significant (all $p > 0.05$; fail to reject H_0). Moreover, the MARIS for all subfactors indicated an average user experience (all ≥2.5, with the exception of $\text{engagement}_{AHHO} \approx 2.4$).

Although the findings were inconclusive for preferred laboratory experience, the conclusion of 'nearly the same' is consistent with the literature.

SUS Questionnaire: Multiple Regression Analysis (Research Questions 2). Multiple regression was used to analyze the relationship between user experience (perceived usability) and the SUS sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, self-efficacy). The general format for the multiple regression model is:

$SUS = \beta_0 + \beta_1 E + \beta_2 F + \beta_3 N + \beta_4 ET + \beta_5 EU + \beta_6 SE$, where β_0= y-intercept; β_i partial slope for specified sub-factor; E = combined MARIS for effectiveness, F = combined MARIS for efficiency; N = combined MARIS for engagement; ET = combined MARIS for error tolerance; EU = combined MARIS for ease of use; and SE = combined MARIS for self-efficacy. Multiple regression analysis assumes that there is a linear relationship between the dependent (perceived usability) and explanatory variables (SUS sub-factors); independence of errors; normality; homoscedasticity (equal or similar variance in different groups being compared); and that the explanatory variables are not highly correlated (Ott & Longnecker, 2016). The diagnostic plots were analyzed to determine if the conditions of linearity, homoscedasticity and normality were met; correlation analysis was performed using SAS OnDemand for Academics to determine if the condition of no multicollinearity was met. The results of the correlation analysis are shown in the next section.

Correlation Analysis – Relationships Between Variables. Table 4.17 shows the Pearson coefficients (r) for each of the correlations and the strengths of the relationship. The strength of each relationship was determined by the following: $r>0.90$ (very strong); $r = 0.70-0.89$ (strong); $r = 0.50 – 0.69$ (moderate); $r = 0.30 – 0.49$ (weak); $r = 0.05 – 0.29$ (very weak); and $r \approx 0$ (no correlation).

Table 4.17

Correlation Coefficients (r) and Strength of Relationships Between Variables

r	E		F		N		ET		EU		SE	
	AHHO	VL	AHHO	VL	AHHO	VL	AHHO	VL	AHHO	VL	AHHO	VL
E	-	-	0.43 W	0.35 W	0.62 M	0.62 M	0.40 W	0.19 VW	0.52 M	0.35 W	0.50 M	0.47 W
F	0.43 W	0.35 W	-	-	0.52 M	0.46 W	0.30 W	0.34 W	0.43 W	0.44 W	0.47 W	0.59 M
N	0.62 M	0.62 M	0.52 M	0.46 W	-	-	0.34 W	0.20 VW	0.54 M	0.51 M	0.56 M	0.67 M
ET	0.40 W	0.19 VW	0.30 W	0.34 W	0.34 W	0.20 VW	-	-	0.50 M	0.51 M	0.55 M	0.43 W
EU	0.52 M	0.35 W	0.43 W	0.44 W	0.54 M	0.51 M	0.50 M	0.51 M	-	-	0.53 M	0.58 M
SE	0.50 M	0.47 W	0.47 W	0.59 M	0.56 M	0.67 M	0.55 M	0.43 M	0.53 M	0.58 M	-	-

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; AHHO = at-home, hands-on laboratory experience; VL = virtual laboratory experience; M = moderate; W = weak; and VW = very weak. The Pearson coefficient (r) for each of the relationships is displayed; there is a weak – moderate, positive correlation between all of the variables. The strongest relationship was between effectiveness/engagement; the weakest relationships were between engagement/error tolerance and efficiency/error tolerance.

The findings show that there is a positive, very weak – moderate correlation between all of the variables. The strongest relationships were between effectiveness/engagement (r=0.62) and between self-efficacy/engagement (r=0.67); the weakest relationships were between engagement/error tolerance (r=0.20) and effectiveness/error tolerance (r=0.19). As the independent variables are not strongly correlated (all <0.70), the condition of independence of explanatory variables is somewhat met. To ensure the evidence is strong enough to support that one or more of the SUS sub-factors played a significant role in participants' user experience (H_0:

$\beta_i = 0$; H_a: $\beta_i \neq 0$), the results from this multiple regression analysis were cross-referenced with the MRIS and MARIS comparison results.

Multiple Regression Analysis: AHHO (Research Question 2a). Table 4.18 shows the U-term, Q-term, and Overall results of the multiple regression analysis for the AHHO version of the Lab 1 protocol; Table 4.19 shows the U-term, Q-term, and Overall results of the multiple regression analysis for the AHHO version of the Lab 2 protocol; and Table 4.20 shows the Overall combined results (Trial 1 + Trial 2) of the multiple regression for AHHO. Each data table includes the following information: model (linear equation), R^2-value (degree of fit), partial slope parameter for each variable (β_i = change in perceived usability (SUS score) with the change in the explanatory variable while the other explanatory variables are held constant), and the p-value for the regression analysis (H_0: $\beta_i = 0$; H_a: $\beta_i \neq 0$; significance based upon $\alpha=0.05$).

AHHO: Trial-1. The degree of fit is determined by the R^2- value; the closer the value to 1, the better the fit. In all cases, the R^2-value was fairly high (>0.90). This means that the model worked well as >90% of the variation was explained by the model. Moreover, an analysis of the residuals (for each test) showed the assumptions of the multiple regression analysis were satisfied (i.e. homoscedasticity and normality) (Ott & Longnecker, 2016).

Table 4.18 shows that the only variable that was significant was engagement ($p= 0.0106$; $\beta_3 = 7.2313$) for the U-term data set. The results of the Q-term dataset showed that all variables were significant ($p<0.05$), except efficiency and self-efficacy. The Overall dataset showed that all variables were significant ($p<0.05$); they are listed in order of the magnitude of the partial slope parameter: engagement ($\beta_3 = 5.2166$), ease of use ($\beta_5 = 4.4073$), error tolerance ($\beta_4 = 4.0207$), effectiveness ($\beta_1 = 4.0050$), efficiency ($\beta_2 = 2.4416$), and self-efficacy $\beta_6 = 2.2977$). The magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott & Longnecker, 2016). As engagement (U-term and Overall datasets) and ease of use (Q-term and Overall datasets) ranked high, these variables appeared to play a significant role in the AHHO user-experience for Lab 1.

Table 4.18

Multiple Regression Analysis: AHHO – Trial 1

$$SUS_{AHHO1} = \beta_0 + \beta_1 E_{AHHO1} + \beta_2 F_{AHHO1} + \beta_3 N_{AHHO1} + \beta_4 ET_{AHHO1} + \beta_5 EU_{AHHO1} + \beta_6 SE_{AHHO1}$$

		E	F	N	ET	EU	SE
U-term	R^2=0.9044						
n=33	Slope (β_i)	2.7072	2.3825	7.2313	4.0325	3.6600	1.8378
	p-value	0.3454	0.3252	0.0106^{*}	0.0983	0.0894	0.4956
Q-term	R^2=0.9578						
n=62	Slope (β_i)	3.7094	1.7784	4.4209	4.4447	6.0738	1.5519
	p-value	0.0024^{*}	0.0711	$<0.0001^{*}$	$<0.0001^{*}$	0.0001^{*}	0.1675
Overall	R^2=0.9392						
n=95	Slope (β_i)	4.0050	2.4416	5.2166	4.0207	4.4073	2.2977
	p-value	0.0001^{*}	0.0056^{*}	$<0.0001^{*}$	$<0.0001^{*}$	$<0.0001^{*}$	0.0090^{*}

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; R^2 = degree of fit; β_i = partial slope; * $p < 0.05$, reject the null = the explanatory variable played a significant role in participants' user experience.

AHHO: Trial-2. The degree of fit is determined by the R^2- value; the closer the value to 1, the better the fit. In all cases, the R^2-value was fairly high (>0.95). This means that the model worked well as >95% of the variation was explained by the model. Moreover, an analysis of the residuals (for each test) showed the assumptions of the multiple regression analysis were satisfied (i.e. homoscedasticity and normality) (Ott & Longnecker, 2016).

Table 4.19

Multiple Regression Analysis: AHHO – Trial 2

$SUS_{AHHO2} = \beta_0 + \beta_1 E_{AHHO2} + \beta_2 F_{AHHO2} + \beta_3 N_{AHHO2} + \beta_4 ET_{AHHO2} + \beta_5 EU_{AHHO2} + \beta_6 SE_{AHHO2}$

		E	F	N	ET	EU	SE
U-term	R^2=0.9654						
n=34	Slope (β)	7.6641	1.8413	5.0505	4.0056	4.8609	3.2277
	p-value	0.0096^*	0.3366	0.0683	0.0185^*	0.0060^*	0.0492^*
Q-term	R^2=0.9741						
n=39	Slope (β)	3.8565	5.1522	2.2258	3.5289	4.0562	4.3765
	p-value	0.0166^*	0.0059^*	0.0434^*	0.0012^*	0.0051^*	0.0065^*
Overall	R^2=0.9537						
n=73	Slope (β)	5.3967	3.5680	3.2683	3.4946	4.9084	3.7034
	p-value	$<0.0001^*$	0.0030^*	0.0027^*	$<0.0001^*$	$<0.0001^*$	0.0017^*

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; R^2 = degree of fit; β_i = partial slope; * $p < 0.05$, reject the null = the explanatory variable played a significant role in participants' user experience.

Table 4.19 shows that all the variables were significant ($p<0.05$), except for efficiency and engagement for the U-term dataset. The significant variables listed in order of magnitude of the partial slope parameter were as follows: effectiveness ($\beta_1 = 7.6641$), ease of use ($\beta_5 = 4.8609$), error tolerance ($\beta_4 = 4.0056$), and self-efficacy ($\beta_6 = 3.2277$). The results of the Q-term dataset showed that all variables were significant ($p<0.05$); efficiency ($\beta_2 = 5.1522$) > self-efficacy ($\beta_6 = 4.37646$) > ease of use ($\beta_5 = 4.0562$) > effectiveness ($\beta_1 = 3.8565$) > error tolerance ($\beta_4 = 3.5289$) > Engagement ($\beta_3 = 2.2258$). Likewise, the Overall dataset showed that all variables were significant ($p<0.05$); effectiveness ($\beta_1 = 5.3967$) > ease of use ($\beta_5 = 4.9084$) > self-efficacy ($\beta_6 = 3.70834$) > efficiency ($\beta_2 = 3.5680$) > error tolerance ($\beta_4 = 3.4646$) > engagement ($\beta_3 = 3.2683$). The magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott & Longnecker, 2016). As effectiveness (U-term and Overall datasets) and ease of use (U-term, Q-term, and Overall datasets), and self-efficacy (Q-term and Overall datasets) ranked high, these variables appeared to play a significant role in the AHHO user experience for Lab 2.

AHHO: Trial-1 + Trial-2. For the final analysis for the AHHO laboratory experience, the trial-1 and trial-2 MARIS values were combined for each variable. This dataset looked at the effect each variable had on the AHHO user experience. The R^2-value was fairly high (R^2 =0.94). This means that the model worked well as >94% of the variation was explained by the model. Moreover, an analysis of the residuals showed the assumptions of the multiple regression analysis were satisfied (i.e. homoscedasticity and normality) (Ott & Longnecker, 2016).

Table 4.20

Multiple Regression Analysis: AHHO – Trial 1 + Trial 2

$SUS_{AHHO} = \beta_0 + \beta_1 E_{AHHO} + \beta_2 F_{AHHO} + \beta_3 N_{AHHO} + \beta_4 ET_{AHHO} + \beta_5 EU_{AHHO} + \beta_6 \ SE_{AHHO}$

		E	F	N	ET	EU	SE
Overall R^2=0.9414	Slope (β) p-value	4.2633 <0.0001*	3.0225 <0.0001*	5.6718 <0.0001*	3.7333 <0.0001*	4.5412 <0.0001*	3.1340 <0.0001*

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; R^2 = degree of fit; β_i = partial slope; * $p < 0.05$, reject the null = the explanatory variable played a significant role in participants' user experience.

Table 4.20 shows that all the variables were significant ($p<0.05$). The variables listed in order of magnitude of the partial slope parameter were as follows: engagement ($\beta_3 = 5.6718$) > ease of use ($\beta_5 = 4.5412$) > effectiveness ($\beta_1 = 4.2633$) > error tolerance ($\beta_4 = 3.7333$) > self-efficacy ($\beta_6 = 3.1340$) > efficiency ($\beta_2 = 3.0225$). The magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott & Longnecker, 2016). As effectiveness (Trial-2 and Combined datasets), engagement (Trial-1 and Combined datasets), and ease of use (Trial-1, Trial-2, and Combined datasets) ranked high, these variables appeared to play a significant role in the AHHO user experience.

SUS Questionnaire: Sub-factors – VL (Research Question 2b). Table 4.21 shows the U-term, Q-term, and Overall results of the multiple regression analysis for the VL version of the Lab 1 protocol; Table 4.22 shows the U-term, Q-term, and Overall results of the multiple regression analysis for the VL version of the Lab 2 protocol; and Table 4.23 shows the Overall combined results (Trial 1 + Trial 2) of the multiple regression for VL. Each data table includes the following information: model (linear equation), R^2-value (degree of fit), partial slope parameter for each variable (β_i = change in perceived usability (SUS score) with the change in the explanatory variable while the other explanatory variables are held constant), and the p-value for the regression analysis (H_0: $\beta_i = 0$; H_a: $\beta_i \neq 0$; significance based upon $\alpha=0.05$).

VL: Trial-1. The degree of fit is determined by the R^2- value; the closer the value to 1, the better the fit. In all cases, the R^2-value was fairly high (>0.95). This means that the model worked well as >95% of the variation was explained by the model. Moreover, an analysis of the residuals (for each test) showed the assumptions of the multiple regression analysis were satisfied (i.e. homoscedasticity and normality) (Ott & Longnecker, 2016).

Table 4.21 shows that all the variables were significant ($p<0.05$), except for efficiency and engagement for the U-term dataset. The significant variables listed in order of magnitude of the partial slope parameter were as follows: ease of use ($\beta_5 = 7.6914$) >> self-efficacy ($\beta_6 =5.9996$) >> effectiveness ($\beta_1 = 2.9385$) > error tolerance ($\beta_4 = 2.7716$). The results of the Q-term dataset showed that ease of use ($\beta_5 =6.8132$), self-efficacy ($\beta_6 = 4.5985$), and error tolerance ($\beta_4 = 3.5010$) were significant ($p<0.05$). All of the variables were significant for the Overall dataset ($p<0.05$), except for engagement ($p=0.3076$); with ease of use ($\beta_5 = 6.6411$) > self-efficacy ($\beta_6 = 5.3410$) >> error tolerance ($\beta_4 = 2.9891$) > effectiveness ($\beta_1 = 2.3817$) > efficiency ($\beta_2 = 1.4732$). The magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott & Longnecker, 2016). As ease of use (U-term, Q-term, and Overall datasets) and self-efficacy (U-term, Q-term, and Overall datasets) ranked high, these variables appeared to play a significant role in the VL user experience for Lab 1.

Table 4.21

Multiple Regression Analysis: VL– Trial 1

$SUS_{VL1} = \beta_0 + \beta_1 E_{VL1} + \beta_2 F_{VL1} + \beta_3 N_{VL1} + \beta_4 ET_{VL1} + \beta_5 EU_{VL1} + \beta_6 SE_{VL1}$

		E	F	N	ET	EU	SE
U-term	$R^2=0.9551$						
n=33	Slope (β)	2.9385	1.7520	1.1544	2.7716	7.6914	5.9996
	p-value	0.0118^*	0.1171	0.6099	0.0424^*	0.0005^*	0.0046^*
Q-term	$R^2=0.9538$						
n=62	Slope (β)	1.3845	1.7202	0.8676	3.5010	6.8132	4.5985
	p-value	0.4798	0.0946	0.4961	0.0008^*	$<0.0001^*$	0.0031^*
Overall	$R^2=0.9561$						
n=95	Slope (β)	2.3817	1.4732	0.8767	2.9891	6.6411	5.3410
	p-value	0.0034^*	0.0251^*	0.3706	$<0.0001^*$	$<0.0001^*$	$<0.0001^*$

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; R^2 = degree of fit; β_i = partial slope; * $p < 0.05$, reject the null = the explanatory variable played a significant role in participants' user experience.

VL: Trial-2. The degree of fit is determined by the R^2- value; the closer the value to 1, the better the fit. In all cases, the R^2-value was fairly high (>0.95). This means that the model worked well as >95% of the variation was explained by the model. Moreover, an analysis of the residuals (for each test) showed the assumptions of the multiple regression analysis were satisfied (i.e. homoscedasticity and normality) (Ott & Longnecker, 2016).

Table 4.22

Multiple Regression Analysis: VL– Trial 2

$$SUS_{VL2} = \beta_0 + \beta_1 E_{VL2} + \beta_2 F_{VL2} + \beta_3 N_{VL2} + \beta_4 ET_{VL2} + \beta_5 EU_{VL2} + \beta_6 SE_{VL2}$$

		E	F	N	ET	EU	SE
U-term	R^2=0.9828						
n=34	Slope (β)	2.2394	-0.8103	5.2620	5.9148	6.9948	2.3421
	p-value	0.1192	0.4386	0.0013*	0.0064*	0.0007*	0.0777
Q-term	R^2=0.9676						
n=39	Slope (β)	3.1231	2.9034	4.7542	3.6004	4.3924	4.5561
	p-value	0.1257	0.0764	0.0038*	0.0111*	0.0034*	0.0150*
Overall	R^2=0.9494						
n=73	Slope (β)	1.9598	1.9269	4.9915	3.3031	6.5068	3.6950
	p-value	0.1526	0.0776	<0.0001*	0.0010*	<0.0001*	<0.0001*

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; R^2 = degree of fit; β_i = partial slope; * $p < 0.05$, reject the null = the explanatory variable played a significant role in participants' user experience.

Table 4.22 shows that ease of use (β_5 =6.9948), error tolerance (β_4 = 5.9148), and engagement (β_2 = 5.2620) were significant ($p<0.05$) for the U-term dataset. The Q-term dataset shows that all of the variables were significant ($p<0.05$), except for effectiveness and efficiency. The significant variables listed in order of magnitude of the partial slope parameter were as follows: engagement (β_2 = 4.7542) > self-efficacy (β_6 =4.5561) > ease of use (β_5 = 4.3924) > error tolerance (β_4 = 3.6004). All of the variables were significant for the Overall dataset ($p<0.05$), except for effectiveness (p=0.1526) and efficiency (p=0.0776); with ease of use (β_5 = 6.5068) > engagement (β_3 = 4.9915) > self-efficacy (β_6 = 3.6950) > error tolerance (β_4 = 3.3031) The magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott & Longnecker, 2016). As ease of use (U-term, Q-term, and Overall datasets), engagement (U-term, Q-term, and Overall datasets), and error tolerance (U-term, Q-term, and Overall datasets) ranked high, these variables appeared to play a significant role in the VL user-experience for Lab 2.

VL: Trial-1 + Trial-2. For the final analysis for the VL laboratory experience, the trial-1 and trial-2 MARIS values were combined for each variable. This dataset looked at the effect each variable had on the VL user-experience. The R^2-value was fairly high (R^2 =0.95). This means that the model worked well as >95% of the variation was explained by the model. Moreover, an analysis of the residuals showed the assumptions of the multiple regression analysis were satisfied (i.e. homoscedasticity and normality) (Ott & Longnecker, 2016).

Table 4.23

Multiple Regression Analysis: VL– Trial 1 + Trial 2

$SUS_{VL} = \beta_0 + \beta_1 E_{VL} + \beta_2 F_{VL} + \beta_3 N_{VL} + \beta_4 ET_{VL} + \beta_5 EU_{VL} + \beta_6 SE_{VL}$

		E	F	N	ET	EU	SE
Overall R^2=0.9470	Slope (β)	1.7575	2.0755	4.2078	2.6078	6.6446	4.6876
	p-value	0.0254*	0.0015*	<0.0001*	<0.0001*	<0.0001*	<0.0001*

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; R^2 = degree of fit; β_i = partial slope; * $p < 0.05$, reject the null = the explanatory variable played a significant role in participants' user experience.

Table 4.23 shows that all the variables were significant ($p<0.05$). The variables listed in order of magnitude of the partial slope parameter were as follows: ease of use ($\beta_5 = 6.446$) >> self-efficacy ($\beta_6 = 4.6876$) > engagement ($\beta_3 = 4.2078$) >> error tolerance ($\beta_4 = 2.6078$) > efficiency ($\beta_2 = 2.0755$) > effectiveness ($\beta_1 = 1.7575$). The magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott & Longnecker, 2016). As ease of use (Trial-1, Trial-2, and Combined datasets), engagement (Trial-2 and

Combined datasets), and self-efficacy (Trial-1 and Combined datasets) ranked high, these variables appeared to play a significant role in the VL user experience.

Multiple Regression Analysis: Summary. Table 4.24 shows a summary of the multiple regression analysis results.

Table 4.24

Summary: Multiple Regression Analysis

$SUS = \beta_0 + \beta_1 E + \beta_2 F + \beta_3 N + \beta_4 ET + \beta_5 EU + \beta_6 SE$

Variables	$AHHO_1$	VL_1	$AHHO_2$	VL_2	AHHO	VL	LE
E (β_1)	4.0050	2.3817	**5.3967**	1.9598^{NS}	**4.2633**	1.7575	2.9813
F (β_2)	2.4416	1.4732	3.5680	1.9269^{NS}	3.0225	2.0755	2.4924
N (β_3)	**5.2166**	0.8767^{NS}	3.2683	**4.9915**	**5.6718**	**4.2078**	**4.2222**
ET (β_4)	4.0207	2.9891	3.4946	3.3031	3.7333	2.6078	3.2277
EU (β_5)	**4.4073**	**6.6411**	**4.9084**	**6.5068**	**4.5412**	**6.6446**	**5.6746**
SE (β_6)	2.2977	**5.3410**	3.7034	3.6950	3.1340	**4.6876**	**3.8683**

Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience; LE = laboratory experience (AHHO + VL); $_1$ = Trial-1; $_2$ = Trial-2; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; β_i = partial slope; ^{NS} = the explanatory variable did not play a significant role in participants' user experience ($p > 0.05$ for the partial slope parameter, failed to reject the null); **bold** = large magnitude, indicates a large effect on the user experience.

The partial slope parameters for Trial-1 ($AHHO_1$ and VL_1) and Trial 2 ($AHHO_2$ and VL_2) reflect the Overall dataset; the slopes for AHHO and VL reflect the combined datasets (Trial-1 + Trail-2). The summary table indicates that effectiveness, engagement, and ease of use play the most important role in the AHHO laboratory experience; and engagement, ease of use, and self-efficacy play the most important role in the VL laboratory experience. It appears that ease of use, engagement, and sell-efficacy play the most important role in participants' laboratory experience (for both AHHO and VL). This was further supported by the regression analysis of users' laboratory experience (LE) with the combined MARISs (AHHO + VL).

Phase I: Summary – SUS, MRIS, MARIS, and Regression Analysis. In Phase I (quasi-experimental crossover phase), the first two questions on the Lab 1 Survey (trial-1) and Lab 2 Survey (trial-2) were used to gather data on perceived usability (SUS score) and its sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy) for the AHHO and VL laboratory experiences. A modified version of Brooke's (1996) SUS questionnaire was used to determine if adult learners in an online, undergraduate, science course for non-majors would prefer an at-home, hands-on or an at-home, virtual laboratory experience (RQ 1-1).

The SUS score was calculated for each participant using the method outlined in Sauro and Lewis (2016). The mean SUS scores for $AHHO_1$, VL_1, $AHHO_2$, VL_2, AHHO, and VL were calculated and analyzed using SAS OnDemand for Academics. The results of the two-sample, t-test (H_0: $\mu_{AHHO} - \mu_{VL} = 0$; H_a: $\mu_{AHHO} - \mu_{VL} \neq 0$) and the non-parametric Wilcoxon rank sum test (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$) showed that the perceived usability was nearly the same for AHHO and VL for Trial-1, Trial-2, and Combined (Trial-1 + Trial-2) datasets (Table 4.25). Although the findings were inconclusive for preferred laboratory experience, the conclusion of 'nearly the same' is consistent with the literature.

Table 4.25

Summary: SUS Scores – Trial 1, Trial 2, and Combined (Trial-1 + Trial-2)

	SUS Score	Grade	User-Experience (based on raw numbers)	User-Experience (based on hypothesis test)
AHHO$_1$	69.039	C	VL$_1$ > AHHO$_1$	VL$_1$ ≈ AHHO$_1$
VL$_1$	75.778	B		No significant difference
AHHO$_2$	60.329	D	AHHO$_2$ > VL$_2$	AHHO$_2$ ≈ VL$_2$
VL$_2$	55.294	D		No significant difference
AHHO	63.3611	C-	VL > AHHO	VL ≈ AHHO
VL	66.9620	C		No significant difference

Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience; $_1$ = Trial-1; $_2$ = Trial-2. The assigned grade is based upon the curved grading scale developed by Sauro and Lewis (2016).

Based upon Lewis and Sauro’s (2019) description of item attributes, each of the SUS items were coded as a measure of effectiveness, efficiency, engagement, error tolerance, ease of use, or self-efficacy. Each variable corresponded to both a positive statement and negative statement, with the exception of error tolerance (negative statement only) and self-efficacy (positive statement only). The mean raw item scores (MRIS) and the mean adjusted raw item scores (MARIS) were used to determine if there was evidence to support that one or more factors played a significant role in the user experience (RQ 1-1’, RQ2a, and RQ2b).

Like the SUS score analysis, Table 4.26 shows there is agreement between the MRIS and MARIS results for effectiveness (positive user experience for both AHHO and VL); the user experience is more positive for VL than AHHO for efficiency, engagement, error tolerance, ease of use, and self-efficacy. However, the findings are inconclusive for preferred laboratory experience as the difference is not significant.

Table 4.26

Summary: SUS Sub-factor Analysis – MRIS, MARIS, and Multiple Regression

V	Analysis	AHHO$_1$	VL$_1$	AHHO$_2$	VL$_2$	AHHO	VL	UX
E	MRIS	SUS=68	SUS=68	SUS=68	SUS=68	SUS=68	SUS=68	Inc. (+)
	MARIS	>2.5	>2.5	>2.5	≈2.5	>2.5	>2.5	Inc. (+)
	MR	S	S	S, β	NS	S, β	S	Inc.
F	MRIS	Not met	SUS=68	Not met	Not met	Not met	Not met	Inc. (-)
	MARIS	>>2.5	>>2.5	<2.5	<2.5	>2.5	>2.5	Inc. (+)
	MR	S	S	S	NS	S	S	Inc.
N	MRIS	SUS=68	**SUS=68**	**Not met**	Not met	Not met	SUS=68	Inc. (mixed)
	MARIS	>2.5	**>2.5**	**<2.5**	<<2.5	<2.5	>2.5	Inc. (mixed)
	MR	S, β	NS	S	S, β	S, β	S, β	Inc.
ET	MRIS	Not met	Not met	Not met	Not met	Not met	Not met	Inc. (-)
	MARIS	>2.5	>2.5	<2.5	<2.5	<2.5	>2.5	Inc. (mixed)
	MR	S	S	S	S	S	S	Inc.
EU	MRIS	SUS=68	**SUS=68**	Not met	Not met	Not met	SUS=68	Inc. (mixed)
	MARIS	>2.5	>>2.5	<2.5	<2.5	>2.5	>2.5	Inc. (+)
	MR	S, β	S, β	S, β	S, β	S, β	S, β	Inc.
SE	MRIS	SUS=68	SUS=68	Not met	Not met	Not met	Not met	Inc. (-)
	MARIS	>>2.5	>>2.5	<<2.5	<<2.5	>2.5	>2.5	Inc. (+)
	MR	S	S, β	S	S	S	S, β	Inc.

Note. V=variable; AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience; $_1$ = Trial-1; $_2$ = Trial-2; UX = user experience; E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; MRIS =

data from mean raw item score analysis; MARIS = data from mean adjusted raw item score analysis; MR = data from multiple regression analysis; SUS=68 = met the SUS=68 benchmark (average user experience); Not met = did not meet the SUS=68 benchmark (below average user experience); >2.5 (average-above average user experience); <2.5 = below average user experience; S = significant; NS = Not significant; Inc. = results inconclusive for preferred laboratory experience; (+) = mainly positive user experience; (-) = mainly negative user-experience; (mixed) = mixed positive/negative user experience; β = large partial slope parameter; **bold** = significant difference between AHHO/VL ($p<0.05$), bolded term is larger value.

The summary table also indicates that there is agreement between the MARIS and multiple regression results (large β corresponds with MARIS > 2.5, with the exception of engagement). As such, it appears that effectiveness and ease of use play the most important role in the AHHO laboratory experience; and engagement, ease of use, and self-efficacy play the most important role in the VL laboratory experience. The question of which factors play a significant role in the user experience was further explored in the qualitative strand of the study.

Phase II: Self-Selection Phase

Description of the Laboratories

Students were given the opportunity to self-select between an at-home, hands-on laboratory experience and an at-home, virtual laboratory experience. While the hands-on and virtual laboratory protocols were both health-related in content and satisfied the same objectives (relationship between correlation and causation), the procedures and pre- and post-lab questions were not the same for each laboratory.

In the hands-on version of Laboratory 4, students used balloons to measure their tidal volume (volume of air moved with a normal respiratory cycle) and vital capacity (maximum volume of air moved following deep inhalation and forced exhalation). Students then compared their respiratory volumes to the standard values, identified any confounding factors that may have affected their measured values (respiratory illness, smoker, athlete, etc.), and discussed whether these factors correlated with abnormal respiratory volumes. Finally, they were asked to determine if there was enough evidence to establish causation (i.e. that the identified factors/conditions were the cause of the differences of their respiratory volumes from standard values).

In the virtual version of Laboratory 4, students were provided with a graph depicting a positive correlation between organic food sales and the prevalence of autism. They were asked perform background research on ‘organic food’, ‘autism’, and to identify any confounding or spurious factors that may have contributed to this relationship. The students were then asked to outline the steps of an experiment that could be performed to test this relationship and discuss what types of evidence would be needed to determine that increased sales of organic food caused the increased prevalence of autism.

Identification of the Laboratory Experience

The first question on the Lab 4 Survey asked participants to identify which version of the laboratory that they completed. This was accomplished by a simple description of the materials used for the laboratory. Participants’ responses were used to identify which perception data should be assigned to the AHHO (“utilized balloons to measure tidal volume and vital capacity”) and VL (“utilized research to examine the relationship between organic food sales and the prevalence of autism”) experiences.

Participant Data

A total of 270 students completed Laboratory 4. All students were given the opportunity to complete the Lab 4 Survey; 49 students (n_U=24; n_Q=25) responded. Table 4.27 shows the number participants (and their laboratory selection) for the U-term, Q-term, and Overall.

Table 4.27

Self-Selection Data: Number of Participants and Their Laboratory Selection

Laboratory Experience	U-term	Q-term	Overall
$AHHO_{raw}$	n=5 (2 removed)	n=12 (4 removed)	n=17 (6 removed)
$AHHO_{cleaned}$	n=3	n=8	n=11
VL_{raw}	n=16 (removed 1)	n = 16 (0 removed)	n=32 (1 removed)
$VL_{cleaned}$	n=15	n=16	n=31

Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience.

It should be noted that 7 participants were removed from the dataset based upon their response to Question 2, “Why did you choose to complete this laboratory exercise?". It seems one or more of the instructors assigned Lab 4 rather than allowing students to choose the version of the lab they wished to complete. Two (2) U-term and 4 Q-term respondents indicated they were assigned the AHHO version of Lab 4; one (1) U-term participant indicated that they were assigned the VL version of Lab 4. As this would skew the self-selection results, these respondents were removed from dataset for the quantitative analysis only (they were included in the qualitative analysis for questions 4-6).

Comparison of Proportions AHHO versus VL (RQ1-2)

The 'cleaned' dataset was used to determine the proportions of respondents that self-selected the AHHO and VL experiences. A two-proportion Z-test ($\alpha=0.05$) was used to determine if there was a significant difference between the proportions (H_0: $\hat{p}_{AHHO} = \hat{p}_{VL}$; H_a: $\hat{p}_{AHHO} \neq \hat{p}_{VL}$). The test statistic was calculated for the Overall dataset only; the sample sizes for the U-term data ($n_{AHHO}=3$; $n_{VL}=15$) and Q-term data ($n_{AHHO} =8$; $n_{VL} = 16$) were too small for an accurate determination. The test statistic ($Z_{test} = 2.8125$) was greater than the critical value ($Z=1.9600$); therefore, we rejected the null hypothesis. This means that the difference in proportions was significant. As the $\hat{p}_{VL} > \hat{p}_{AHHO}$; the self-selection data indicated that participants preferred the virtual laboratory experience.

However, the proportions were adjusted based on responses to Question 2 ("Why did you choose to complete this laboratory exercise?") and Question 3 ("Was there something about the other laboratory exercise that made it less appealing to you?"). There were 11 respondents (5 U-term; 6 Q-term) that indicated they would have preferred to perform the AHHO version of the laboratory, but did not have access to the materials for the lab experiment. For example, one respondent stated*, "I chose this lab exercise because I did not have all of the materials needed for the hands on. I would have preferred to complete the hands-on versus the virtual one as I am not a big fan of doing a lot of research".* Another said, *"I did not have the items available to do the home lab, so I did the virtual. I actually thought the hands-on lab might have been easier to complete"*. Finally, another indicated, *"Didn't have time to go get what was needed".*

After adjusting the numbers (by +11 for AHHO; -11 for VL), the test statistic for the adjusted proportions ($Z_{test} = 0.3018$) was less than $Z=1.9600$ (critical value); therefore, we fail to reject the null hypothesis. This means that the difference in proportions is not significant. As $\hat{p}_{AHHO} = \hat{p}_{VL}$, the self-selection data was inconclusive. Table 4.28 shows the results of the Z-test.

Table 4.28

Summary of Results for two-proportion Z-test

	n_i	$\hat{p}_i$	Z_{test}	Conclusion
$VL_{cleaned}$	31	0.7381	2.8125	2.8125 > 1.9600 (critical value)
$AHHO_{cleaned}$	11	0.2619		Reject H_0; $\hat{p}_{AHHO} \neq \hat{p}_{VL}$
	total = 42			Preferred VL
$VL_{adjusted}$	20	0.4762	0.3081	0.3018 < 1.9600 (critical value)
$AHHO_{adjusted}$	22	0.5238		Fail to reject H_0; $\hat{p}_{AHHO} = \hat{p}_{VL}$
	total = 42			Nearly the same

Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience. The 'adjusted' numbers were calculated by adding 11 to AHHO and subtracting 11 from the VL.

The adjusted Phase II findings were consistent with the Phase I findings (SUS score and sub-factor analysis). As these data all arrived at the same conclusion (AHHO≈VL), this supports the use of the SUS as an effective tool to measure perceived usability of a laboratory experiment. Moreover, the Phase I and Phase II results of 'nearly the same' are also consistent with the literature. The next section moves from the quantitative findings to the qualitative findings.

Qualitative Findings

The surveys for the crossover phase (Lab 1 Survey and Lab 2 Survey) included open-ended reflection questions modified from Reck et al. (2019) (with permission, Appendix G); the open-ended reflection questions for the self-selection phase were self-developed based upon an extensive review of the literature (Attardi et al., 2015; Meintzer et al., 2017; Nolen & Koretsky, 2018; Pal & Vanijja, 2020; Reck et al., 2019; Stuckey-Mickell & Stuckey-Danner, 2007). The results of the Phase I analysis are presented in the next section.

Phase I: Quasi-experimental Crossover Phase

Crossover Phase: Hypothesis and Emergent Coding

Hypothesis coding was used in the first round of data analysis. This means that a pre-determined set of codes (based upon factors identified in the research questions) were initially applied to the responses to the open-ended reflection questions (Creswell & Creswell, 2018). The pre-determined codes utilized in this study were the SUS sub-factors tested in the quantitative strand; effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy. Like the alternating positive- and negative-tone for the items in the SUS, each response was coded as either a positive (+) or negative (-) statement for its associated variable.

Many of the responses were assigned multiple codes. For example, engagement (-), ease of use (-), and efficiency (-) was applied to the following response: *"I didn't really enjoy using a virtual ruler. I had to take a picture of an object and then send it to my email. I then had to upload it to the virtual ruler. I then had to measure it digitally. It would have been a lot easier just to use a real ruler and quickly measure the object."* Table 4.29 illustrates the alignment between the SUS items, ± sub-factors (hypothesis codes) used to code the responses, and some sample quotes from the Lab 1 and Lab 2 Surveys.

Table 4.29

Definitions Used for Hypothesis Coding

Code	SUS Item *Sample Quote from Survey*	Code	SUS Item *Sample Quote from Survey*
E+	I found the various parts of this lab experiment to be well integrated.	E-	I thought there was too much inconsistency in this lab experiment.
	“The questions were clear and easy to answer.”		*“Had to use my imagination when instructions and worksheet were in conflict.”*
F+	I would imagine that most people would be able to do lab experiments like this very quickly.	F-	I needed to learn a lot of things before I could do this experiment.
	“Easy and fast to do so it fit with my schedule		*“I thought it would be easy but it wasn't and it made the lab take a lot longer to complete”*
N+	I think I would like to do more lab experiments like this one.	N-	I found this lab experiment very awkward to do.
	“As a adult learner…this lab was an excellent way to learn practical information while still completing the requirements in a remote setting.”		*“It was a little akward at first thinking that I was measuring a couch that was only 8 cm, but once you get past that it was easy.”*
ET+	No SUS Item	ET-	I think I would need more support to be able to do more lab experiments like this.
	“I was a little confused at first on what to do…Once I got it, it was fine.”		*“I felt I needed a little more direction, but It may be different if I was in an in person class.”*
EU+	I thought this lab experiment was easy to do.	EU-	I found this lab experiment unnecessarily complex.
	“The simplicity and ease of use.”		*“It was a little challenging trying to figure out the virtual aspect”*
SE+	I felt very confident doing this lab experiment.	SE-	No SUS item
	“….everything fell into place and made my confidence grow.”		*“I just wasn't sure if i was doing it correctly.“*

Note. E=effectiveness; F=efficiency; N=engagement; ET=error tolerance; EU=ease of use; SE=self-efficacy. The underlined portions of the sample responses (from Lab 1 and Lab 2 Surveys) align with the positive (+) or negative (-) aspects of the code.

Emergent coding was used for the second round of data analysis. All responses were re-evaluated for recurring and unexpected themes that emerged from the data analysis. For example, the previous quote, *"I didn't really enjoy using a virtual ruler. I had to take a picture of an object and then send it to my email. I then had to upload it to the virtual ruler. I then had to measure it digitally. It would have been a lot easier just to use a real ruler and quickly measure the object."*, was also assigned emergent codes 'materials', 'data collection', and 'prefer physical manipulatives'. Other themes that emerged from the data included: 'math' (>70% of the responses mentioned math, calculations, or statistics), "*..the coin tosses were easy, however, I need a refresher on the math side of it";* 'clarity of instructions' (≈40% of the responses mentioned either the clarity or lack of clarity of the instructions), *"The instructions were not 100% clear";*, and 'kids' (several of the respondents mentioned that their kids helped them with data collection), *"I thought the lab was very clear and easy. I actually did it with my boys. They had fun with it"*). The results of the analysis for each open-ended question for trial-1 and trial-2 is presented in the next section.

Question 3: Trial-1 and Trial-2

Question 3 (Q3) on the Lab 1 Survey and Lab 2 Survey asked participants to reflect on "What aspects of this laboratory experience met your expectations?". For trial-1 (n=98), 18 respondents (18.4%) chose not to answer Q3; 5 respondents (5.1%) indicated that Lab 1 met all expectations; and 12 (12.2%) indicated that they did not know what to expect for or had no expectations of the laboratory experience. Similarly, for trial-2 (n=73), 14 respondents (19.2%)

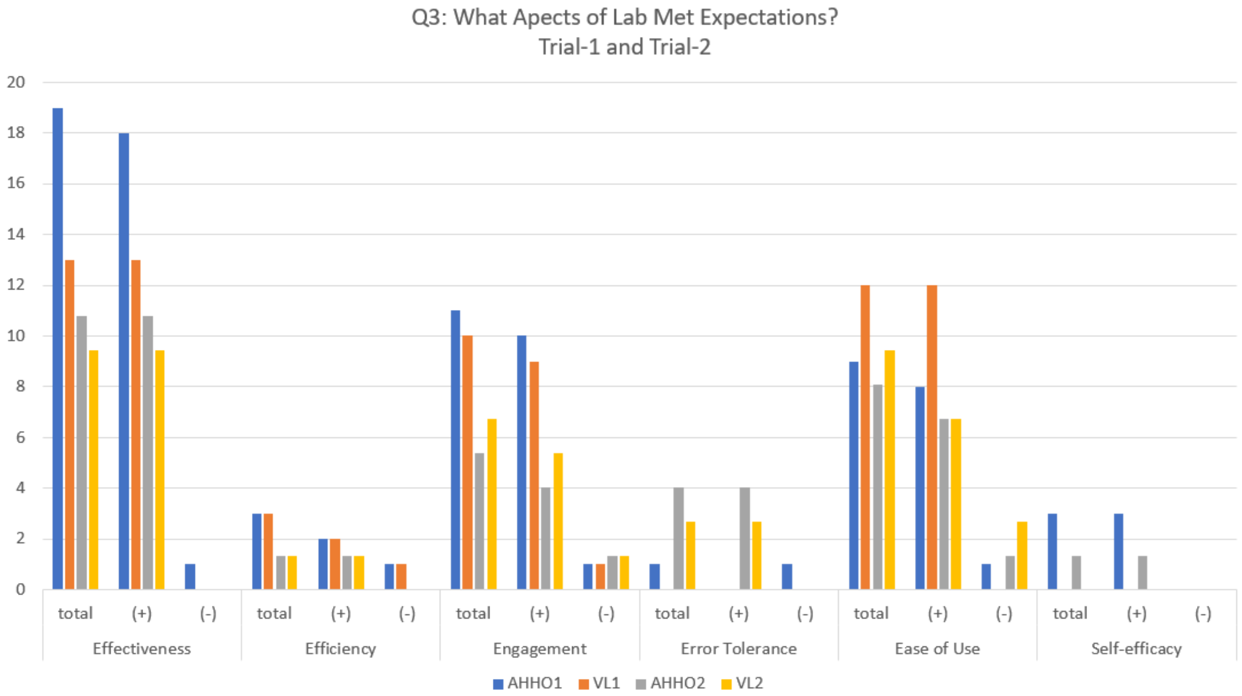
chose not to answer Q3; 2 respondents (2.7%) indicated that Lab 2 met all expectations; and 5 (6.8%) had no expectations. Additionally, 2 respondents (2.7%) that completed the AHHO and 3 respondents (4.1%) that completed the VL indicated that nothing about Lab 2 met their expectations. These negative reactions were consistent with the quantitative findings; the SUS, MRIS, and MARIS data all indicated a below average user experience for Lab 2.

Likewise, the results of the hypothesis coding for Q3 were also consistent with the quantitative findings. The MARIS and multiple regression analysis identified that effectiveness, engagement, and ease of use played a significant role in the laboratory experience; the same three sub-factors emerged from the responses to Q3. Furthermore, the results from Q3 show that the Lab 1 user-experience was more positive than the Lab 2 user experience. Figure 4.9 shows the results for the hypothesis coding for this question.

It was interesting that (while the differences were not significant) the MRIS and MARIS for $VL_1 > AHHO_1$ for all sub-factors; however, the participants' responses to Q3 seem to indicate that $AHHO_1 > VL_1$ for effectiveness and engagement. It seems that the coin-flip simulator became annoying after a certain number of flips, “*it was a little daunting to flip it online 50+ times, I wish there was a button or number you could type in so it could automatically do it. Overall it was not bad, but just difficult if you needed to do it more than 25 times”*. The participants that completed the hands-on version of the lab were able to break up the monotony by involving their children in the lab (‘*I actually did [the lab] with my boys. They had fun with it”*), chasing the coin(s) when it landed “*off the surface of my desk and onto the floor”*, and “*trying to flip the coin the same way each time*”. These responses were also associated with the emergent codes ‘data collection’ (daunting to flip coin 50+ times); ‘kids’ (did lab with my boys); and ‘materials’ (chasing coins and flipping real/virtual coins).

Figure 4.9

Hypothesis Coding: Question 3 – Trial-1 and Trial-2



Note. $AHHO_1$ = at-home, hands-on laboratory experience, trial-1 (n=45); $AHHO_2$ = at-home, hands-on laboratory experience, trial-2 (n=39); VL_1 = at-home, virtual laboratory experience, trial-1 (n=53); VL_2 = at-home, virtual laboratory experience, trial-2 (n=34). The bar graph shows the total number of responses for each sub-factor and the number of responses that corresponded to the positive (+) and negative (-) aspects of the hypothesis code; the number of trial-2 responses were scaled by a factor of 1.35 to allow for comparison across trials.

Respondents (for both $AHHO_2$ and VL_2) did not seem to enjoy Lab 2. The following quote captures the overall frustration with the laboratory experience, "*I absolutely hated this lab. I hate that there are things that we have to do that take up so much time. I also had to try to learn how to do these formulas on my own and I am not even sure that I did it correctly….Math is horrid and this class does not meet my expectations of what I thought I would be doing in a* *science class"*. (The underlined parts of the statement corresponded to engagement (-), efficiency (-), error tolerance (-), self-efficacy (-), and math).

More than 70% of the responses indicated that 'math' (the number of calculations and the calculations themselves) were a huge dissatisfier; ≈30% of the responses indicated that they did not see the 'relevance' of measuring objects; ≈30% indicated that they did not know how to (or were not confident in their ability to) correctly read a ruler ('measurements'); and that the lab took longer to complete than anticipated because of issues with the ruler/calculations. This was also reflected in the responses to question 4.

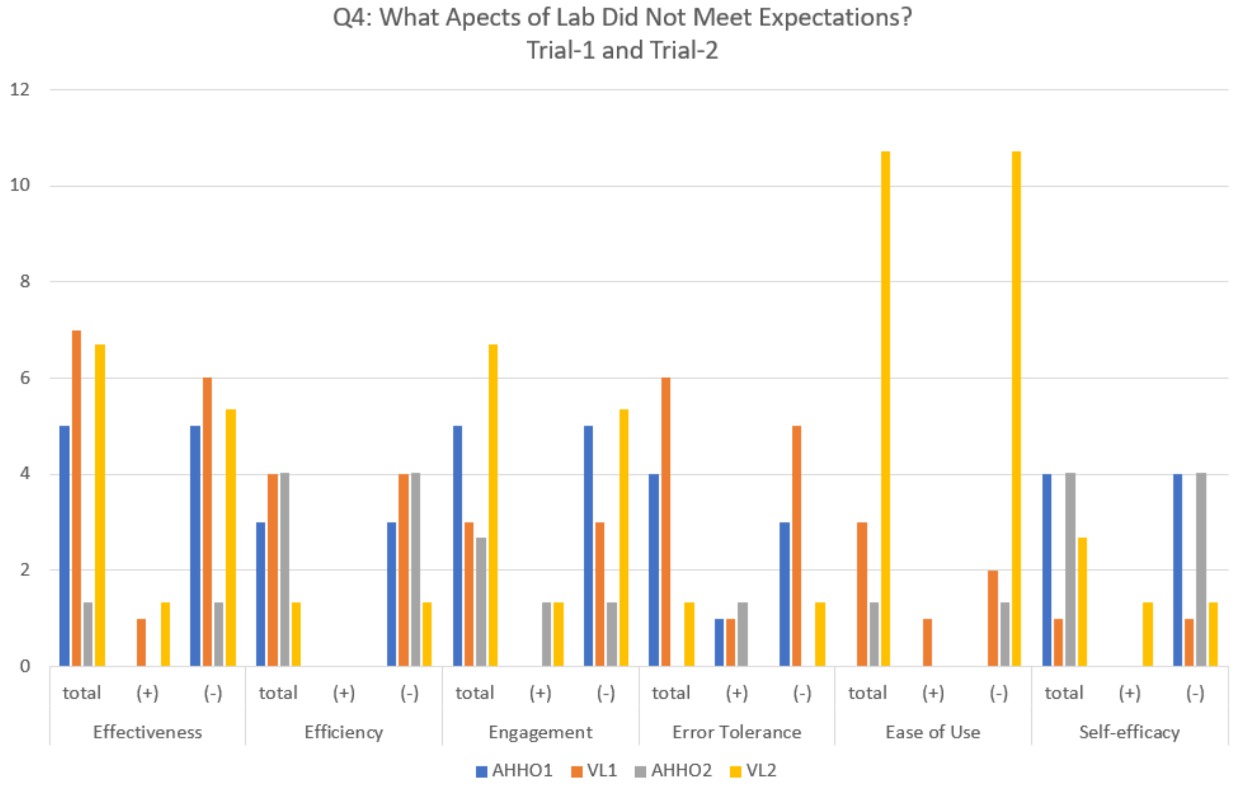
Question 4: Trial-1 and Trial-2

Question 4 (Q4) on the Lab 1 Survey and Lab 2 Survey asked participants to reflect on "What aspects of this laboratory experience did not meet your expectations?". For trial-1 (n=98), 19 respondents (19.4%) chose not to answer Q4; 29 respondents (29.6%) indicated that Lab 1 met all expectations; and 7 (7.1%) indicated that they did not know what to expect for or had no expectations of the laboratory experience. Similarly, for trial-2 (n=73), 14 respondents (19.2%) chose not to answer Q4; 15 respondents (20.5%) indicated that Lab 2 met all expectations; and 4 (5.5%) had no expectations. Additionally, 2 respondents (2.7%) that completed the AHHO and 1 respondent (1.4%) that completed the VL indicated that nothing about Lab 2 met their expectations. It should be noted there were fewer responses to code for Q4 due to the increased non-response rate and a large number of 1-2-word responses (i.e. none, all, or no expectations).

As Q4 was the negative foil to Q3, the results of the hypothesis coding for Q4 were fairly consistent with Q3 (and the quantitative findings). The MARIS, multiple regression analysis, and Q3 identified that effectiveness, engagement, and ease of use played a significant role in the laboratory experience; the same three sub-factors emerged from the responses to Q4. Error tolerance (or need for more support) also emerged as an important factor for the Lab 1 experience. Although respondents indicated they would have liked more "*assistance with the lab questions, a little more support*", the Lab 1 user experience was more positive than the Lab 2 user experience. Figure 4.10 shows the results for the hypothesis coding for this question.

Figure 4.10

Hypothesis Coding: Question 4 – Trial-1 and Trial-2



Note. AHHO$_1$ = at-home, hands-on laboratory experience, trial-1 (n=45); AHHO$_2$ = at-home, hands-on laboratory experience, trial-2 (n=39); VL$_1$ = at-home, virtual laboratory experience, trial-1 (n=53); VL$_2$ = at-home, virtual laboratory experience, trial-2 (n=34). The bar graph shows the total number of responses for each sub-factor and the number of responses that corresponded to the positive (+) and negative (-) aspects of the hypothesis code; the number of trial-2 responses were scaled by a factor of 1.34 to allow for comparison across trials.

The responses to Q4 helped to clarify many of the issues participants encountered during the Lab 1 experience (‘math’, ‘clarity of instructions’, and ‘data collection’) and Lab 2 (‘math’, ‘clarity of instructions’, ‘relevance’, and ‘measurements’). The emergent code ‘clarity of instructions’ was strongly tied to ‘math’ (and was also tied to error tolerance). Although the laboratory instructions provided the definitions, formulas, and example calculations, respondents largely indicated that the instruction sheets were not clear enough and that they needed more support as they were “*unprepared for this level of math*”.

‘Data collection’ also emerged as an issue for Lab 1. A number of respondents indicated that they had trouble keeping an accurate tally of the heads/tails results; respondents also mentioned that the large number of coin flips was exhausting. These issues are exemplified by the following quotes: “*I was unsure of about the results as I started getting tired of flipping the coins”, “I understand why the amount of tosses was needed, it just became cumbersome”,* and *“The amount of times I had to flip the coin was pretty exhausting”*. It is interesting that keeping an accurate tally was identified as an issue that detracted from the overall laboratory experience in Q4, however, quite a few respondents (n=7) identified keeping tally of the results was an enjoyable aspect of the laboratory experience in question 5.

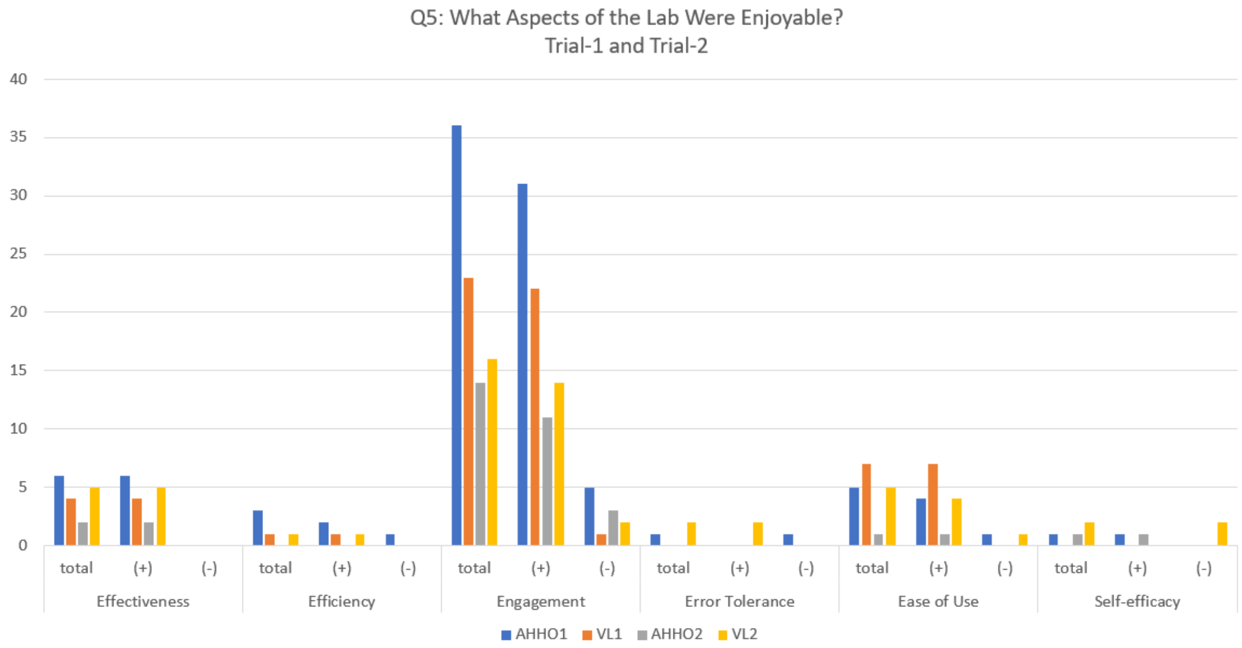
Question 5: Trial-1 and Trial-2

Question 5 (Q5) on the Lab 1 Survey and Lab 2 Survey asked participants to reflect on “What aspects of this laboratory experience did you find enjoyable?”. For trial-1 (n=98), 14 respondents (14.3%) chose not to answer Q5; 3 respondents (3.1%) indicated that they enjoyed everything about Lab; and 4 (4.1%) indicated that they did not enjoy anything about the Lab 1 experience. 9.2%) Similarly, for trial-2 (n=73), 15 respondents (20.5%) chose not to answer Q5; 3 respondents (4.1%) indicated that they enjoyed everything about Lab 2; and 4 (5.5 %)

indicated that they did not enjoy anything about the Lab 2 experience. Figure 4.11 shows the results of the hypothesis coding for this question.

Figure 4.11

Hypothesis Coding: Question 5 – Trial-1 and Trial-2



Note. AHHO$_1$ = at-home, hands-on laboratory experience, trial-1 (n=45); AHHO$_2$ = at-home, hands-on laboratory experience, trial-2 (n=39); VL$_1$ = at-home, virtual laboratory experience, trial-1 (n=53); VL$_2$ = at-home, virtual laboratory experience, trial-2 (n=34). The bar graph shows the total number of responses for each sub-factor and the number of responses that corresponded to the positive (+) and negative (-) aspects of the hypothesis code; the number of trial-2 responses were scaled by a factor of 1.44 to allow for comparison across trials.

Like Q3 and Q4 (and the quantitative findings), effectiveness, engagement, and ease of use emerged as the primary factors that played an important role in the laboratory experience from the responses to Q5. Again, the findings show that the Lab 1 user experience was more positive than the Lab 2 user experience; and (like Q3, but unlike the quantitative findings) effectiveness and engagement for $AHHO_1 > VL_1$. As this question was focused on aspects that participants enjoyed, it is not surprising that the bulk of the responses were coded for engagement. Table 4.30 provides a representative sample of the engagement responses.

Table 4.30a

Summary of Engagement Responses – Trial 1

	Engagement (+)	Engagement (-)
$AHHO_1$	*"I enjoy doing little experiments. As a much older student (55), when I went to college after high school and got my 2 year degree, everything was done in class, and there were "lab" times for many of the classes. Doing schooling online is definitely a different experience for me, but adding little labs like this break up the constant "online" work."*	*"While I generally would rather conduct experiments by hands-on means. This lab would have been much more enjoyable by virtual means. So with that being said, this lab was not really enjoyable. As previously mentioned, I found the task of flipping the coins to be tedious."*
	"My 9 year old daughter helping me out with the coin flip"	*"This lab was not enjoyable. I spent most of my time having to look for YouTube tutorials on how to calculate the probability."*
	"I enjoyed tracking the data."	*"I understood the purpose, however did not enjoy the math"*
VL_1	*"The user-friendly application and readily available demonstration of the module material into a practical exercise."*	*I am not a fan of math, so my favorite part was completing the lab assignment*
	"The idea of being able to do a lab experiment virtually was different and made me go into the experiment with a positive attitude. The simplicity of clicking to flip the coin was enjoyable"	
	"I enjoyed getting the data and using the simulator was easy".	

Note. $AHHO_1$ = at-home, hands-on laboratory experience, trial-1; VL_1 = at-home, virtual laboratory experience, trial-1.

Table 4.30b

Summary of Engagement Responses – Trial 2

	Engagement (+)	Engagement (-)
$AHHO_2$	*"Going around my house to find the best object made me feel like I was really doing an experiment."* *"I found it enjoyable to understand how measuring can be more complex then I originally thought."* *"The hands on experience was pleasant."*	*"Didn't really enjoy it at all"* *"I don't really care to do lab experiments"*
VL_2	*"It was enjoyable to be able to pick out an image to have as the background."*	*"There was not much of this experience I found enjoyable for the simple fact I wasn't sure if I was doing the lab right throughout completing it."*
	"Finding the objects around my house and using the digital tool to measure them. It was new and interesting as I have never done that before."	*"Science labs are not something I would consider enjoyable. I approached it as something that had to be completed."*

Note. $AHHO_2$ = at-home, hands-on laboratory experience, trial-2; VL_2 = at-home, virtual laboratory experience, trial-2.

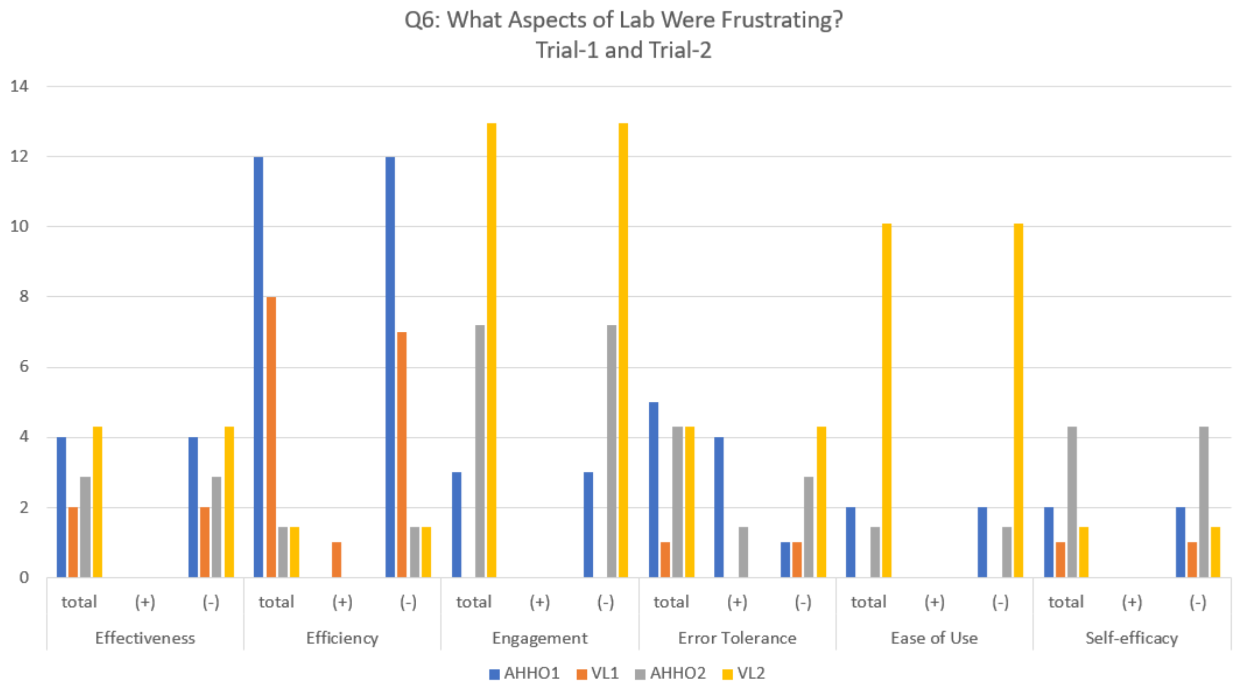
Question 6: Trial-1 and Trial-2

Question 6 (Q6) on the Lab 1 Survey and Lab 2 Survey asked participants to reflect on "What aspects of this laboratory experience frustrated you?". For trial-1 (n=98), 16 respondents (16.3%) chose not to answer Q6; 29 respondents (29.6%) indicated that nothing about Lab 1 frustrated them; and 1 (1.0%) indicated that everything about the VL frustrated them. Similarly, for trial-2 (n=73), 14 respondents (19.2%) chose not to answer Q6; 16 respondents (21.9%) indicated that nothing about Lab 2 frustrated them; and 2 respondents (2.7%) that completed the

AHHO and 3 respondents (4.1%) that completed the VL indicated that everything about the laboratory frustrated them. Again, these negative reactions were consistent with the quantitative findings; the SUS, MRIS, and MARIS data all indicated a below average experience for Lab 2. Figure 4.12 shows the results of the hypothesis coding for this question.

Figure 4.12

Hypothesis Coding: Question 6 – Trial-1 and Trial-2



Note. AHHO$_1$ = at-home, hands-on laboratory experience, trial-1 (n=45); AHHO$_2$ = at-home, hands-on laboratory experience, trial-2 (n=39); VL$_1$ = at-home, virtual laboratory experience, trial-1 (n=53); VL$_2$ = at-home, virtual laboratory experience, trial-2 (n=34). The bar graph shows the total number of responses for each sub-factor and the number of responses that corresponded to the positive (+) and negative (-) aspects of the hypothesis code; the number of trial-2 responses were scaled by a factor of 1.44 to allow for comparison across trials.

The bar graph indicates that efficiency played an important role in the Lab 1 experience (AHHO$_1$ and VL$_1$); engagement and ease of use played a significant role in the VL$_2$ experience.

This is consistent with previous responses. Participants (in both $AHHO_1$ and VL_1) indicated displeasure with the number of coin flips in their responses to Q3 and Q4; a large number of Q6 respondents (n=17) indicated frustration with the excessive amount of time it took them to flip the coins. This sentiment was captured best in the following quote: *"I feel like the amount of time we had to flip the coins was unnecessary"*.

Likewise, participants in VL_2 previously indicated exasperation with virtual ruler (and measurements in general) in their responses to Q3 and Q4. A large number of responses to Q6 (n=12) mentioned frustration with the simulator. The negative impact on the overall laboratory experience is best captured in the following quotes: "*It was frustrating trying to measure objects on a virtual ruler and being able to grasp or understand true measurements which in turn made the lab more confusing and harder to understand"* and "*Measuring objects with the virtual ruler was a bit exasperating."*. The VL_2 results for Q6 were echoed in question 7.

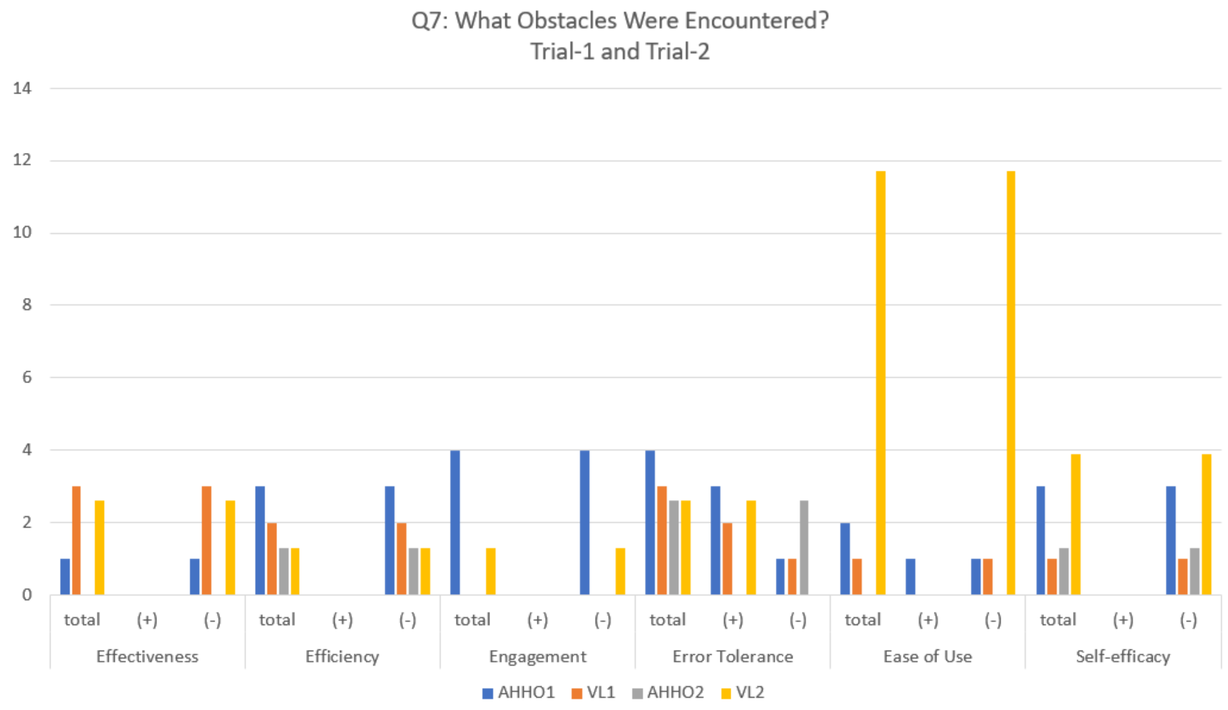
Question 7: Trial-1 and Trial-2

Question 7 (Q7) on the Lab 1 Survey and Lab 2 Survey asked participants to reflect on "What obstacles did you encounter?" For trial-1 (n=98), 24 respondents (24.5%) chose not to answer Q7; 28 respondents (28.63%) indicated there were no obstacles encountered for Lab 1; and 1 (1.0%) indicated that everything about the VL was an obstacle. Similarly, for trial-2 (n=73), 16 respondents (21.9%) chose not to answer Q7; 16 respondents (21.9%) indicated that there were no obstacles encounter in Lab 2.

Figure 4.13 shows the results of the hypothesis coding for this question. The most glaring result (and the only key take-away for Q7) was that ease of use played a significant role in the VL_2 experience. All responses to Q7 were nearly identical to those in Q3-Q6. As there was nothing new (or unique) gleaned from the responses to Q7, it appeared that data saturation was achieved.

Figure 4.13

Hypothesis Coding: Question 7 – Trial-1 and Trial-2



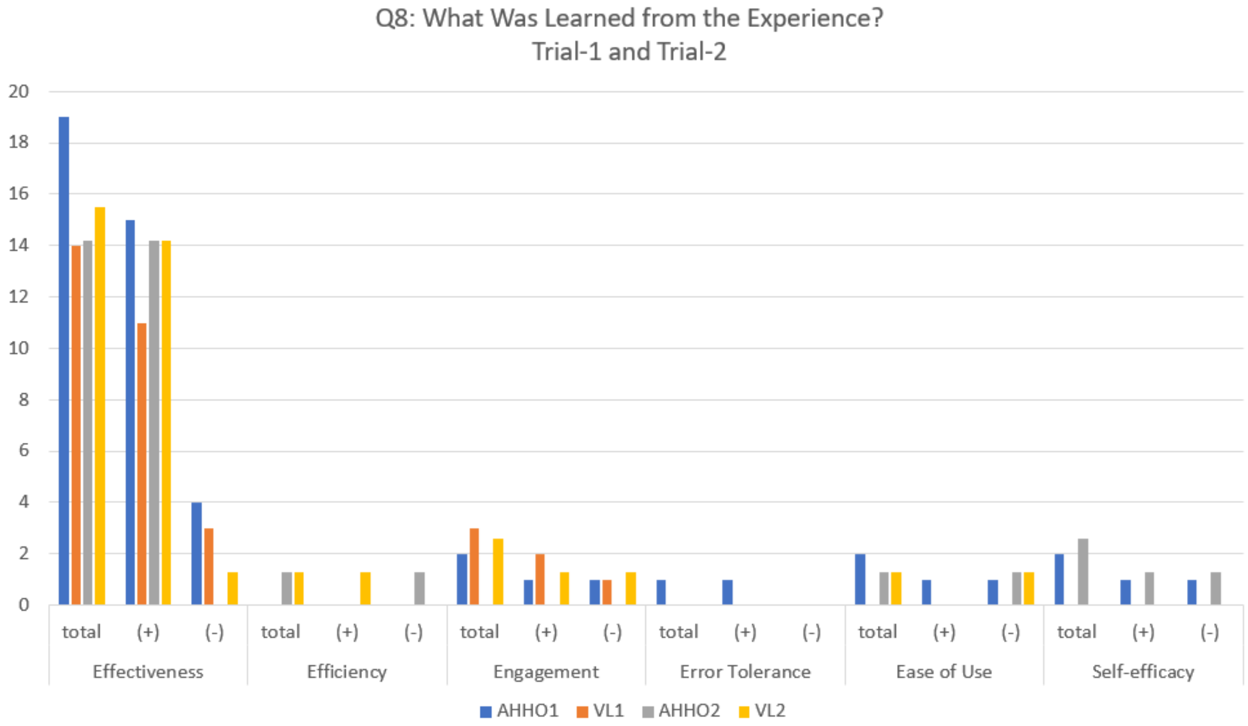
Note. AHHO$_1$ = at-home, hands-on laboratory experience, trial-1 (n=45); AHHO$_2$ = at-home, hands-on laboratory experience, trial-2 (n=39); VL$_1$ = at-home, virtual laboratory experience, trial-1 (n=53); VL$_2$ = at-home, virtual laboratory experience, trial-2 (n=34). The bar graph shows the total number of responses for each sub-factor and the number of responses that corresponded to the positive (+) and negative (-) aspects of the hypothesis code; the number of trial-2 responses were scaled by a factor of 1.30 to allow for comparison across trials.

Question 8: Trial-1 and Trial-2

Question 8 (Q8) on the Lab 1 Survey and Lab 2 Survey asked participants to reflect on "What did you learn from this experience?". For trial-1 (n=98), 22 respondents (22.4%) chose not to answer Q8; and 1 (1.0%) indicated that they were not sure what they learned. Similarly, for trial-2 (n=73), 14 respondents (19.2%) chose not to answer Q8; and 1 respondent (1.4%) indicated they learned nothing. Figure 4.14 shows a comparison of the results for the hypothesis coding for this question.

Figure 4.14

Hypothesis Coding: Question 8 – Trial-1 and Trial-2



Note. AHHO₁ = at-home, hands-on laboratory experience, trial-1 (n=45); AHHO₂ = at-home, hands-on laboratory experience, trial-2 (n=39); VL₁ = at-home, virtual laboratory experience, trial-1 (n=53); VL₂ = at-home, virtual laboratory experience, trial-2 (n=34). The bar graph shows the total number of responses for each sub-factor and the number of responses that corresponded to the positive (+) and negative (-) aspects of the hypothesis code; the number of trial-2 responses were scaled by a factor of 1.29 to allow for comparison across trials.

The bar graph shows that respondents found the $AHHO_1$ laboratory experience to be the most effective; however, participants indicated they were able to achieve the learning goals for all laboratory experiences. Unfortunately, participants were so focused on the calculations that they did not seem to realize what they were supposed to have learned by the laboratory experience. The most common response for Lab 1(n=18, 25%) and for Lab 2 (n=19, 33%) was that they learned math!

Question 9: Trial-1 and Trial-2

Question 9 (Q9) on the Lab 1 Survey and Lab 2 Survey asked participants to reflect on "What questions do you still have?" 93 respondents for trial-1 (94.9%) and 64 respondents for trial-2 (87.7%) stated that they did not have any lingering questions. Some examples of questions that were posed include: *"The relevance?"*; *"What is the purpose?"*, *"Will there be anymore of this?"*, and *"Why this experiment?"*. While these questions show a general sense of frustration, they represent a small proportion (≈8%) of the overall number of respondents. This supports the results from Q8; the laboratory experiences were perceived to be effective.

Crossover Phase: Summary (RQ 4 and RQ5)

The results of the hypothesis coding for the crossover phase were consistent with the quantitative findings. The MARIS and multiple regression analysis identified that effectiveness, engagement, and ease of use played a significant role in the laboratory experience; the same three sub-factors emerged from the responses to Q3-Q9. Furthermore, the results from Q3-Q9 show that the Lab 1 user experience was more positive than the Lab 2 user experience. Math had a negative impact on the $AHHO_2$ experience; math and the virtual ruler had a negative impact on the VL_2 experience. Participants also indicated that they failed to see the relevance of measuring virtual objects. As adult learning theories emphasize the importance of using relevant, real-world experiences (Chen, 2013; Knowles, 1980, 1984; Lewis & Bryan, 2021), the use of the virtual ruler should be reconsidered in future iterations of the course. The factors that play a significant role in the laboratory experience was further explored in Phase II.

Phase II: Self-Selection Phase

Students were given the opportunity to self-select between an at-home, hands-on laboratory experience and an at-home, virtual laboratory experience for Lab 4. While the hands-on and virtual laboratory protocols were both health-related in content and satisfied the same objectives (relationship between correlation and causation), the procedures and pre- and post-lab questions were not the same for each laboratory. Following this final laboratory experience, all students were given the opportunity to complete the Lab 4 Survey; 49 students (n_U=24; n_Q=25) responded.

The open-ended reflection questions for the self-selection phase were self-developed based upon an extensive review of the literature (Attardi et al., 2015; Meintzer et al., 2017; Nolen & Koretsky, 2018; Pal & Vanijja, 2020; Reck et al., 2019; Stuckey-Mickell & Stuckey-Danner, 2007). The questions were grouped into two major categories: (1) what factors prompted participants to select one laboratory protocol over the other (questions 2-3); and (2) what factors do participants consider to be the most important for a good laboratory experience (questions 4-6). The responses were summarized and presented in the following sections.

Category I: Responses to Question 2 and Question 3 (RQ3)

Question 2 (Q2) on the Lab 4 Survey asked participants to indicate "Why did you choose to complete this laboratory exercise?". The primary reasons for choosing the AHHO laboratory experience included: the hands-on aspect of the lab (n=6); the lab seemed fun/interesting (n=6); the lab seemed easier than the VL (n=3); and respondent had the supplies (n=3). On the other hand, the main reasons for choosing the VL experience included: the virtual aspect of the lab (n=9); the lab seemed fun/interesting (n=7); the lab seemed easier the AHHO (n=5); convenience (n=4); respondent did not want to perform measurements (n=1); respondent did not have the supplies at home (n=11); and latex allergy (n=3). Table 4.31 shows a representative sample of the Q2 responses.

Table 4.31

Sample Quotes: Lab 4 Survey, Question 2

Sample Quotes: Why At-home Hands-on Laboratory Experience?
“I generally prefer hands-on experiments. And I had balloons!”
“I chose to complete the hands-on do to we are virtual, and we don’t get to do something like this often. It was fun.”
“I thought it would be more interesting and fun.”
“Found it interesting. And I would much rather do a hands-on experiment over a virtual one”
“It seemed easier than the other one”
“I thought it would help give the best results”
Sample Quotes: Why Virtual Laboratory Experience
“I chose the virtual because it is easier for me to do things online”
“The convenience of the virtual lab”
“Easier for me to research than do a physical experiment in my busy household”
“I don’t like to measure items”
“I am allergic to latex”
“I enjoy doing exercises virtually and I think they are better than trying to gather balloons, cans, rulers, etc.”
“I did not have a balloon handy and the topic on this one seemed more worth my time.”
“I completed this laboratory exercise because I found it much easier to gather information given the time I had to complete the homework.”
“It was online and more feasible to complete for me with my busy schedule”.

Note. This table shows representative sample of the responses to question 2 (Q2).

To further clarify why respondents chose to complete the AHHO or VL experiences, Question 3 (Q3) on the Lab 4 Survey asked participants "Was there something about the other laboratory exercise that made it less appealing to you?". As Q3 is the foil to Q2, it was not surprising that the responses were consistent. Table 4.32 shows a representative sample of the Q3 responses.

Table 4.32

Sample Quotes: Lab 4 Survey, Question 3

Sample Quotes: Why not the Virtual Laboratory Experience?
"When I read over the online experiment, I had a hard time understanding the instructions."
"I really didn't want to do the research."
"It was virtual."
Not really. Just want to do the hands-on lab. I didn't really look at it due to deciding before hand"
"All of the word-answer questions made it less appealing to me"
Sample Quotes: Why not the At-home Hands-on Laboratory Experience?
"It required physical interaction"
"Trying to experiment with a balloon seemed like it would be annoying"
"I do not like going to the store and did not want to purchase a balloon".
"The subject matter"
"No, the ease and accessibility of the online was the only reason"
"Latex allergy"
"I saw some of the math problems, I was not interested in doing, even though I had completed them previously".
"Going out and buying products to use for the lab then waste the extra materials"
"The other lab exercise is more of a solution-based exercise"
"I felt the virtual research lab seemed more interesting and less involved"
"Yes, going on a scavenger hunt throughout my house trying to find the items required to complete the assignment."

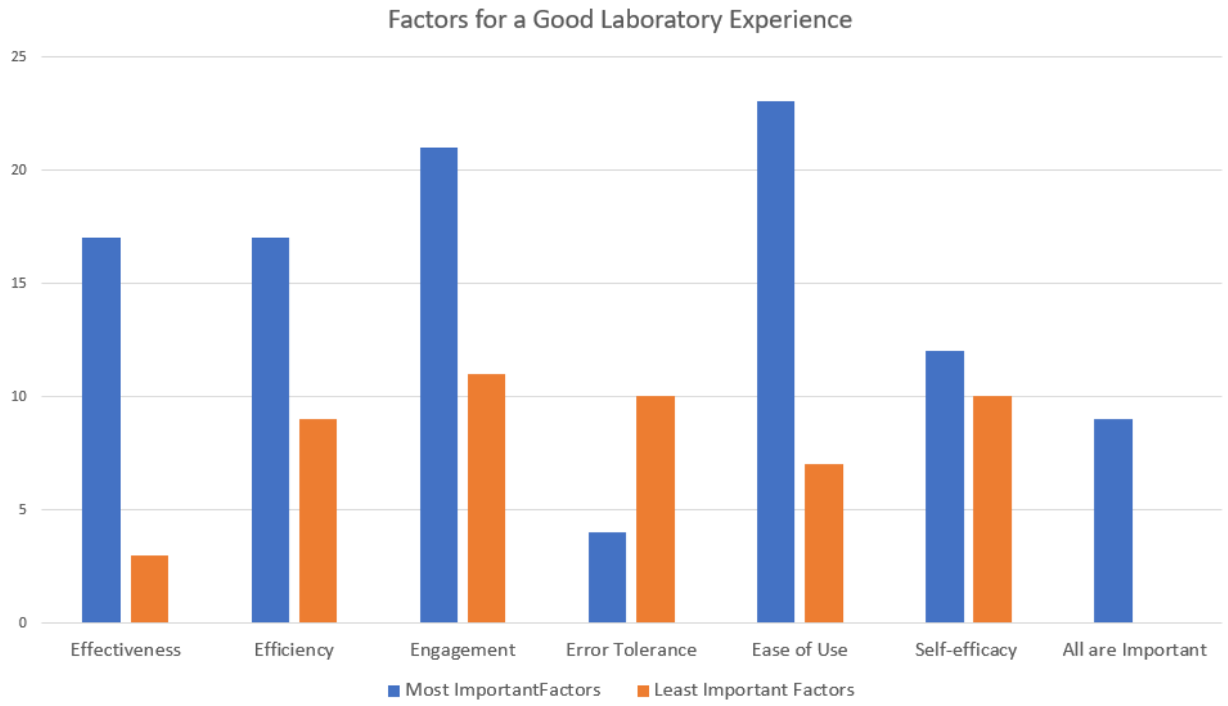
Note. This table shows representative sample of the responses to question 3 (Q3).

Category II: Responses to Question 4, Question 5, and Question 6 (RQ5)

Question 4 (Q4) on the Lab 4 Survey asked participants to indicate "In your opinion, what factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) are the most important for a good laboratory experience?"; and Question 5 (Q5) asked participants to indicate "In your opinion, what factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) are the least important for a good laboratory experience?". Figure 15 shows the results of the Q4 and Q5 analysis.

Figure 4.15

Q4 and Q5 on the Lab 4 Survey: Factors for a Good Laboratory Experience



Note. The bar graph shows the most important and least important factors for a good laboratory experience. Respondents indicated that effectiveness, efficiency, engagement, and ease of use are the most important; engagement, error tolerance, and self-efficacy are the least important.

The results of this analysis are consistent with the quantitative findings (MRIS, MARIS, and multiple regression analysis) and the Phase I qualitative findings: effectiveness, engagement, and ease of use play a key role in the laboratory experience. Additionally, the results for Q4 also identified efficiency as an important factor. It is interesting to note that engagement was identified as both one of the most important factors and least important factors. The importance of engagement seemed to be negatively correlated with age.

Respondents in the <25, 25-29, and 30-34 age groups tended to indicate that engagement was more important than respondents in the 35-39, 40-44, 45-49, 50-54, and 55+ age groups. For example, a respondent in the 35-39 age group stated, *"Engagement might be my least important factor because I know this is school work and has to be done."*; another, in the 45-49 age group indicated that "*engagement [is the least important] as it doesn't matter if I like what I am doing or not it is part of the assignment"*. Conversely, a respondent in the 30-34 age group stated "*I would say engagement and self-efficacy [are the most important]. It is easier to learn in a lab when people are engaged and can complete relatively easy tasks*"; another in the same age group responded *"I would say engagement is probably the biggest factor. You have to enjoy or have fun with what you are doing"*.

This might help explain the trend in SUS scores and age identified in the quantitative findings. Figure 4.5 showed that the mean SUS scores were fairly consistent for the 35-39, 40-44, 45-49, 50-54, and 55+ age groups for both trials (SUS score ≈ 65 ±5); however, there was steep decline in the SUS scores from trial-1 to trial-2 in the <25, 25-29, and 30-34 age groups. As noted in the quantitative findings (MRIS and MARIS) and qualitative findings (Phase I Q3-Q8), the user experience for Lab 1 was better than Lab 2. The 35+ participants did not see the laboratory experience as something to be enjoyed, it was an assignment to be completed (hence the stable SUS score); the SUS scores most likely dropped from Lab 1 to Lab 2 for participants in the <35 group because the actual experience was more important to this group.

Question 6 (Q6) on the Lab 4 Survey asked participants to indicate "What other factors are important for a good laboratory experience?". The results of this analysis are in Table 4.33.

Table 4.33

Lab 4 Survey: Q6 – Other factors

Other Factors	Number of Responses	Reason
Clarity of Instructions* -easy to understand -thorough -straightforward -consistent^H	16	*"It is important to understand what it is you are required to do"* *"Since this is all done online and not in person where questions can be asked"*
Interesting^H	4	*"The experiment should be interesting and exciting for the student to be fully engaged and willing to successfully complete the assignment"*
Streamlined^H	3	*"If the lab looks hard or long, then I will not want to do it. I will dread it."*
Materials Provided*	3	*"It is hard to do the lab without the materials needed for the lab"*
Physicality*	2	*"I like when I physically get to do something. One thing about online school is it doesn't happen too often"*
Access to Professor^H	1	*"Access to professor for questions"*
Relevance*	1	*"The relationship of the exercise with the coursework"*

Note. This table shows a summary of the factors identified in question 6 (Q6); it also shows representative sample of the responses for the reason why the factor was identified. The factors identified with (*) correspond with one of the emergent codes from Phase I analysis; factors identified with (^{H}) correspond with one of the hypothesis codes from the Phase I analysis.

It should be noted that all of the ‘other factors’ respondents identified in Q6 corresponded to either one of the hypothesis codes (SUS sub-factors) or to one of the codes that emerged from the Phase I data analysis (or both). ‘Clarity of Instructions’ aligned with the emergent code ‘clarity of instructions’ and with the hypothesis code ‘effectiveness’; ‘interesting’ aligned with the hypothesis code ‘engagement’; ‘streamlined’ aligned with the hypothesis code ‘efficiency’; ‘materials provided’ aligned with the emergent code ‘materials’; ‘physicality’ aligned with the emergent code ‘prefer hands-on’; ‘access to the professor’ aligned with the hypothesis code ‘error tolerance’; and ‘relevance’ aligned with the emergent code ‘relevance’. The alignment of the ‘other factors’ with those already identified during the coding process means that the open-ended questions used in Phase I (Q3-Q8) and Phase II, category II (Q4-Q5) were effective (i.e. they successfully identified all of the factors that were important to respondents). It also lends an air of trustworthiness to the coding methodology and data analysis (i.e. a high degree of confidence in the data and interpretation as all factors important to respondents were identified and in alignment).

It should also be noted that ‘materials’ played a more significant role in the user-experience than anticipated. In the self-selection phase, 11 respondents (26.2%) indicated that they would have preferred to the AHHO experience, but they had to select the VL experience due to a lack of materials. One of the assumptions of the study was that students would have access to the supplies necessary to complete the lab as they are common household items (coins, ruler, and balloons). However, respondents remarked that finding a coin or their ruler (in the crossover phase) or having to go to the store to purchase balloons (in the self-selection phase) was a huge dissatisfier.

Furthermore, ‘materials’ also included the virtual simulators. The selection criteria for the virtual simulators utilized in this study, included: (a) simulator was free; (b) simulator did not require students to set up an account (i.e. did not require a username or password); and (c) simulator was not too “glitchy” (i.e. did not freeze up or kick you out during the data collection). The simulator for Lab 1 was not as large a dissatisfier as Lab 2. The below average user experience for Lab 2 was in part due to the negative ease of use of the simulator.

These findings support the purpose of this study – investigating the use of an instrument that could help determine students’ preferred laboratory experience (hands-on or virtual) to make data-driven decisions for course design and resource allocation. Laboratories are expensive. Standard laboratory kits cost $150-$300 per kit per student; good quality simulators cost $50-$100 per simulation per student. While it is important to control costs for the University and to students (especially for adult learners), the quality of the laboratory experience should not be limited by the lack of physical supplies or poor quality of the free simulators.

Summary of Merged Results

This convergent mixed methods comparison study utilized Brooke’s (1996) System Usability Scale (SUS) as a framework to evaluate the preferred at-home, laboratory experience for adult learners in an online, undergraduate science course for non-majors at a private, professionally-focused university in the Midwestern United States. The study occurred in two phases. In Phase I (quasi-experimental cross-over phase), students completed the Lab 1 and Lab 2 Surveys after performing an at-home hands-on (AHHO) and after performing an at-home virtual experience (VL). In Phase II (self-selection phase), students were given the opportunity to self-select between AHHO or VL experience for their last laboratory assignment; participants then completed the Lab 4 survey in which they identified which laboratory experience they selected and why.

Summary of the Quantitative Strand: Phase I

In Phase I (quasi-experimental crossover phase), the first two questions on the Lab 1 Survey (trial-1) and Lab 2 Survey (trial-2) were used to gather data on perceived usability (SUS score) and its sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy) for the AHHO and VL laboratory experiences.

Phase I: Perceived Usability (RQ 1-1)

The SUS score was calculated for each participant using the method outlined in Sauro and Lewis (2016). The mean SUS scores for $AHHO_1$, VL_1, $AHHO_2$, VL_2, AHHO, and VL were calculated and analyzed using SAS OnDemand for Academics. The results of the two-sample, t-test (H_0: $\mu_{AHHO} - \mu_{VL} = 0$; H_a: $\mu_{AHHO} - \mu_{VL} \neq 0$) and the non-parametric Wilcoxon rank sum test (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$) showed that the perceived usability was nearly the same for AHHO and VL for Trial-1, Trial-2, and Combined (Trial-1 + Trial-2) datasets. Although the findings for RQ 1-1 were inconclusive for preferred laboratory experience, the conclusion of ‘nearly the same’ is consistent with the literature.

Phase I: Sub-factors (RQ 1-1’)

Based upon Lewis and Sauro’s (2019) description of item attributes, each of the SUS items were coded as a measure of effectiveness, efficiency, engagement, error tolerance, ease of use, or self-efficacy. Each variable corresponded to both a positive statement and negative statement, with the exception of error tolerance (negative statement only) and self-efficacy (positive statement only). The mean raw item scores (MRIS) and the mean adjusted raw item scores (MARIS) were used to determine if there was evidence to support preferred laboratory experience based upon one or more factors.

MRIS Results. The results of the two-sample, t-test (H_0: $MRIS_{AHHO} – MRIS_{VL} = 0$; H_a: $MRIS_{AHHO} - MRIS_{VL} \neq 0$) and the non-parametric Wilcoxon rank sum test (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$) showed the differences in the MRISs (for trial-1) were not significant, except for item 2 (ease of use) and the differences in MRISs (for trial-2) were not significant except for item 8 (engagement). Therefore, the MRIS analysis not yield conclusive results for participants' preferred laboratory experience. However, the analysis did identify variables that played a role on the user experience; the variables that met the SUS=68 benchmark (average user experience) were effectiveness, engagement, ease of use, and self-efficacy.

Based upon the one-way ANOVA test (H_0: MRISs are equal; H_a: at least one of the MRISs is different) and the Kruskal-Wallis test (H_0: medians the same; H_a: at least one of the medians is different), there were significant differences in the MRIS values for Group A ($AHHO_1/VL_2$) for item 8 (engagement) and item 9 (self-efficacy); Group B ($AHHO_2/VL_1$) showed significant differences for item 1 (engagement), item 3 (ease of use), item 8 (engagement), item 9 (self-efficacy), and item 10 (efficiency). Again, the analysis did not yield conclusive results for preferred laboratory experience; however, the tests did identify engagement, ease of use, self-efficacy, and efficiency are important factors in the laboratory experience.

MARIS Results. The results of the two-sample, t-test (H_0: $MARIS_{AHHO} – MARIS_{VL} = 0$; H_a: $MARIS_{AHHO} - MARIS_{VL} \neq 0$) and the non-parametric Wilcoxon rank sum test (H_0: $\Delta = 0$; H_a: $\Delta \neq 0$) showed the differences in the MARISs (for trial-1) were not significant, except for engagement ($p=0.0042$) and ease of use ($p=0.0182$) with $VL_1 > AHHO_1$. This means that the students preferred the VL laboratory experience for trial-1 based upon engagement and ease of use; the preferred laboratory experience for the other sub-factors (effectiveness, efficiency, error tolerance, and self-efficacy) was inconclusive. For trial-2, the tests showed the differences in the MARISs (for trial-2) were not significant, except for engagement (p=0.0280) with $AHHO_2 > VL_2$. This means that the students preferred the AHHO laboratory experience for trial-2 based upon engagement; the preferred laboratory experience for the other sub-factors (effectiveness, efficiency, error tolerance, ease of use, and self-efficacy) was inconclusive.

Based upon the one-way ANOVA test (H_0: MARISs are equal; H_a: at least one of the MARISs is different) and the Kruskal-Wallis test (H_0: medians the same; H_a: at least one of the medians is different), there were significant differences between combined MARIS values for Group A for all of the variables except error tolerance (p=0.7018) and ease of use (p=0.0682); Group B showed significant differences for all of the variables except for error tolerance (p=0.0756).

Finally, a two-tailed, two sample t-test ($\alpha=0.05$) and Wilcoxon rank sum test was conducted to determine if the differences combined MARIS values for AHHO (trial 1 + trial 2) and VL (trial 1 + trial 2) were significant. The findings showed that all of the sub-factors were nearly the same for AHHO and VL. Again, the findings were inconclusive for preferred laboratory experience; however, the analysis did identify effectiveness, engagement, and ease of use as variables that play an important role based upon a MARIS ≥ 2.5 and t-tests.

Phase I: Sub-factors (RQ 2)

The mean raw item scores (MRIS), mean adjusted raw item scores (MARIS), and multiple regression analysis (SUS and MARIS) were used to determine if there was evidence to support that one or more factors played a significant role in the user experience. As discussed above, the MRIS analysis identified effectiveness, engagement, ease of use, and self-efficacy as important factors in the laboratory experience; the MARIS analysis identified the same variables (with the exception of self-efficacy).

Multiple regression (H_0: $\beta_i = 0$; H_a: $\beta_i \neq 0$) was used to analyze the relationship between user experience (perceived usability) and the SUS sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, self-efficacy). The general format for the multiple regression model is: $SUS = \beta_0 + \beta_1 E + \beta_2 F + \beta_3 N + \beta_4 ET + \beta_5 EU + \beta_6 SE$, where β_0= y-intercept; β_i partial slope for specified sub-factor; E = combined MARIS for effectiveness, F = combined MARIS for efficiency; N = combined MARIS for engagement; ET = combined MARIS for error tolerance; EU = combined MARIS for ease of use; and SE = combined MARIS for self-efficacy.

The results for the AHHO showed that all the variables were significant ($p<0.05$). The variables listed in order of magnitude of the partial slope parameter were as follows: engagement ($\beta_3 = 5.6718$) > ease of use ($\beta_5 = 4.5412$) > effectiveness ($\beta_1 = 4.2633$) > error tolerance ($\beta_4 = 3.7333$) > self-efficacy ($\beta_6 = 3.1340$) > efficiency ($\beta_2 = 3.0225$). As the magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott & Longnecker, 2016), effectiveness, engagement, and ease of use appeared to play a significant role in the AHHO user experience.

The results for VL showed that all of the variables were significant ($p<0.05$). The variables listed in order of magnitude of the partial slope parameter were as follows: ease of use ($\beta_5 = 6.446$) >> self-efficacy ($\beta_6 = 4.6876$) > engagement ($\beta_3 = 4.2078$) >> error tolerance ($\beta_4 = 2.6078$) > efficiency ($\beta_2 = 2.0755$) > effectiveness ($\beta_1 = 1.7575$). Again, as the magnitude of the slope parameter is indicative of the size of the effect of the variable on the user experience (Ott

& Longnecker, 2016), ease of use, engagement, and self-efficacy appeared to play a significant role in the VL user experience. Both the AHHO and VL results were consistent with the MRIS and MARIS results.

Summary of the Quantitative Strand: Phase II

During Phase II (self-selection phase), the first question on the Lab 4 Survey was used to calculate the proportion of respondents that chose to complete AHHO and VL experiences. These data were compared to determine students preferred at-home laboratory experience.

Phase II: Preferred Laboratory Experience (RQ 1-2)

A two-proportion Z-test ($\alpha=0.05$) was used to determine if there was a significant difference between the proportions (H_0: $\hat{p}_{AHHO} = \hat{p}_{VL}$; H_a: $\hat{p}_{AHHO} \neq \hat{p}_{VL}$). The test statistic ($Z_{test} = 2.8125$) was greater than the critical value ($Z=1.9600$); therefore, we rejected the null hypothesis. This means that the difference in proportions was significant. As the $\hat{p}_{VL} > \hat{p}_{AHHO}$; the self-selection data indicated that participants preferred the virtual laboratory experience.

However, the proportions were adjusted based on responses to Question 2 (“Why did you choose to complete this laboratory exercise?") and Question 3 (“Was there something about the other laboratory exercise that made it less appealing to you?”). There were 11 respondents that indicated they would have preferred to perform the AHHO version of the laboratory, but did not have access to the materials for the lab experiment. After adjusting the numbers (by +11 for AHHO; -11 for VL), the test statistic for the adjusted proportions ($Z_{test} = 0.3018$) was less than $Z=1.9600$ (critical value); therefore, we fail to reject the null hypothesis. This means that the difference in proportions is not significant. As $\hat{p}_{AHHO} = \hat{p}_{VL}$, the self-selection data was inconclusive.

The adjusted Phase II findings were consistent with the Phase I findings (SUS score and sub-factor analysis). As these data all arrived at the same conclusion (AHHO≈VL), this supports the use of the SUS as an effective tool to measure perceived usability of a laboratory experiment. Moreover, the Phase I and Phase II results of 'nearly the same' are also consistent with the literature.

Summary of the Qualitative Strand: Phase I (RQ 4 and RQ 5)

The surveys for the crossover phase (Lab 1 Survey and Lab 2 Survey) included open-ended reflection questions (Q3-Q9) modified from Reck et al. (2019) (with permission, Appendix G). The results of the hypothesis coding for the crossover phase identified that effectiveness, engagement, and ease of use played a significant role in the laboratory experience. Emergent codes 'math', 'materials', and 'relevance' seemed to have a significant, negative impact on the VL_2 experience. Other codes that emerged from the data analysis were: clarity of instructions, worked with kids, data collection, measurements, prefer hands-on, prefer virtual, and convenience.

Summary of the Qualitative Strand: Phase II

The open-ended reflection questions for the self-selection phase were self-developed based upon a systematic review of the literature (Attardi et al., 2015; Meintzer et al., 2017; Nolen & Koretsky, 2018; Pal & Vanijja, 2020; Reck et al., 2019; Stuckey-Mickell & Stuckey-Danner, 2007). The questions were grouped into two major categories: (1) what factors prompted participants to select one laboratory protocol over the other (questions 2-3); and (2) what factors do participants consider to be the most important for a good laboratory experience (questions 4-6). The responses were summarized and presented in the following sections.

Category I: RQ 3

Based upon the responses to Q2 and Q3 on the Lab 4 Survey, the primary reasons for choosing the AHHO laboratory experience included: the hands-on aspect of the lab (n=6); the lab seemed fun/interesting (n=6); the lab seemed easier than the VL (n=3); and respondent had the supplies (n=3). On the other hand, the main reasons for choosing the VL experience included: the virtual aspect of the lab (n=9); the lab seemed fun/interesting (n=7); the lab seemed easier the AHHO (n=5); convenience (n=4); respondent did not want to perform measurements (n=1); respondent did not have the supplies at home (n=11); and latex allergy (n=3).

Category II: RQ4 and RQ 5

Based upon the responses to Q4 and Q5 on the Lab 4 Survey, the factors that played a key role on the laboratory experience were effectiveness, engagement, and ease of use. This was consistent with the factors identified in the Phase I findings. Additionally, the ‘other factors’ responses identified in Q6 corresponded to either one of the hypothesis codes (SUS sub-factors) or to one of the codes that emerged from the Phase I data analysis (or both). ‘Clarity of Instructions’ aligned with the emergent code ‘clarity of instructions’ and with the hypothesis code ‘effectiveness’; ‘interesting’ aligned with the hypothesis code ‘engagement’; ‘streamlined’ aligned with the hypothesis code ‘efficiency’; ‘materials provided’ aligned with the emergent code ‘materials’; ‘physicality’ aligned with the emergent code ‘prefer hands-on’; ‘access to the professor’ aligned with the hypothesis code ‘error tolerance’; and ‘relevance’ aligned with the emergent code ‘relevance’. The alignment of the ‘other factors’ with those already identified during the coding process means that the open-ended questions used in Phase I (Q3-Q8) and Phase II, category II (Q4-Q5) were effective (i.e. they successfully identified all of the factors that were important to respondents). It also lends an air of trustworthiness to the coding methodology and data analysis (i.e. a high degree of confidence in the data and interpretation as all factors important to respondents were identified and in alignment).

Merged Findings (Overarching Research Questions 1 and 2)

While the SUS analysis, sub-factor analysis, and self-selection data analysis were inconclusive for preferred laboratory experience (AHHO ≈ VL for all statistical analyses; RQ1), the quantitative findings (MRIS, MARIS, and multiple regression analysis) and qualitative findings (phase I and phase II) identified that effectiveness, engagement, and ease of use play a key role in the laboratory experience (RQ 2). The next chapter will provide a deeper discussion of the research findings, limitations of the study, recommendations for future research, and practical implication of the study.

Chapter 5 – Results, Conclusions and Recommendations

Leaders in higher education face an increasingly complicated set of challenges. Intense competition for student enrollment, a turbulent financial environment, changing expectations of the workforce, issues of equity and inclusion, access to technology, and the rapid pace innovation are just a few of the issues that must be addressed to remain competitive in this volatile market (Kim & Rehg, 2018; Neuwirth et al., 2020; Sunderman et al., 2020). This requires Program Chairs to stay current in their field so that they can maintain relevant, high-quality curricula that meets the specific needs of the learners at their institution (Weaver et al., 2019). For science programs. this includes how and where laboratories are performed (Raman et al., 2022; Reeves & Crippen, 2021).

As the laboratory experience is often considered to be the cornerstone of STEM education (Clough, 2002), it should not be surprising that the efficacy of alternative laboratory experiences has been extensively studied. Since the early 2000s, there have been more than 20,000 publications (journal articles, editorials, books, and conference proceedings) in which science educators from a wide variety of STEM disciplines (Biology, Chemistry, Physics, and Engineering), from multiple educational levels (K-12, undergraduate, and graduate) and from more than 25 countries have studied the efficacy of alternative laboratory experiences (Chapter 2). Even with a spike in the number of studies between 2020-2022 due to campus closures in response to the COVID-19 pandemic (Raman et al., 2022), there is still a paucity of research on adult learners' perceptions of these alternative laboratory experiences. Given that, in the United States alone, adults over the age of 25 accounted for 25% of total student enrollment in higher education (NCES, 2021b), this represents a significant gap in the literature.

Furthermore, 85 of the 95 studies selected for review in Chapter 2 utilized home-grown assignments (quizzes, exams, laboratory reports, projects, pre-/post-tests, etc.), final grades, or self-developed surveys/questionnaires/interview questions to assess the effectiveness and/or students' perceptions of the laboratory experiences. The pervasive use metrics that were developed by a specific instructor for a specific section of a specific class to meet the specific learning objectives at a specific institution makes it nearly impossible to generalize these findings, even to the same course at another institution (Gopalan et al., 2020). This highlighted the need for a normed and validated tool that can be used to evaluate student perception data across a myriad of laboratory experiences.

This convergent mixed-methods study utilized Brooke's (1996) System Usability Scale (SUS) as a framework to evaluate the preferred at-home, laboratory experience for adult learners in an online, undergraduate science course for non-majors at a private, professionally-focused university in the Midwestern United States. The SUS is a normed, validated, and product-agnostic questionnaire (Bevan, 1995; Brooke, 2013; Lewis, 2018a) that has been used to measure the perceived usability of a wide range of products (Lacerda & von Wangenheim, 2018; Lewis, 2018b; Lewis & Sauro, 2019). Although this tool has been used to evaluate students' perceptions of various educational technologies (Granić & Ćukušić, 2011; Harrati et al., 2016; Orfanou et al., 2015; Vlachogianni & Tselios, 2021, 2022), to the researcher's knowledge, the SUS has never been used to evaluate user experiences for laboratory experiments. The SUS measures the quality of the product by users' perceptions of its effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy (Bangor et al., 2008; Bevan, 1995; Brookes,1996; Lewis, 2018b; Quesenbery, 2001; Sauro & Lewis, 2016). As these variables closely align with factors that support adult learners' independent learning styles (Arghode et al., 2017; Calabrese & Capraro, 2021; Chen, 2014; Knowles, 1980, 1984), it was proposed that a slightly modified version of the SUS would be able to provide valuable insights into the laboratory experiences for this student population.

The study occurred in two phases, a quasi-experimental crossover phase followed by a self-selection phase. The results of the self-selection phase were used to corroborate the results of the crossover phase; and the integration of the quantitative and qualitative data strands provided the opportunity to corroborate the results of the modified SUS, seek complementarity, and identify areas of dissonance between the crossover and self-selection datasets. The results in both phases showed no significant difference between the at-home, hands-on (AHHO), and virtual (VL) laboratory experiences. This is consistent with the literature reviewed in Chapter 2 in which 93% of the selected comparison studies reported that the laboratory experiences were ‘nearly the same’.

While the results for adult learners’ preferred laboratory experience were inconclusive, the quantitative and qualitative analyses consistently identified effectiveness, engagement, and ease of use as factors that played an important role in the laboratory experience. Moreover, as the self-selection data and crossover data arrived at the same conclusion, the modified SUS proved to be an effective instrument to measure perceived usability, its sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy), and the quality of the user experience for laboratory experiments. This chapter will provide an in-depth discussion of the study’s findings and their design implications, limitations (and delimitations), and recommendations for future research.

Discussion of the Findings and Implications for Instructional Design

The overarching goals of this mixed-methods study were to investigate adult learners preferred at-home laboratory experience in terms of perceived usability and self-selection responses; and to investigate which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, self-efficacy, or other) played a significant role in the participants' experiences. To achieve these goals, data was collected from a convenience sample of students registered for SCIE 211: Introduction to Scientific Analysis and Reasoning, a 12-week online, undergraduate, science course for non-majors at a private, professionally-focused university in the Midwestern United States. During the data collection period (Fall 2022), the research site offered 5 sections of SCIE 211 in the U-term (with a start date in August) and 6 sections in the Q-term (with a start date in September); the staggered start dates allowed for replication of the study within the same trimester.

All students were given the opportunity to complete the anonymous surveys for both trials of the crossover phase (Lab 1 Survey and Lab 2 Survey) and the anonymous survey for the self-selection phase (Lab 4 Survey). The response rate was fairly high (36.3% for the Lab 1 Survey; 27.0% for the Lab 2 Survey; and 18.1% for the Lab 4 Survey) and the demographic data indicated that a majority of the respondents (>85%) met the definition of an adult learner. All three surveys included both closed-ended and open-ended questions. The quantitative results were reported for U-term, Q-term, and Overall (when appropriate based on U-/Q-term sample sizes); the qualitative results were reported for the Overall population only as all statistical analyses revealed no significant difference between the U-term and Q-term sub-populations.

The next section will discuss the findings for mixed-methods research question 1: Will adult learners in an online, undergraduate, science course for non-majors at a private, professionally-focused university in the Midwest prefer at-home, hands-on laboratory experiences or at-home, virtual laboratory experiences? This research question was supported by evidence from both Phase I (quasi-experimental crossover phase) and Phase II (self-section phase).

Adult Learners Preferred At-home Laboratory Experience

Phase I: Quasi-Experimental Crossover Phase

In terms of perceived usability, this study found that the quality of the AHHO and VL user experiences were nearly the same. A modified version of Brooke's (1996) SUS questionnaire was used to gather data on perceived usability (SUS score) and its sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy) for the AHHO and VL laboratory experiences. The results of the two-tailed, two-sample, t-test (and the non-parametric Wilcoxon rank sum test) showed there was no significant difference in the perceived usability for AHHO and VL for each trial, for each sub-population (U-term, Q-term, and Overall), and for the Combined (trial-1 + trial-2) dataset.

Although the finding of 'nearly the same' does not provide a clear answer to the question of preferred laboratory experience, this result can still be used to inform instructional design decisions. On the one hand, no significant difference means that respondents found instructional value in both the hands-on and virtual laboratory experiences. This is consistent with similar studies that found replacing physical interactions with virtual experiences had little to no impact on users' satisfaction and conceptual learning gains (Brewer et al., 2013; Gnesdilow & Putambekar, 2022; Klahr et al., 2007; Oser & Fraser, 2015; Pyatt & Sims, 2011; Zacharia et al., 2008). This implies that the laboratory medium (AHHO or VL) may not be important to the user experience for the adult learner.

On the other hand, the findings of the SUS and sub-factor analysis revealed that the quality of the user experience was heavily influenced by the perceived ease of use of the laboratory materials and the perceived effectiveness (or clarity) of the laboratory protocol. There was a steep decline in perceived usability between trial-1 (*Lab 1: Probability and Statistics*) and trial-2 (*Lab 2: Measurements and Significant Figures*) for both AHHO (SUS_{AHHO1} = 69.039, C-grade; SUS_{AHHO2} = 60.329, D-grade) and VL (SUS_{VL1} = 75.778, B-grade; SUS_{VL2} = 55.294, D-grade); with a significant difference in perceived usability between VL_1/VL_2 for each sub-population (for U-term, Q-term, and Overall). Moreover, the results of the one-way ANOVA sub-factor analysis (and the non-parametric Kruskal-Wallis test) showed that there were significant differences in perceived effectiveness, efficiency, engagement, ease of use, and self-efficacy between $AHHO_1/VL_2$, $AHHO_2/VL_1$, and VL_1/VL_2 for each sub-population (U-term, Q-term, and Overall); and there were significant differences in perceived ease of use and self-efficacy between $AHHO_1/AHHO_2$ for each sub-population (U-term, Q-term, and Overall).

The results of the statistical analyses implied that while participants in both mediums (AHHO and VL) were successfully able to complete Lab 2, the experience was not enjoyable as respondents found the lab to be difficult, time-consuming, and made them feel unsure of their performance. These findings were corroborated by the analysis of the open-ended responses. Hypothesis coding showed that ease of use had a negative impact on the VL_2 experience. The Lab 2 simulator was harder to use and required more work than the Lab 1 simulator. Students had to upload virtual images, manipulate the virtual ruler to measure each object’s length, width, and height multiple times; while students only had to click on a virtual coin for Lab 1. Frustration with the virtual ruler (and measurements in general) was captured by the following quote: “*It was frustrating trying to measure objects on a virtual ruler and being able to grasp or understand true measurements which in turn made the lab more confusing and harder to understand”.*

However, the most telling statistic was that >70% of the responses indicated that math (the number of calculations and the calculations themselves) was the main source of frustration and the primary obstacle to the completion of Lab 2. The emergent code ‘math’ was closely tied to the emergent code ‘clarity of instructions’ (and the hypothesis code ‘effectiveness’); respondents largely indicated that the formulas (and examples) provided in the laboratory instructions were not adequate to support their ability to perform the calculations. Studies have demonstrated there is a strong correlation between math self-efficacy and STEM interest/perception (Blotnicky et al., 2018; Rozgonjuk et al., 2020; Simpkins et al., 2006). As such, the negative user experiences for Lab 2 may have more to do with participants’ math self-efficacy (and perceived lack of clarity of the instructions) than the actual laboratory experience (manipulation of real or virtual objects).

Furthermore, correlation analysis revealed a moderate, positive relationship between ‘efficiency’ and ‘self-efficacy’ ($r = 0.5858$). As such, the perceived inefficiency of Lab 2 may have been influenced by participants’ lack of self-confidence in measurement skills and mathematical abilities (i.e. participants spent additional time repeating/redoing work because they doubted the accuracy of their measurements and calculations). This observation is salient as a large majority of respondents indicated that they are full-time students (70%) and work more than 36 hours/week (86%). Laboratory experiences that take more time (and effort) to complete than anticipated can create additional stress and decrease overall satisfaction (Andrews et al., 2020; Easdon, 2020; Moosvi et al., 2019) especially for adult learners with limited time for schoolwork (Osam et al., 2017).

These findings were consistent with similar studies that found student interest and engagement in the laboratory experience were more closely tied to the quality of the experimental design and materials than to the laboratory medium (Johnson & Barr, 2021; Koretsky et al., 2011; Nolan & Koretsky, 2018; Son & Narguizian, 2016). This implies that laboratory medium (AHHO or VL) is not as important to adult learners as the clarity of the laboratory instructions, access to ancillary instructional materials (videos/examples), and the usability of the laboratory materials. As respondents for Lab 1 seemed to enjoy the convenience and ease of use of the coin-flip simulator, but respondents for Lab 2 seemed to be more comfortable using a real ruler to measure real objects, a blended approach may be best (i.e. each experiment should utilize the manipulative, physical or virtual, that best fits the task at hand).

Phase II: Self-Selection Phase

In terms of self-selection responses, this study found that the proportion of participants that selected the AHHO was nearly the same as the proportion that selected the VL experience. The first question on the Lab 4 Survey was used to calculate the proportion of respondents that chose to complete AHHO and VL experiences. A two-proportion Z-test (α=0.05) showed that there was a significant difference between $\hat{p}_{AHHO}=0.2619$ and $\hat{p}_{VL}=0.7381$ for the Overall dataset (the sample sizes for U-term and Q-term were too small for an accurate determination). It initially appeared that the self-selection dataset indicated that participants preferred the virtual laboratory experience as $\hat{p}_{VL} > \hat{p}_{AHHO}$.

However, the proportions were adjusted based upon the responses to Question 2 (“Why did you choose to complete this laboratory exercise?") and Question 3 (“Was there something about the other laboratory exercise that made it less appealing to you?”). There were 11 respondents that indicated they would have preferred to perform the AHHO version of the laboratory, but did not have access to the materials for the lab experiment. After adjusting the numbers (by +11 for AHHO; -11 for VL), the two-proportion Z-test showed there was no significant difference between $\hat{p}_{AHHO}$=0.5238 and $\hat{p}_{VL}$=0.4762. As $\hat{p}_{AHHO} \approx \hat{p}_{VL}$, the self-selection data was inconclusive for preferred laboratory experience.

Again, even though the finding of 'nearly the same' does not provide a clear answer to the question of preferred laboratory experience, these results (in conjunction with the open-ended responses) can still be used to inform instructional design decisions. First, this further supports that respondents found instructional value in both the hands-on and virtual laboratory experiences and that the laboratory medium (AHHO or VL) may not be important to the user experience for the adult learner.

Additionally, these findings further support the impact that 'materials' had on the user experience. One of the assumptions for this study was that students would have access to the supplies necessary to complete the lab as they are common household items (coins, a ruler, and balloons). However, respondents remarked that having to take the time to find a coin or a ruler (in the crossover phase) or to go to the store to purchase balloons (in the self-selection phase) was a huge dissatisfier. This is consistent with similar studies that found the AHHO experiences were negatively impacted by students' lack of easy access to home supplies (Easdon, 2020; Kelley, 2020; Moosvi et al., 2019; Youssef et al., 2021).

More importantly, Vogt et al. (2013) found that the persistence rate for adult learners in their blended AHHO experience was significantly lower than for students in the traditional section. The adult learners indicated that the added time stressors of the AHHO experience, on top of their full-time work schedules and extensive family commitments, led them to withdraw from the course or to transfer to the on-campus section (where they had easy access to supplies and just-in-time instructor support). While the data collected for this study did not include persistence (or retention) rates for SCIE 211, correlation analysis did reveal a moderate, positive relationship between ‘efficiency’ and ‘self-efficacy’ (r = 0.5858); and that self-efficacy was closely tied to the quality (and clarity) of the laboratory materials. This, in conjunction with Tinto’s (2017) assertion that persistence is influenced by self-efficacy, implies that laboratory experiences that take more time (and effort) to complete than anticipated not only can create additional stress for the adult learner (Osam et al., 2017); but also, can undermine the learner’s belief in their ability to successfully complete the laboratory-science course.

However, Tinto (2017) also posited that self-efficacy is “malleable” (p.3) and can be girded by a supportive University environment. Effective leaders use their influence to shape a culture that is ready and able to respond to changes in (or new understandings of) stakeholders’ needs (Adero & Odiyo, 2020; Cote, 2017; Iordanoglou, 2018). This is why it is important for lead faculty and program chairs to have effective instruments to measure students’ perceptions and to collaborate with senior leadership in order to develop a coordinated university response to support students’ needs and preferences (Weaver et al., 2019). In terms of laboratory experiences, the data on the impact of materials on adult learners’ self-efficacy (and possibly persistence) implies that students should be able to opt-in for a university provided kit that includes all of the materials to complete each laboratory activity. A customized kit could either be purchased by the student via the bookstore (or other University channel) or be provided directly to students by the University. There are strengths and weaknesses for both of these options.

Orozco (2017) tested a laboratory kit purchased from a popular vendor. While students had access to all of the materials for the 13 laboratory activities (and reported a positive user experience in terms of efficiency and ease of use), the kit was expensive and the laboratory manual did not provide instructional materials. This highlights that vendors must be carefully selected for price, convenience, instructional materials provided, and the ability to customize the kit (to meet the laboratory goals and limit costs to students). Depending on the vendor, the cost of a customized kit can be in line with the laboratory fees ($100-$300). Furthermore, some vendors not only provide equipment and supplies but also access to an electronic textbook and laboratory manual in a format that is ready to "plug-into" the learning management system. This option is cost effective for both students and the University.

On the other hand, Honig et al. (2020) packaged and sent their own kits to students. While this ensured that the students designated as "Process Engineers" (p.87) had access to the materials for the laboratory experiences, it was labor-intensive for both the University and the students. University staff had to develop, build, package, and track the kits (to ensure students received and returned the materials); Process Engineers had to fabricate equipment from parts (due to packaging constraints), make it "work" in their home environments (with help from teammates and university staff via video conferencing), and break down equipment to return the kit. Unfortunately, not all of the students qualified as a Process Engineer due to environmental factors, including space constraints, incompatible water tap fitting, issues with power supply, and lack of a second person at home in case of emergency. These students were assigned as a "Design Engineer" (p.87) and placed in a group with a Process Engineer. Honig et al. (2020) illustrates that while this option may cheaper for the student, the laboratory experience can be inequitable and can have significant time costs for all stakeholders.

Furthermore, 'materials' also includes the virtual manipulatives. The selection criteria for the virtual simulators utilized in the crossover phase required the that the simulator: (a) was free; (b) did not require students to set up an account (i.e. did not require a username or password); and (c) was not too "glitchy" (i.e. did not freeze up or kick you out during the data collection).

As previously noted, the below average user experience for Lab 2 was in large part due to difficulties utilizing the simulator.

In the self-selection phase, students were given the opportunity to explore another type of virtual experience. Lab 4 (*Correlation and Causation*) incorporated the use of a pre-collected dataset. While respondents appreciated the convenience of the VL, they were not fully satisfied with this passive laboratory experience. This is consistent with similar studies that found students had trouble connecting to datasets that had been provided by the instructor; they felt that collecting the data was a critical component to understanding the laboratory goals and objectives (Finne et al., 2022; Hsu & Rowland-Goldsmith, 2020; Johnson & Barr, 2021; Klein et al., 2021). This implies that program chairs should advocate for high-quality VL simulations (or other VL activities) that incorporate opportunities for students to actively collect their own data.

High-quality simulations are expensive, ranging from $50-$100 per simulation per student (Senapati, 2022). The laboratory simulations could be purchased by students from a vendor or provided to students by the University. Fortunately, costs to the students and University can be defrayed by purchasing a university site license. Moreover, the simulations could be cost-shared across programs that include "laboratory-type" experiences (i.e. exercise science, health sciences, education, etc.). While high-quality simulations have proven to be an effective and engaging learning experience, numerous studies have found students preferred the VL be used in combination with hands-on experiences (Finne et al., 2022; Gnesdilow & Putambekar, 2022; Johnson & Barr, 2021; Sithole et al., 2022).

The adult learners in this study seemed to agree with these sentiments. As previously mentioned, respondents enjoyed the convenience and ease of use of the coin-flip simulator for Lab 1. However, respondents for Lab 2 would have been more comfortable using a real ruler to measure real objects. Vendors have also noticed; several laboratory kit vendors now provide a blend of VL and AHHO experiences (personal communication, quotes from vendors, 2022).

While it is important to control costs for the University and to students (especially for adult learners), the quality of the laboratory experience should not be limited by the lack of home supplies or by the poor quality of the free simulators. As such, this deeper understanding of adult learners' experiences can be used to help lead faculty and program chairs better manage their science curriculum and the costs for laboratory resources.

The next section will discuss the findings for mixed-methods research question 2: Which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) play a significant role in participants' user experiences? Again, the research question was supported by evidence from both Phase I (quasi-experimental crossover phase) and Phase II (self-section phase).

Factors that Played a Significant Role in User Experiences

Phase I: Quasi-Experimental Crossover Phase

Each of the ten items on the modified SUS were coded as a measure of effectiveness, efficiency, engagement, ease of use, error tolerance, ease of use, or self-efficacy according to Lewis and Sauro's (2019) item descriptions. The mean raw item scores (MRIS), mean adjusted raw item scores (MARIS), and multiple regression analysis (SUS and MARIS) were used to determine if there was evidence to support that one or more factors played a significant role in the user experience.

Table 5.1 shows the summary of the Phase I sub-factor analysis. They include the results for: one-way ANOVA test (and non-parametric Kruskal-Wallis test) for the MRIS an MARIS analysis for trial-1 ($AHHO_1/VL_1$), trial-2 ($AHHO_2/VL_2$), Group A ($AHHO_1/VL_2$) and Group B

($\text{AHHO}_2/\text{VL}_1$); multiple regression analysis for AHHO and VL for trial-1 ($\text{AHHO}_1/\text{VL}_1$), trial-2 ($\text{AHHO}_2/\text{VL}_2$), and combined (trial-1 + trial-2); and the factors identified by hypothesis coding for the open-ended questions on the Lab 1 Survey and Lab 2 Survey. It should be noted that the hypothesis tests were conducted for the U-term and Q-term for each trial; the data for these sub-populations was consistent with the Overall dataset. For clarity in presentation, only the Overall dataset is presented in Table 5.1.

Table 5.1

Summary of Sub-Factor Analysis: Phase I – Quantitative and Qualitative Findings

Analysis	Comparison	E	F	N	ET	EU	SE
MRIS	Trial-1	√	√	√+	---	√+	√
	Trial-2	√	---	√	---	---	---
	Group A	√	---	√	---	---	√
	Group B	---	√	√	---	√	√
MARIS	Trial-1	√	√	√+	√	√+	√
	Trial-2	√	---	√	---	---	---
	Group A	√	√	√	---	---	√
	Group B	√	√	√	---	√	√
MRA							
AHHO	Trial-1	√+	√	√+	√	√+	√
	Trial-2	√+	√	√+	√	√	√
	Combined	√+	√	√+	√	√+	√
VL	Trial-1	√	√	---	√+	√+	√+
	Trial-2	---	---	√+	√	√+	√+
	Combined	√	√	√+	√	√+	√+
HC	Q3-Q8	√+	---	√+	---	√+	---
Overall		**	---	**	---	**	*

Note. E = effectiveness; F = efficiency; N = engagement; ET = error tolerance; EU = ease of use; SE = self-efficacy; MRIS = data from mean raw item score analysis; MARIS = data from mean adjusted raw item score analysis; MRA = data from multiple regression analysis; AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience; HC = data from qualitative hypothesis code analysis; √ = identified as important by one of the following: met benchmark (MRIS and MARIS), significant based on hypothesis test (MRIS, MARIS, and MRA), large partial slope (MRA), or large number of responses in at least two open-ended questions; √ + = identified as important by more than one of the above; ** = very important factor in user experience; * = important factor in user experience.

The summary table shows that effectiveness, engagement, and ease of use were identified as the primary factors that played an important role in the user-experience; self-efficacy (and efficiency to a lesser extent) was also important. It is worth noting that self-efficacy (or rather, low self-efficacy) played a more important role in the user-experience for Lab 2 than Lab 1. The objectives for Lab 2 were to collect and analyze measurement data (using either a real or virtual ruler); to use the correct number of significant figures for their specific measurement tool (to the 0.01 cm place); and to use their measurement data to investigate the strengths and weaknesses of the three measures of central tendency (mean, median, and mode). This researcher did not anticipate that the use of a metric ruler (real or virtual) or that simple calculations (mean and standard deviation) would have such a negative impact on respondents' belief in their ability to perform the specified tasks required to complete the laboratory experiment.

As previously discussed, participants perceived the Lab 2 simulator as difficult to use. Furthermore, a closer look at the open-ended responses revealed that: 14% of the responses indicated participants did not understand the purpose for (and were frustrated by) measuring the length, width, and height of the same object multiple times (and were shocked to discover that their belief that the measurements would/should be the same for all 3 trials was not correct); 30% of the responses indicated that they did not know how to (or had difficulty) reading the ruler (both real and virtual) to the 0.01 place; and 29% (of responses for those assigned to VL_2) did not see the relevance of measuring virtual objects.

This has two major instructional design implications. First, participants' lack of self-confidence in measurement skills indicated that the laboratory materials (laboratory protocol and/or ancillary instructional materials) should include a demonstration of all laboratory skills. While the instruction sheet for Lab 2 had a brief description of how to measure objects using the correct number of significant figures, it was assumed that participants knew how to use a metric ruler. This was not a good assumption as >85% of the respondents were adult learners (who may not have used a metric ruler since high school) and SCIE 211 is a non-major's course (for students that may not be used to making accurate measurements). Several respondents indicated they had to search for a tutorial on using a metric ruler; and that providing a video demonstration as preparation for the laboratory activities would have saved them time.

Additionally, participants failed to see the relevance of measuring virtual objects; the measured lengths, widths, and heights were not meaningful as they did not reflect reality. For example, one respondent shared their frustration by the following statement: *"measuring a couch that was only 8 cm was awkward"*. This observation is salient as adult learning theories emphasize the importance of using relevant, real-world examples (Chen, 2014; Knowles, 1980, 1984; Lewis & Bryan, 2021). This means that virtual tools should simulate real-life experiences.

Phase II: Self-Selection Phase

Participants were asked to indicate which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and/or self-efficacy) they considered to be the most important (and least important) for a good laboratory experience. The results for the most important factors were consistent with the quantitative findings (MRIS, MARIS, and multiple regression analysis) and the Phase I qualitative findings: effectiveness, engagement, and ease of use. Respondents also identified efficiency as an important factor. This is not surprising as adult learners have a significant number of time commitments (Bowers & Bergman, 2016; Osam et al., 2017); maintaining work/school/life balance is challenged when laboratory experiences take more time than anticipated (Al-Soufi et al., 2020; Andrews et al., 2020; Doughan & Shamuradyan, 2021; Moosvi et al., 2019). Adult learners struggle for time-balance is exemplified by the following quote from a respondent: *"I work 40+ hours a week and repeating or starting over would cost me a lot of time I do not have".*

It was surprising that engagement was identified as one of the most important factors and as one of the least important factors. The importance of engagement seemed to be negatively correlated with age. Respondents in the <25, 25-29, and 30-34 age groups tended to indicate that engagement was more important than respondents in the 35-39, 40-44, 45-49, 50-54, and 55+ age groups. Participants that were more than 35 years old reported that they did not see the laboratory experience as something to be enjoyed, it was just another assignment that had to be completed. This implies that older adults do not need to be entertained; they just want a simple lab experience with clear instructions that can be completed in a timely manner.

Respondents were also asked to identify other important factors for a good laboratory experience. All of the 'other factors' respondents identified corresponded to either one of the hypothesis codes or to one of the codes that emerged from the Phase I data analysis (or both). The alignment of the 'other factors' with those already identified during the coding process means that the open-ended questions used in Phase I and Phase II effectively identified all of the factors that were important to respondents. The complementarity of the responses also lends an air of trustworthiness to the coding methodology and data analysis.

While the quantitative and qualitative arrived at the same conclusions (AHHO≈VL; the most important factors were effectiveness, engagement, and ease of use; other important factors were efficiency and self-efficacy), these findings are most likely not generalizable. The following sections will discuss the study's limitations and delimitations.

Limitations on the Study

A major limitation of the study was laboratory materials. The only difference between the laboratory protocols in the crossover phase was that the hands-on experiment utilized real objects while the virtual experiment utilized virtual objects. Therefore, it was assumed that student perception data would reflect their preference for manipulating physical objects versus virtual objects. However, participants' lack of access to the household items necessary to complete the hands-on experiment (coins and a ruler) and technical issues with the simulators for the virtual experiment impacted their perceptions. This contributed to the study's other major limitation.

Although the protocols had been reviewed for content validity (and had been used in previous terms), Table 5.2 shows that the user experience for Lab 1 >> Lab 2.

Table 5.2

Summary of Crossover Experience: Mean SUS and Grade

Group	Lab 1	Mean SUS	Grade	Lab 2	Mean SUS	Grade
A	AHHO	69.039 ± 21.421	C	VL	55.294 ± 20.824	D
B	VL	75.778 ± 17.563	B	AHHO	60.239 ± 20.044	D

Note. AHHO = at-home, hands-on laboratory experience; VL = at-home, virtual laboratory experience. Each lab was assigned a grade based on the curved grading scale developed by Sauro and Lewis (2016).

The crossover design was implemented to control for order-bias of the AHHO and VL experiences. Lindsay et al. (2009) posited that a students' first interaction with equipment builds a mental model of the laboratory experience and establishes the reality of what a 'laboratory experience' should be. As such, it could be argued that the Lab 2 user experience was not as positive as the Lab 1 user experience (for both crossover group A and B) because the second lab experiences were not coherent with the participants' mental model of their first laboratory experiences. However, based upon the students' open-ended responses, if order-bias occurred, it was masked by participants overall dissatisfaction with the clarity of instructions, the use of the metric ruler, and the number and complexity of the mathematical calculations for Lab 2. Due to the large disparity between the nature of the laboratory experiments, the researcher was not able to determine if the order of AHHO and VL experiences affected participants' perception.

Delimitations of the Study

A major delimitation of the study was that data was collected from a convenience sample of students registered for SCIE 211. Convenience samples are prone to sampling bias and the findings are not generalizable (Pajo, 2017). While efforts were made to minimize sampling bias (by using the learning management system to randomly assign students to Group A and Group B in the crossover phase), all of the adult learners surveyed in this study were from the same research site. As such, their perceptions may not be representative of adult learners at another institution.

Other efforts to minimize bias included: replicating the study in both the U-term and Q-term in Fall 2022; employing a mixed-methods design in which multiple data strands were utilized (and corroborated) in an effort to answer each of the research questions; utilizing a normed and validated tool to collect quantitative data; and gaining permission to utilize open-ended questions that had been tested in a previous study. While these efforts may increase the reliability (and trustworthiness) of the findings, ultimately, the data was still collected from a convenience sample.

Another delimitation of the study was the laboratory protocols themselves. The laboratory protocols were designed so that students could collect data using common household items or free simulators. As such, experiments that required specialized equipment or a more sophisticated laboratory simulation were not considered for use in this study. Consequently, the modified SUS questionnaire was only tested for simple laboratory experiments that required simple manipulatives. Therefore, the findings may not be generalizable to more advanced laboratory protocols. This (and other recommendations for future research) will be discussed in the next section.

Recommendations for Future Research

The modified SUS was shown to be an effective tool to measure the perceived usability of the SCIE 211 laboratory experiments and to identify which factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy) played an important role in the Lab 1 and Lab 2 user experiences. However, this version of the modified SUS has only been tested in this study. The instrument needs to be further tested to determine if it is an effective tool for other non-major's science laboratories (at this research site and other institutions) and majors-level science laboratories (at other institutions and across the different disciplines).

This researcher would like to replicate this study in a non-majors Anatomy and Physiology course. While the students registered in this course are not science majors, the course is required for science-related majors (exercise science and health sciences). To test the efficacy of the modified SUS, it necessary to investigate if it is as effective for a different set of non-major’s science laboratories; and to see if the perceived usability (and factors for a good laboratory experience) are influenced by participants’ interest in a science-related field.

Conclusions

This study investigated using a modified version of Brooke’s (1996) System Usability Scale (SUS) as a framework to determine adult learners preferred at-home laboratory experience. While the modified SUS did not provide a clear answer to the question of preferred laboratory experience (as both the Phase I and Phase II results showed AHHO $\approx$ VL), the quantitative and qualitative analyses clearly identified effectiveness, engagement, and ease of use as the most important factors for a good laboratory experience. Furthermore, the instrument was sensitive enough to detect which factors (efficiency and self-efficacy) contributed to the below average user experience for Lab 2. These findings provided valuable insights into adult learners’ laboratory needs and expectations.

A deeper understanding of adult learners’ laboratory needs is especially important for Program Chairs at primarily online institutions. Although these institutions cater to working students over the age of 25 (Chen, 2014; NCES, 2021b), most courses are still designed for traditional students (Chen, 2014; Greene & Larsen, 2018; Lewis & Bryan, 2021; Remenick & Goralnik, 2019). Studies have shown that teaching approaches that are considered to be highly effective in traditional student populations may be not be as effective (or even ineffective) with an adult student population (Arghode et al., 2017; Chen, 2014; Gutruf et al., 2021; Knowles, 1984; Lewis & Bryan, 2021; Loeng, 2018; Pratt, 1993; Remenick & Goralnik, 2019; Taylor & Hamdy, 2013). As Program Chairs are expected to maintain relevant, high-quality courses that meet the specific needs of the students at their institution (Weaver et al., 2019), the findings of this study can be used to support the design of student-centered online science courses.

Although the results for adult learners' preferred laboratory experience were inconclusive, the finding of 'nearly the same' can still be used to inform instructional design decisions. On the one hand, no significant difference means that respondents found instructional value in both the hands-on and virtual laboratory experiences. On the other hand, the findings of the SUS and sub-factor analysis revealed that the quality of the adult learners' experiences were heavily influenced by the perceived ease of use of the laboratory materials and the perceived effectiveness (or clarity) of the laboratory protocol. These findings were consistent with similar studies that found student interest and engagement in the laboratory experience were more closely tied to the quality of the experimental design and materials than to the laboratory medium (Johnson & Barr, 2021; Koretsky et al., 2011; Nolan & Koretsky, 2018; Son & Narguizian, 2016). This implies that laboratory medium (AHHO or VL) is not as important to adult learners as the clarity of the laboratory instructions, access to ancillary instructional materials (skills demonstration videos/problem-solving examples), and the usability of the laboratory materials.

Furthermore, 'materials' played a more significant role in the user-experience than anticipated. One of the assumptions of the study was that students would have access to the supplies necessary to complete the lab as they are common household items (coins, a ruler, and balloons) and that the students registered in the online course would have the technical capabilities to operate the simple simulators. However, AHHO respondents remarked that finding a coin or a ruler (in the crossover phase) or having to go to the store to purchase balloons (in the self-selection phase) was a huge dissatisfier; and, VL respondents reported difficulties using the simulator for Lab 2 (and were not satisfied with the passive experience of being given a dataset for Lab 4). This implied that supplies for AHHO experiences should be provided to students in an easy-to-access customized kit; and that program chairs should advocate for access to high-quality VL simulations (or other VL activities) that incorporate opportunities for students to actively collect their own data.

Moreover, while respondents did seem to enjoy the convenience and ease of use of the coin-flip simulator for Lab 1; most would have been more comfortable using a real ruler to measure real objects to complete Lab 2. As such, a blended approach may be best (i.e. each laboratory activity should utilize the manipulative, physical or virtual, that best fits the task at hand). This is consistent with previous studies that found students preferred a combination of VL and AHHO experiences (Finne et al., 2022; Gnesdilow & Putambekar, 2022; Johnson & Barr, 2021; Sithole et al., 2022).

Laboratory kit vendors have also noticed this trend; several now offer customized kits that include a blend of high-quality simulations (and/or virtual reality experiences) and all of the supplies for the hands-on activities. Depending on the vendor, the cost is commensurate with the amount assessed for the lab fee. Applying what was learned from the analysis of the modified SUS, these kits should provide a cost-effective laboratory experience that adult learners in SCIE 211 should find effective, engaging, and easy to use.

Additionally, as the self-selection data and crossover data arrived at the same conclusion, the modified SUS proved to be an effective instrument to measure perceived usability, its sub-factors (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy), and the quality of the user experience for laboratory experiments. The SUS is a normed, validated, and product-agnostic questionnaire (Bevan, 1995; Brooke, 2013; Lewis, 2018a) that has been used to measure the perceived usability of a wide range of products (Lacerda & von Wangenheim, 2018; Lewis, 2018b; Lewis & Sauro, 2019). Although this tool has been used to evaluate students' perceptions of various educational technologies (Granić & Ćukušić, 2011;

Harrati et al., 2016; Orfanou et al., 2015; Vlachogianni & Tselios, 2021, 2022), to the researcher’s knowledge, the SUS has never been used to evaluate user experiences for laboratory experiments.

Although the modified SUS used in this study needs to be further tested, these findings are exciting for several reasons. First, in just 10-simple questions, this instrument was able to accurately determine the overall quality of the adult learners’ experience for both Lab 1 and Lab 2 based on the curved grading scale developed by Sauro and Lewis (2016); and was able to accurately determine which factors played an important role in the Lab 1 and Lab 2 adult learners’ experiences based on the mean raw item score benchmarks developed Lewis and Sauro (2019). Furthermore, the factors that can be measured by this tool (effectiveness, efficiency, engagement, error tolerance, ease of use, and self-efficacy) are closely aligned with the variables that support adult learners’ independent learning styles as reported in the literature (Arghode et al., 2017; Calabrese & Capraro, 2021; Chen, 2014; Knowles, 1980, 1984); and as reported by the adult learners in this study.

Finally, as the questions are not specific to any course or discipline, (if the results prove true with additional testing) the modified SUS can be used to evaluate students’ perceptions across a wide-array of laboratory experiences, across the curriculum, across disciplines, and across institutions. In the sage and timeless words of V. F. Ridgway (1956), what gets measured (correctly) gets managed (appropriately). As applied to the study, this simple, yet powerful, instrument could be used by program chairs at primarily online universities to assess their learners’ preferences correctly and manage their curriculum appropriately.

References

Achuthan, K., Raghavan, D., Shankar, B., Francis, S. P., & Kolil, V. K. (2021). Impact of remote experimentation, interactivity and platform effectiveness on laboratory learning outcomes. *International Journal of Educational Technology in Higher Education*, *18*(1). https://doi.org/10.1186/s41239-021-00272-z

Adero, F. A., & Odiyo, W. O. (2020). Leadership theories and the desired competencies of the 21st century organization leader. *International Journal of Business Management, Entrepreneurship and Innovation*, *2*(3), 16–35. https://doi.org/10.35942/jbmed.v2i3.115

Al-Soufi, W., Carrazana-Garcia, J., & Novo, M. (2020). When the kitchen turns into a physical chemistry lab. *Journal of Chemical Education*, *97*(9), 3090–3096. https://doi.org/10.1021/acs.jchemed.0c00745

Almalki, S. (2016). Integrating quantitative and qualitative data in mixed methods research—Challenges and benefits. *Journal of Education and Learning*, *5*(3), 288–296. https://doi.org/10.5539/jel.v5n3p288

Almeida, F. (2018). Strategies to perform a mixed methods study. *European Journal of Educational Studies*, *5*(1), 137–151. https://doi.org/10.5281/zenodo.1406214

Ambusaidi, A., Al Musawi, A., Al-Balushi, S., & Al-Balushi, K. (2018). The impact of virtual lab learning experiences on 9th grade students' achievement and their attitudes towards science and learning by virtual lab. *Journal of Turkish Science Education*, *15*(2), 13–29. https://doi.org/10.12973/tused.10227a

Andrews, J. L., de Los Rios, J. P., Rayaluru, M., Lee, S., Mai, L., Schusser, A., & Mak, C. H. (2020). Experimenting with at-home general chemistry laboratories during the COVID-

19 pandemic. *Journal of Chemical Education*, *97*, 1887–1894. https://doi.org/10.1021/acs.jchemed.0c00483

Arghode, V., Brieger, E. W., & McLean, G. N. (2017). Adult learning theories: Implications for online instruction. *European Journal of Training and Development*, *41*(7), 593–609. https://doi.org/10.1108/ejtd-02-2017-0014

Attardi, S. M., Choi, S., Barnett, J., & Rogers, K. A. (2016). Mixed methods student evaluation of an online systemic human anatomy course with laboratory. *Anatomical Sciences Education*, *9*(3), 272–285. https://doi.org/10.1002/ase.1584

Attardi, S. M., & Rogers, K. A. (2015). Design and implementation of an online systemic human anatomy course with laboratory. *Anatomical Sciences Education*, *8*(1), 53–62. https://doi.org/10.1002/ase.1465

Avcı, F. (2022). Teaching the "acid–base" subject in biochemistry via virtual laboratory during the COVID -19 pandemic. *Biochemistry and Molecular Biology Education*, *50*(3), 312–318. https://doi.org/10.1002/bmb.21625

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

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Barbeau, M. L., Johnson, M., Gibson, C., & Rogers, K. A. (2013). The development and assessment of an online microscopic anatomy laboratory course. *Anatomical Sciences Education*, *6*(4), 246–256. https://doi.org/10.1002/ase.1347

Barthet, M. M. (2021). Teaching molecular techniques at home: Molecular biology labs that can be performed anywhere and enable hands-on learning of restriction digestion/ligation and

DNA amplification. *Biochemistry and Molecular Biology Education*, *49*(4), 598–604. https://doi.org/10.1002/bmb.21517

Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal*, *4*(2), 115–130. https://doi.org/10.1007/bf00402715

Bhute, V. J., Inguva, P., Shah, U., & Brechtelsbauer, C. (2021). Transforming traditional teaching laboratories for effective remote delivery—A review. *Education for Chemical Engineers*, *35*, 96–104. https://doi.org/10.1016/j.ece.2021.01.008

Blotnicky, K. A., Franz-Odendaal, T., French, F., & Joy, P. (2018). A study of the correlation between STEM career knowledge, mathematics self-efficacy, career interests, and career activities on the likelihood of pursuing a STEM career among middle school students. *International Journal of STEM Education*, *5*(1). https://doi.org/10.1186/s40594-018-0118-3

Bortnik, B., Stozhko, N., Pervukhina, I., Tchernysheva, A., & Belysheva, G. (2017). Effect of virtual analytical chemistry laboratory on enhancing student research skills and practices. *Research in Learning Technology*, *25*(0), 1–20. https://doi.org/10.25304/rlt.v25.1968

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. https://doi.org/10.19173/irrodl.v14i3.1446

Bowers, A., & Bergman, M. (2016). Affordability and the return on investment of college completion: Unique challenges and opportunities for adult learners. *The Journal of Continuing Higher Education*, *64*(3), 144–151. https://doi.org/10.1080/07377363.2016.1229102

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. https://doi.org/10.1016/j.compedu.2015.07.003

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40. https://www.researchgate.net/publication/285811057_SUS_a_retrospective

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, *1*(1), 8–22. https://doi.org/10.1177/2345678906290531

Calabrese, J. E., & Capraro, R. M. (2021). The autonomy of informal STEM and benefits of andragogy with gifted children. *Journal of Research in Innovative Teaching & Learning*. https://doi.org/10.1108/jrit-08-2021-0060

Chang, M., Liao, M., Lue, J., & Yeh, C. (2022). The impact of asynchronous online anatomy teaching and smaller learning groups in the anatomy laboratory on medical students' performance during the Covid-19 pandemic. *Anatomical Sciences Education*, *15*(3), 476–492. https://doi.org/10.1002/ase.2179

Chen, J. C. (2014). Teaching nontraditional adult students: Adult learning theories in practice. *Teaching in Higher Education*, *19*(4), 406–418. https://doi.org/10.1080/13562517.2013.860101

Chen, S. H. (2022). Learning lab skills online: Lessons from implementing video-based instruction for a remote biotechnology lab. *Biochemistry and Molecular Biology Education*, *50*(6), 633–640. https://doi.org/10.1002/bmb.21667

Chen, S., Chang, W.-H., Lai, C.-H., & Tsai, C.-Y. (2014). A comparison of students' approaches to inquiry, conceptual learning, and attitudes in simulation-based and microcomputer-based laboratories. *Science Education*, *98*(5), 905–935. https://doi.org/10.1002/sce.21126

Childers, G., & Jones, M. G. (2017). Learning from a distance: High school students' perceptions of virtual presence, motivation, and science identity during a remote microscopy investigation. *International Journal of Science Education*, *39*(3), 257–273. https://doi.org/10.1080/09500693.2016.1278483

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*(4), 417-451.

Çivril, H., & Özkul, A. E. (2021). Investigation of the factors affecting open and distance education learners' intentions to use a virtual laboratory. *The International Review of Research in Open and Distributed Learning*, *22*(2), 143–165. https://doi.org/10.19173/irrodl.v22i2.5076

Clough, M. (2002). Using the laboratory to enhance student learning. In R. W. Bybee (Ed.), *Learning science and the science of learning* (pp. 85–94). National Science Teachers Association.

Corter, J., Nickerson, J., Esche, S., Chassapis, C., Im, S., & Ma, J. (2007). Constructing reality: A study of remote, hands-on, and simulated laboratories. *ACM Transactions on Computer-Human Interaction*, *14*(2). https://doi.org/10.1145/1275511.1275513

Cossovich, R., Hargis, J., & Chun, H. (2020). Working with electrons: Integrating "kits" for hands-on online learning in homes. *The Online Journal of New Horizons in Education*, *10*(4), 232–246. http://www.tojned.net/journals/tojned/volumes/tojned-volume10-i04.pdf#page=39

Cote, R. (2017). Vision of effective leadership. *International Journal of Business Administration*, *8*(6), 1–10. https://doi.org/10.5430/ijba.v8n6p1

Crandall, P. G., O'Bryan, C. A., Killian, S. A., Beck, D. E., Jarvis, N., & Clausen, E. (2015). A comparison of the degree of student satisfaction using a simulation or a traditional wet lab to teach physical properties of ice. *Journal of Food Science Education*, *14*(1), 24–29. https://doi.org/10.1111/1541-4329.12049

Creswell, J. W. (2022). *A concise introduction to mixed methods research* (2nd ed.). Sage.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications, Inc.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage.

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. https://doi.org/10.1016/j.compedu.2009.05.005

Davenport, J. L., Rafferty, A. N., & Yaron, D. J. (2018). Whether and how authentic contexts using a virtual chemistry lab support learning. *Journal of Chemical Education*, *95*(8), 1250–1259. https://doi.org/10.1021/acs.jchemed.8b00048

Davidson, K. J., Haddrill, P., Casali, F., Murphy, B., Gibson, L., Robinson, M., Clunie, A., Christie, J., Curran, L., & Carlysle-Davies, F. (2022). Lockdown labs: Pivoting to remote learning in forensic science higher education. *Science & Justice*, *62*(2022), 805–813. https://doi.org/10.21428/cb6ab371.3e2a4e98

Davies, K. (2019). The efficacy of flipped laboratory multiperspective videos in skill acquisition. *Journal of College Science Teaching*, *049*(02), 62–67. https://doi.org/10.2505/4/jcst19_049_02_62

de Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and virtual laboratories in science and engineering education. *Science*, *340*(6130), 305–308. https://doi.org/10.1126/science.1230579

de Leeuw, E.D. and Lugtig, P. (2015). Dropouts in longitudinal surveys. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat06661.pub2

de Toledo Durand, M., Baraldi, C., Restini, A., Wolff, A., Faria, M., Couto, L., & Bestetti, R. (2019). Students' perception of animal or virtual laboratory in physiology practical classes in PBL medical hybrid curriculum. *Advances in Physiological Education*, *43*, 451–457. https://doi.org/10.1152/advan.00005.2019

Deboer, G. (1991). *A history of ideas in science education.* Teachers College Press.

DeChenne-Peters, S. E., Sargent, E., Mateer, S., Machingura, M., Zettler, J., Ness, T., DeMars, G., Cannon, S., & BrofttBailey, J. (2022). Comparison of student outcomes in a course-based undergraduate research experience: Face-to-face, hybrid, and online delivery of a biology laboratory. *International Journal for the Scholarship of Teaching and Learning*, *16*(1). https://doi.org/10.20429/ijsotl.2022.160105

Desa, S., Zuhaidi, N. D., & Nordin, N. (2022). The transition from the biochemistry laboratory to home discovery during COVID -19. *Biochemistry and Molecular Biology Education*, *50*(4), 401–402. https://doi.org/10.1002/bmb.21636

Dewhurst, D. G., Macleod, H. A., & Norris, T. A. M. (2000). Independent student learning aided by computers: An acceptable alternative to lectures? *Computers & Education*, *35*(3), 223–241. https://doi.org/10.1016/s0360-1315(00)00033-6

DiBiase, D. (2000). Is distance teaching more work or less work? *American Journal of Distance Education*, *14*(3), 6–20. https://doi.org/10.1080/08923640009527061

Doughan, S., & Shahmuradyan, A. (2021). At-home real-life sample preparation and colorimetric-based analysis: A practical experience outside the laboratory. *Journal of Chemical Education*, *98*(3), 1031–1036. https://doi.org/10.1021/acs.jchemed.0c01299

Easdon, J. (2020). Stay at home laboratories for chemistry courses. *Journal of Chemical Education*, *97*(9), 3070–3073. https://doi.org/10.1021/acs.jchemed.0c00760

Elmer, S. J., Carter, K. R., Armga, A. J., & Carter, J. R. (2016). Blended learning within an undergraduate exercise physiology laboratory. *Advances in Physiology Education*, *40*(1), 64–69. https://doi.org/10.1152/advan.00144.2015

Erasmus, D. J., Brewer, S. E., & Cinel, B. (2014). Assessing the engagement, learning, and overall experience of students operating an atomic absorption spectrophotometer with remote access technology. *Biochemistry and Molecular Biology Education*, *43*(1), 6–12. https://doi.org/10.1002/bmb.20838

Faulconer, E. K., & Gruss, A. B. (2018). A review to weigh the pros and cons of online, remote, and distance science laboratory experiences. *The International Review of Research in Open and Distributed Learning*, *19*(2), 155–168. https://doi.org/10.19173/irrodl.v19i2.3386

Fetters, M. D. (2020). *The mixed methods research workbook : activities for designing, implementing, and publishing projects*. Sage.

Fetters, M. D., & Freshwater, D. (2015). The 1 + 1 = 3 integration challenge. *Journal of Mixed Methods Research*, *9*(2), 115–117. https://doi.org/10.1177/1558689815581222

Finne, L. T., Gammelgaard, B., & Christiansen, F. V. (2022). When the lab work disappears: Students' perception of laboratory teaching for quality learning. *Journal of Chemical Education*, *99*(4), 1766–1774. https://doi.org/10.1021/acs.jchemed.1c01113

Folmer, E., & Bosch, J. (2004). Architecting for usability: a survey. *Journal of Systems and Software*, *70*(1-2), 61–78. https://doi.org/10.1016/s0164-1212(02)00159-0

Galustyan, O. V., Borovikova, Y. V., Polivaeva, N. P., Bakhtiyor, K. R., & Zhirkova, G. P. (2019). E-learning within the Field of Andragogy. *International Journal of Emerging Technologies in Learning (IJET)*, *14*(09), 148–156. https://doi.org/10.3991/ijet.v14i09.10020

Gnesdilow, D., & Puntambekar, S. (2022). Comparing middle school students' science explanations during physical and virtual Laboratories. *Journal of Science Education and Technology*, *31*(2022), 191–202. https://doi.org/10.1007/s10956-021-09941-0

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, *44*(1), 218–243. https://doi.org/10.3102/0091732x20903302

Goudsouzian, L. K., Riola, P., Ruggles, K., Gupta, P., & Mondoux, M. A. (2018). Integrating cell and molecular biology concepts: Comparing learning gains and self-efficacy in corresponding live and virtual undergraduate laboratory experiences. *Biochemistry and Molecular Biology Education*, *46*(4), 361–372. https://doi.org/10.1002/bmb.21133

Granić, A., & Ćukušić, M. (2011). Usability testing and expert inspections complemented by educational evaluation: A case study of an e-Learning platform. *Educational Technology*

& Society, *14*(2), 1176–3647. http://www.jstor.org/stable/jeductechsoci.14.2.107

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274. https://doi.org/10.3102/01623737011003255

Greene, K., & Larsen, L. (2018). Virtual andragogy: A new paradigm for serving adult online learners. *International Journal for Digital Society*, *9*(2), 1376–1381. https://doi.org/10.20533/ijds.2040.2570.2018.0169

Gunawan, G., Harjono, A., Hermansyah, H., & Herayanti, L. (2019). Guided inquiry model through virtual laboratory to enhance students' science process skills on heat concept. *Cakrawala Pendidikan*, *38*(2), 259–268. https://doi.org/10.21831/cp.v38i2.23345

Gutruf, P., Utzinger, U., & Subbian, V. (2021). Moving from pedagogy to andragogy in biomedical engineering design: Strategies for lab-at-home and distance learning. *Biomedical Engineering Education*, *1*(2), 301–305. https://doi.org/10.1007/s43683-020-00043-8

Hamed, G., & Aljanazrah, A. (2020). The effectiveness of using virtual experiments on students' learning in the general physics lab. *Journal of Information Technology Education: Research*, *19*, 977–996. https://doi.org/10.28945/4668

Harrati, N., Bouchrika, I., Tari, A., & Ladjailia, A. (2016). Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis. *Computers in Human Behavior*, *61*(2016), 463–471. https://doi.org/10.1016/j.chb.2016.03.051

Heng, Z. S., Koh, D. W., Yeo, J. Y., Ooi, C., & Gan, S. K. (2022). Effects of different delivery modes on teaching biomedical science practical skills in higher education during the 2021 pandemic measures. *Biochemistry and Molecular Biology Education*, *50*(4), 403–413. https://doi.org/10.1002/bmb.21637

Heradio, R., Luis, Galan, D., Cabrerizo, F. J., HerreraViedma, E., & Dormido, S. (2016). Virtual and remote labs in education: A bibliometric analysis. *Computers & Education*, *98*, 14–38. https://doi.org/https://doi.org/10.1016/j.compedu.2016.03.010

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

Hofstein, A., & Mamlok-Naaman, R. (2007). The laboratory in science education: The state of the art. *Chemistry Education Research and Practice*, *8*(2), 105–107. https://doi.org/10.1039/b7rp90003a

Honig, C. D., Sutton, C. C., & Bacal, D. M. (2022). Off-campus but hands-on: Mail out practicals with synchronous online activities during COVID-19. *Education for Chemical Engineers*, *39*(2022), 84–93. https://doi.org/10.1016/j.ece.2022.03.004

Howard, D., & Meier, M. (2021). Meeting laboratory course learning goals remotely via custom home experiment kits. *The Physics Teacher*, *59*(6), 404–409. https://doi.org/10.1119/5.0021600

Hsu, J. L., & Rowland-Goldsmith, M. (2020). Student perceptions of an inquiry-based molecular biology lecture and lab following a mid-semester transition to online teaching. *Biochemistry and Molecular Biology Education*, *49*(1), 15–25. https://doi.org/10.1002/bmb.21478

Husmann, P. R., O'Loughlin, V. D., & Braun, M. W. (2009). Quantitative and qualitative changes in teaching histology by means of virtual microscopy in an introductory course in human anatomy. *Anatomical Sciences Education*, *2*(5), 218–226. https://doi.org/10.1002/ase.105

International Organization for Standardization. (2018). *Usability: Definitions and concepts*. Iso.org. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Iordanoglou, D. (2018). Future trends in leadership development practices and the crucial leadership skills. *Journal of Leadership, Accountability and Ethics*, *15*(2). https://doi.org/10.33423/jlae.v15i2.648

Iwarsson, S., & Ståhl, A. (2003). Accessibility, usability and universal design—positioning and definition of concepts describing person-environment relationships. *Disability & Rehabilitation*, *25*(2), 57–66. https://doi.org/10.1080/713813473

Jara, C. A., Candelas, F. A., Torres, F., Dormido, S., Esquembre, F., & Reinoso, O. (2009). Real-time collaboration of virtual laboratories through the internet. *Computers & Education*, *52*(1), 126–140. https://doi.org/10.1016/j.compedu.2008.07.007

Javier, C. B., & Lomuntad, Ma. A. T. (2018). A study on the effectiveness of using computer simulations in teaching circuits vs the use of hands-on laboratory equipment. *Contemporary Social Sciences*, *27*(4), 120–134. https://doi.org/10.29070/27/58314

Johnson, J. E., & Barr, N. B. (2021). Moving hands-on mechanical engineering experiences online: Course redesigns and student perspectives. *Online Learning*, *25*(1). https://doi.org/10.24059/olj.v25i1.2465

Kapici, H. O., Akcay, H., & de Jong, T. (2019). Using hands-on and virtual laboratories alone or together―Which works better for acquiring knowledge and skills? *Journal of Science Education and Technology*, *28*(3), 231–250. https://doi.org/10.1007/s10956-018-9762-0

Kasworm, C. E. (2009). Adult learners in a research university: Negotiating undergraduate student identity. *Adult Education Quarterly*, *60*(2), 143–160. https://doi.org/10.1177/0741713609336110

Kelley, E. W. (2020). Reflections on three different high school chemistry lab formats during COVID-19 remote learning. *Journal of Chemical Education*, *97*(9), 2606–2616. https://doi.org/10.1021/acs.jchemed.0c00814

Kim, H., & Rehg, M. (2018). Faculty Performance and Morale in Higher Education: A Systems Approach. *Systems Research and Behavioral Science*, *35*(3), 308–323. https://doi.org/10.1002/sres.2495

Klahr, D., Triona, L. M. T., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching*, *44*(1), 183–203. https://doi.org/10.1002/tea.20152

Klein, P., Ivanjek, L., Dahlkemper, M. N., Jeličić, K., Geyer, M.-A. ., Küchemann, S., & Susac, A. (2021). Studying physics during the COVID-19 pandemic: Student assessments of learning achievement, perceived effectiveness of online recitations, and online laboratories. *Physical Review Physics Education Research*, *17*(1), 0101171–01011711. https://doi.org/10.1103/physrevphyseducres.17.010117

Knowles, M. S. (1980). *The Modern Practice of Adult Education: from Pedagogy to Andragogy*. Cambridge Adult Education.

Knowles, M. S. (1984). *Andragogy in action*. Jossey-Bass.

Koretsky, M. D., Amatore, D., Barnes, C., & Kimura, S. (2008). Enhancement of student learning in experimental design using a virtual laboratory. *IEEE Transactions on Education*, *51*(1), 76–85. https://doi.org/10.1109/te.2007.906894

Koretsky, M., Kelly, C., & Gummer, E. (2011). Student perceptions of learning in the laboratory: Comparison of industrially situated virtual laboratories to capstone physical laboratories. *Journal of Engineering Education*, *100*(3), 540–573. https://doi.org/10.1002/j.2168-9830.2011.tb00026.x

Lacerda, T. C., & von Wangenheim, C. G. (2018). Systematic literature review of usability capability/maturity models. *Computer Standards & Interfaces*, *55*, 95–105. https://doi.org/10.1016/j.csi.2017.06.001

Lewis, J. R. (2018a). The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307

Lewis, J. R. (2018b). Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. *International Journal of Human–Computer Interaction*, *35*(15), 1404–1419. https://doi.org/10.1080/10447318.2018.1533152

Lewis, J., & Sauro, J. (2019). Item benchmarks for the system usability scale. *Journal of Usability Studies*, *13*(3), 158–167. https://www.researchgate.net/profile/James-Lewis-8/publication/330225055_Item_Benchmarks_for_the_System_Usability_Scale/links/5c34fb41a6fdccd6b59ce868/Item-Benchmarks-for-the-System-Usability-Scale.pdf

Lewis, N., & Bryan, V. (2021). Andragogy and teaching techniques to enhance adult learners' experience. *Journal of Nursing Education and Practice*, *11*(11), 31–40. https://doi.org/10.5430/jnep.v11n11p31

Lindsay, E. D., & Good, M. C. (2005). Effects of laboratory access modes upon learning outcomes. *IEEE Transactions on Education*, *48*(4), 619–631. https://doi.org/10.1109/TE.2005.852591

Lindsay, E. D., Murray, S., Liu, D. K., Lowe, D. B., & Bright, C. G. (2009). Establishment reality vs. maintenance reality: how real is real enough? *European Journal of Engineering Education*, *34*(3), 229–234. https://doi.org/10.1080/03043790902902906

Loeng, S. (2018). Various ways of understanding the concept of andragogy. *Cogent Education*, *5*(1), 1–15. https://doi.org/10.1080/2331186x.2018.1496643

Londino-Smolar, G., & Hansel, C. (2021). Let's solve it: designing an interactive online forensic science lab. *Journal of Higher Education Theory and Practice*, *21*(2), 73-88. https://doi.org/10.33423/jhetp.v21i2.4121

Lowe, D., Newcombe, P., & Stumpers, B. (2013). Evaluation of the use of remote laboratories for secondary school science education. *Research in Science Education*, *43*(3), 1197–1219. https://doi.org/10.1007/s11165-012-9304-3

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

Magin, D., & Kanapathipillai, S. (2000). Engineering students' understanding of the role of experimentation. *European Journal of Engineering Education*, *25*(4), 351–358. https://doi.org/10.1080/03043790050200395

Martínez, G., Naranjo, F. L., Pérez, Á. L., Suero, M. I., & Pardo, P. J. (2011). Comparative study of the effectiveness of three learning environments: Hyper-realistic virtual simulations, traditional schematic simulations and traditional laboratory. *Physical Review Special*

Topics - Physics Education Research, *7*(2). https://doi.org/10.1103/physrevstper.7.020111

Mawn, M. V., Carrico, P., Charuk, K., Stote, K. S., & Lawrence, B. (2011). Hands-on and online: Scientific explorations through distance learning. *Open Learning: The Journal of Open, Distance and E-Learning*, *26*(2), 135–146. https://doi.org/10.1080/02680513.2011.567464

McKim, C. A. (2017). The value of mixed methods research. *Journal of Mixed Methods Research*, *11*(2), 202–222. https://doi.org/10.1177/1558689815607096

Meagher, B. J., Cataldo, K., Douglas, B. J., McDaniel, M. A., & Nosofsky, R. M. (2018). Training of rock classifications: The use of computer images versus physical rock samples. *Journal of Geoscience Education*, *66*(3), 221–230. https://doi.org/10.1080/10899995.2018.1465756

Meintzer, C., Sutherland, F., & Kennepohl, D. (2017). Evaluation of student learning in remotely controlled instrumental analyses. *The International Review of Research in Open and Distributed Learning*, *18*(6). https://doi.org/10.19173/irrodl.v18i6.3093

Miller, T. A., Carver, J. S., & Roy, A. (2018). To go virtual or not to go virtual, that is the question. *Journal of College Science Teaching*, *48*(2), 59-67. https://www.jstor.org/stable/2661627

Moosvi, F., Reinsberg, S., & Rieger, G. (2019). Can a hands-on physics project lab be delivered effectively as a distance lab? *The International Review of Research in Open and Distributed Learning*, *20*(1), 21–42. https://doi.org/10.19173/irrodl.v20i1.3782

National Center for Education Statistics. (2021a). *Digest of education statistics, 2021: Number and percentage of students enrolled in degree-granting postsecondary institutions, by* *distance education participation, location of student, level of enrollment, and control and level of institution: Fall 2019 and fall 2020*. Nces.ed.gov. https://nces.ed.gov/programs/digest/d21/tables/dt21_311.15.asp

National Center for Education Statistics. (2021b). *Digest of education statistics, 2021: Total fall enrollment in degree-granting postsecondary institutions, by level of enrollment, control and level of institution, attendance status, and age of student: 2019*. Nces.ed.gov. https://nces.ed.gov/programs/digest/d21/tables/dt21_303.50.asp

National Center for Education Statistics (2021c). *Digest of education statistics, 2021: Average undergraduate tuition, fees, room, and board charges for full-time students in degree-granting postsecondary institutions, by control and level of institution and state or jurisdiction: 2019-20 and 2020-21.* Nces.ed.gov. https://nces.ed.gov/programs/digest/d21/tables/dt21_330.20.asp

National Center for Education Statistics (2021d). *Distance education in college: What do we know from IPEDS?* Nces.ed.gov. https://nces.ed.gov/blogs/nces/post/distance-education-in-college-what-do-we-know-from-ipeds

National Center for Education Statistics (n.d.). *Nontraditional undergraduates/Definitions and data.* Nces.ed.gov. https://nces.ed.gov/pubs/web/97578e.asp

Neuwirth, L. S., Jović, S., & Mukherji, B. R. (2020). Reimagining higher education during and post-COVID-19: Challenges and opportunities. *Journal of Adult and Continuing Education*, *27*(2), 147797142094773. https://doi.org/10.1177/1477971420947738

Nguyen, J. G., & Keuseman, K. J. (2020). Chemistry in the kitchen laboratories at home. *Journal of Chemical Education*, *97*(9), 3042–3047. https://doi.org/10.1021/acs.jchemed.0c00626

Nolen, S. B., & Koretsky, M. D. (2018). Affordances of virtual and physical laboratory projects for instructional design: Impacts on student engagement. *IEEE Transactions on Education*, *61*(3), 226–233. https://doi.org/10.1109/te.2018.2791445

Olympiou, G., & Zacharia, Z. C. (2011). Blending physical and virtual manipulatives: An effort to improve students' conceptual understanding through science laboratory experimentation. *Science Education*, *96*(1), 21–47. https://doi.org/10.1002/sce.20463

Orfanou, K., Tselios, N., & Katsanos, C. (2015). Perceived usability evaluation of learning management systems: Empirical evaluation of the system usability scale. *The International Review of Research in Open and Distributed Learning*, *16*(2). https://doi.org/10.19173/irrodl.v16i2.1955

Orozco, D. (2017). An evaluation of eScience lab kits for online learning. *Themis: Research Journal of Justice Studies and Forensic Science*, *5*(1), 115–136. https://doi.org/10.31979/themis.2017.0508

Osam, E. K., Bergman, M., & Cumberland, D. M. (2016). An integrative literature review on the barriers impacting adult learners' return to college. *Adult Learning*, *28*(2), 54–60. https://doi.org/10.1177/1045159516658013

Oser, R., & Fraser, B. J. (2015). Effectiveness of virtual laboratories in terms of learning environment, attitudes and achievement among high-school genetics students. *Curriculum and Teaching*, *30*(2), 65–80. https://doi.org/10.7459/ct/30.2.05

Ott, R. L. & Longnecker, M. (2016). *An introduction to statistical methods & data analysis* (7th ed). Cengage.

Pajo, B. (2017). *Introduction to research methods: A hands-on approach*. Sage Publications.

Pal, D., & Vanijja, V. (2020). Perceived usability evaluation of Microsoft Teams as an online learning platform during COVID-19 using system usability scale and technology acceptance model in India. *Children and Youth Services Review*, *119*, 105525. https://doi.org/10.1016/j.childyouth.2020.105535

Paxinou, E., Panagiotakopoulos, C. T., Karatrantou, A., Kalles, D., & Sgourou, A. (2020). Implementation and evaluation of a three-dimensional virtual reality biology lab versus conventional didactic practices in lab experimenting with the photonic microscope. *Biochemistry and Molecular Biology Education*, *48*(1), 21–27. https://doi.org/10.1002/bmb.21307

Pheeney, P. (1997). Hands-on, minds-on: Activities to engage our students. *Science Scope*, *21*(4), 30-33.

Pratt, D. D. (1993). Andragogy after twenty-five years. *New Directions for Adult and Continuing Education*, *1993*(57), 15–23. https://doi.org/10.1002/ace.36719935704

Pyatt, K., & Sims, R. (2011). Virtual and physical experimentation in inquiry-based science labs: Attitudes, performance and access. *Journal of Science Education and Technology*, *21*(1), 133–147. https://doi.org/10.1007/s10956-011-9291-6

Quesenbery, W. (2001). What does usability mean: Looking beyond "ease of use." *In Annual Conference- Society for Technical Communication*, *48*, 432–436. https://doi.org/10.1.1.460.4258

Quiñones, D., Rusu, C., & Rusu, V. (2018). A methodology to develop usability/user experience heuristics. *Computer Standards & Interfaces*, *59*, 109–129. https://doi.org/10.1016/j.csi.2018.03.002

Raman, R., Achuthan, K., Nair, V. K., & Nedungadi, P. (2022). Virtual Laboratories- A historical review and bibliometric analysis of the past three decades. *Education and Information Technologies*, *27*(8), 11055–11087.

https://doi.org/10.1007/s10639-022-11058-9

Ramsden, P. (1979). Student learning and perceptions of the academic environment. *Higher Education*, *8*(4), 411–427. https://doi.org/10.1007/bf01680529

Rayment, S., Garrie, K., Kaur, I., McVicker, G., Storey, E., Winter, J., De Girolamo, L. A., Rimmer, C., Negus, D., Nelson, C., Thomas, J., Loughlin, M., & Dale, J. (2022). Investigating student engagement and making science real during a pandemic: Bioskills at home, a case study. *Education Sciences*, *12*(2), 106. https://doi.org/10.3390/educsci12020106

Reck, R., Sreenivas, R. S., & Loui, M. (2019). Evaluating the effectiveness of an affordable and portable laboratory kit for an introductory control systems course. *Advances in Engineering Education*, *Fall 2019*(1), 1–35. https://eric.ed.gov/?id=EJ1236922

Redelmeier, D. A., & Tibshirani, R. J. (1997). Interpretation and bias in case-crossover studies. *Journal of Clinical Epidemiology*, *50*(11), 1281–1287. https://doi.org/10.1016/s0895-4356(97)00196-0

Reece, A., & Butler, M. (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. https://doi.org/10.2505/4/jcst17_046_03_83

Reeves, S. M., & Crippen, K. J. (2021). Virtual laboratories in undergraduate science and engineering courses: A systematic review, 2009–2019. *Journal of Science Education and Technology*, *30*(1), 16–30. https://doi.org/10.1007/s10956-020-09866-0

Remenick, L., & Goralnik, L. (2019). Applying andragogy to an outdoor science education event. *The Journal of Continuing Higher Education*, *67*(1), 24–36. https://doi.org/10.1080/07377363.2019.1629804

Reuter, R. (2009). Online versus in the classroom: Student success in a hands-on lab class. *American Journal of Distance Education*, *23*(3), 151–162. https://doi.org/10.1080/08923640903080620

Ridgway, V. F. (1956). Dysfunctional consequences of performance measurements. *Administrative science quarterly*, *1*(2), 240-247.

Rosen, D. J., & Kelly, A. M. (2020). Epistemology, socialization, help seeking, and gender-based views in in-person and online, hands-on undergraduate physics laboratories. *Physical Review Physics Education Research*, *16*(2). https://doi.org/10.1103/physrevphyseducres.16.020116

Rowe, R. J., Koban, L., Davidoff, A. J., & Thompson, K. H. (2017). Efficacy of online laboratory science courses. *Journal of Formative Design in Learning*, *2*(1), 56–67. https://doi.org/10.1007/s41686-017-0014-0

Rozgonjuk, D., Kraav, T., Mikkor, K., Orav-Puurand, K., & Täht, K. (2020). Mathematics anxiety among STEM and social sciences students: the roles of mathematics self-efficacy, and deep and surface approach to learning. *International Journal of STEM Education*, *7*(1). https://doi.org/10.1186/s40594-020-00246-z

Saitta, E. K. H., Bowdon, M. A., & Geiger, C. L. (2011). Incorporating service-learning, technology, and research supportive teaching techniques into the university chemistry classroom. *Journal of Science Education and Technology*, *20*(6), 790–795. https://doi.org/10.1007/s10956-010-9273-0

Saldivar, B., Carter, C., Filipp, S., Gurka, M., & Davis, M. (2019). Patient satisfaction surveys in the Outpatient clinic setting: The variability of response with positively or negatively toned questions. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, *56*, 1–7. https://doi.org/10.1177/0046958019875554

Sari Ay, Ö., & Yilmaz, S. (2015). Effects of virtual experiments oriented science instruction on students' achievement and attitude. *Elementary Education Online*, *14*(2). https://doi.org/10.17051/io.2015.25820

Sasmito, A. P., & Sekarsari, P. (2022). Enhancing students' understanding and motivation during Covid-19 pandemic via development of virtual laboratory. *Journal of Turkish Science Education*, *19*(1), 180–193. https://doi.org/10.36681/tused.2022..117

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Morgan Kaufmann.

Schoonenboom, J., Johnson, R. B., & Froehlich, D. E. (2018). Combining multiple purposes of mixing within a mixed methods research design. *International Journal of Multiple Research Approaches*, *10*(1), 271–282. https://doi.org/10.29034/ijmra.v10n1a17

Schultz, M., Callahan, D. L., & Miltiadous, A. (2020). Development and use of kitchen chemistry home practical activities during unanticipated campus closures. *Journal of Chemical Education*, *97*(9), 2678–2684. https://doi.org/10.1021/acs.jchemed.0c00620

Seery, M. K. (2020). Establishing the laboratory as the place to learn how to do chemistry. *Journal of Chemical Education*, *97*(6), 1511–1514. https://doi.org/10.1021/acs.jchemed.9b00764

Senapati, S. (2022). Peeking into the sophisticated world of interactive science simulations. *Resonance*, *27*(11), 1971-1983. https://doi.org/10.1007/s12045-02201493-0

Shachar, M., & Neumann, Y. (2010). Twenty years of research on the academic performance differences between traditional and distance learning: Summative meta-analysis and trend examination. *MERLOT Journal of Online Learning and Teaching*, *6*(2). https://jolt.merlot.org/vol6no2/shachar_0610.htm

Shackel, B. (2009). Usability – Context, framework, definition, design and evaluation. *Interacting with Computers*, *21*(5-6), 339–346. https://doi.org/10.1016/j.intcom.2009.04.007

Sharma, S., & Ahluwalia, P. K. (2018). Can virtual labs become a new normal? A case study of Millikan's oil drop experiment. *European Journal of Physics*, *39*(6), 065804. https://doi.org/10.1088/1361-6404/aada39

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, *42*(1), 70–83. https://doi.org/10.1037/0012-1649.42.1.70

Sindelar, M. E., & Witkowski, R. J. (2021). Development and assessment of an online plant science laboratory course: Comparison to current in-person laboratory. *Natural Sciences Education*, *50*(2). https://doi.org/10.1002/nse2.20070

Sithole, A., Chiyaka, E. T., & Mabwe, K. (2022). Assessing students' approaches and perceptions to learning in physics experiments based on simulations and at-home lab kits. *Higher Education Studies*, *12*(1), 96–104. https://doi.org/10.5539/hes.v12n1p96

Son, J. Y., Narguizian, P., Beltz, D., & Desharnais, R. A. (2016). Comparing physical, virtual, and hybrid flipped labs for general education biology. *Online Learning*, *20*(3). https://doi.org/10.24059/olj.v20i3.687

Sotelo, D., Sotelo, C., Ramirez-Mendoza, R. A., López-Guajardo, E. A., Navarro-Duran, D., Niño-Juárez, E., & Vargas-Martinez, A. (2022). Lab-Tec@ Home: A cost-effective kit for online control engineering education. *Electronics*, *11*(6), 907. https://doi.org/10.3390/electronics11060907

Spencer, S. (2021). *The effect of distance education biology laboratories on student success at one California community college* (pp. 1–91) [Dissertation]. https://www.proquest.com/dissertations-theses/effect-distance-education-biology-laboratories-on/docview/2623033987/

Špernjak, A., & Šorgo, A. (2018). Differences in acquired knowledge and attitudes achieved with traditional, computer-supported and virtual laboratory biology laboratory exercises. *Journal of Biological Education*, *52*(2), 206–220. https://doi.org/10.1080/00219266.2017.1298532

Stefanovic, M., Tadic, D., Nestic, S., & Djordjevic, A. (2013). An assessment of distance learning laboratory objectives for control engineering education. *Computer Applications in Engineering Education*, *23*(2), 191–202. https://doi.org/10.1002/cae.21589

Stokes, J. A., & Silverthorn, D. U. (2021). Updating anatomy and physiology lab delivery: Shifting from a paper-based to an online lab instruction platform, just in time for a global pandemic. *Advances in Physiology Education*, *45*(2), 290–298. https://doi.org/10.1152/advan.00190.2020

Stuckey-Mickell, T., & Stuckey-Danner, B. (2007). Virtual labs in the online biology course: Student perceptions of effectiveness and usability. *MERLOT Journal of Online Learning and Teaching*, *3*(2), 105–111. https://jolt.merlot.org/vol3no2/stuckey.pdf

Sunderman, H. M., Headrick, J., & McCain, K. (2020). Addressing complex issues and crises in higher education with an adaptive leadership framework. *Change: The Magazine of Higher Learning*, *52*(6), 22–29. https://doi.org/10.1080/00091383.2020.1839322

Sung, S. H., Li, C., Huang, X., & Xie, C. (2021). Enhancing distance learning of science—Impacts of remote labs 2.0 on students' behavioural and cognitive engagement. *Journal of Computer Assisted Learning*, *37*(6), 1606–1621. https://doi.org/10.1111/jcal.12600

Swan, A. E., & O'Donnell, A. M. (2009). The contribution of a virtual biology laboratory to college students' learning. *Innovations in Education and Teaching International*, *46*(4), 405–419. https://doi.org/10.1080/14703290903301735

Sypsas, A., & Kalles, D. (2018). Virtual laboratories in biology, biotechnology and chemistry education. *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*, 1–6. https://doi.org/10.1145/3291533.3291560

Tatli, Z., & Ayas, A. (2013). Effect of a virtual chemistry laboratory on students' achievement. *Educational Technology & Society*, *16*(1), 1176–3647. https://www.jstor.org/stable/jeductechsoci.16.1.159

Taylor, D. C. M., & Hamdy, H. (2013). Adult learning theories: Implications for learning and teaching in medical education: AMEE Guide No. 83. *Medical Teacher*, *35*(11), e1561–e1572. https://doi.org/10.3109/0142159x.2013.828153

Tekbıyık, A., & Ercan, O. (2015). Effects of the physical laboratory versus the virtual laboratory in teaching simple electric circuits on conceptual achievement and attitudes towards the subject. *International Journal of Progressive Education*, *11*(3), 77-89. https://dergipark.org.tr/en/pub/ijpe/issue/26330/277435

Tho, S. W., & Yeung, Y. Y. (2018). An implementation of remote laboratory for secondary science education. *Journal of Computer Assisted Learning*, *34*(5), 629–640. https://doi.org/10.1111/jcal.12273

Tinto, V. (2017). Reflections on student persistence. *Student Success*, *8*(2), 1. https://doi.org/10.5204/ssj.v8i2.376

Toth, E. E., Morrow, B. L., & Ludvico, L. R. (2009). Designing blended inquiry learning in a laboratory context: A study of incorporating hands-on and virtual laboratories. *Innovative Higher Education*, *33*(5), 333–344. https://doi.org/10.1007/s10755-008-9087-7

Tsihouridis, C., Batsila, M., Vavougios, D., & Ioannidis, G. S. (2018). The timeless controversy between virtual and real laboratories in science education–And the winner is…. In M. E. Auer, D. Guralnick, & I. Simonics (Eds.), *Advances in Intelligent Systems and Computing, Vol 2 Teaching and Learning in a Digital World* (pp. 1539–1550.). Springer.

Unger, S., & Meiran, W. (2020). Student attitudes towards online education during the COVID-19 viral outbreak of 2020: Distance learning in a time of social distance. *International Journal of Technology in Education and Science*, *4*(4), 256–266. https://doi.org/10.46328/ijtes.v4i4.107

Uribe, M. del R., Magana, A. J., Bahk, J.-H., & Shakouri, A. (2016). Computational simulations as virtual laboratories for online engineering education: A case study in the field of thermoelectricity. *Computer Applications in Engineering Education*, *24*(3), 428–442. https://doi.org/10.1002/cae.21721

VanMeter-Adams, A., Frankenfeld, C. L., Bases, J., Espina, V., & Liotta, L. A. (2014). Students who demonstrate strong talent and interest in STEM are initially attracted to STEM through extracurricular experiences. *CBE—Life Sciences Education*, *13*(4), 687–697. https://doi.org/10.1187/cbe.13-11-0213

Vasiliadou, R. (2020). Virtual laboratories during coronavirus (COVID-19) pandemic. *Biochemistry and Molecular Biology Education*, *48*(5), 482–483. https://doi.org/10.1002/bmb.21407

Vennix, J., den Brok, P., & Taconis, R. (2018). Do outreach activities in secondary STEM education motivate students and improve their attitudes towards STEM? *International Journal of Science Education*, *40*(11), 1263–1283. https://doi.org/10.1080/09500693.2018.1473659

Vlachogianni, P., & Tselios, N. (2021). Perceived usability evaluation of educational technology using the system usability scale (SUS): A systematic review. *Journal of Research on Technology in Education*, 1–18. https://doi.org/10.1080/15391523.2020.1867938

Vlachogianni, P., & Tselios, N. (2022). The relationship between perceived usability, personality traits and learning gain in an e-learning context. *The International Journal of Information and Learning Technology*, *39*(1), 70–81. https://doi.org/10.1108/ijilt-08-2021-0116

Vogt, N. P., Cook, S. P., & Muise, A. S. (2013). A new resource for college distance education astronomy laboratory exercises. *American Journal of Distance Education*, *27*(3), 189–200. https://doi.org/10.1080/08923647.2013.795365

Weaver, L. D., Ely, K., Dickson, L., & DellAntonio, J. (2019). The changing role of the department chair in the shifting landscape of higher education. *International Journal of Higher Education*, *8*(4), 175. https://doi.org/10.5430/ijhe.v8n4p175

Wei, J., Treagust, D. F., Mocerino, M., Lucey, A. D., Zadnik, M. G., & Lindsay, E. D. (2019). Understanding interactions in face-to-face and remote undergraduate science laboratories:

A literature review. *Disciplinary and Interdisciplinary Science Education Research*, *1*(1). https://doi.org/10.1186/s43031-019-0015-8

West, J., & Veenstra, A. (2012). Cane toad or computer mouse? Real and computer-simulated laboratory exercises in physiology classes. *Australian Journal of Education*, *56*(1), 56–67. https://doi.org/10.1177/000494411205600105

Xu, D., Solanki, S., & Fink, J. (2021). College acceleration for all? Mapping racial gaps in advanced placement and dual enrollment participation. *American Educational Research Journal*, *58*(5), 954–992. https://doi.org/10.3102/0002831221991138

Young, J. R., Ortiz, N., & Young, J. L. (2016). STEMulating interest: A meta-analysis of the effects of out-of-school time on student STEM interest. *International Journal of Education in Mathematics, Science and Technology*, *5*(1), 62–74. https://doi.org/10.18404/ijemst.61149

Youngblood, J. P., Webb, E. A., Gin, L. E., van Leusen, P., Henry, J. R., VandenBrooks, J. M., & Brownell, S. E. (2022). Anatomical self-efficacy of undergraduate students improves during a fully online biology course with at-home dissections. *Advances in Physiology Education*, *46*(1), 125–139. https://doi.org/10.1152/advan.00139.2021

Youssef, F. F., Mohan, J., Davis, E. M., & John, M. (2021). Home-based physiology labs in the time of COVID-19 prove popular with medical students. *Advances in Physiology Education*, *45*(4), 709–714. https://doi.org/10.1152/advan.00048.2021

Zacharia, Z. C. (2007). Comparing and combining real and virtual experimentation: An effort to enhance students' conceptual understanding of electric circuits. *Journal of Computer Assisted Learning*, *23*(2), 120–132. https://doi.org/10.1111/j.1365-2729.2006.00215.x

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. https://doi.org/10.1002/tea.20260

Zhang, X., Al-Mekhled, D., & Choate, J. (2021). Are virtual physiology laboratories effective for student learning? A systematic review. *Advances in Physiology Education*, *45*(3), 467–480. https://doi.org/10.1152/advan.00016.2021

Zidny, R., Fadhilah, G. A., Melda, G. E., Sholihah, I. I., Widiastuti, N. L., Haerunnisa, N., & El Islami, R. A. Z. (2019). Simple and low-cost chemical experiment kits to observe the concept of gas laws. *Jurnal Penelitian Dan Pembelajaran IPA*, *5*(1), 16–25. https://doi.org/10.30870/jppi.v5i1.3917

Appendix A - Source Tables

Source Table A-I

Summary of Comparison Studies

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Lindsay & Good (2005) Are ranges not reported	3 formats TL RL VL (simulation)	Engineering Data Acquisition and Control	UG; M T4Y: I Public RU	Effectiveness Engagement Australia	N/A	Evaluation of Laboratory Reports (behavior and outcomes analysis); Post-test (open-ended questions) n=118 Students perceptions of their learning outcomes were nearly the same for all 3 modes. There were substantial differences in students' perceptions of lab objectives with actual outcomes; the presence of hardware seemed to focus students' attention to hardware, students that performed the simulation were not focused on hardware. TL: engaged by novelty of the experience RL: engaged by linking lab to theory VL: engaged by reinforcing lecture material Most students indicated that TL was preferred mode.	Students perceptions of what they achieved and what they were supposed to achieve differed by modality. Presence of lab equipment (or manipulating equipment from a distance) captured the attention of students (concerned with how to operate the equipment); students that did not operate "real" equipment focused more how to learn from the experiment. Students largely prefer TL.
Corter et al. (2007) Age ranges not reported	3 formats TL RL VL (simulation)	Engineering Engineering & Design II	UG; M T4Y Private Ivy	Effectiveness Engagement Ease of Use Self-efficacy United States	Standardized spacial ability test; scores on class assignments; grades; GPA; SAT scores; lab preferences questionnaire n=292 No significant difference in quiz and exam scores. No significant correlation between ability to learn from a particular lab type and previous grades, SAT scores, or spacial ability. Lab section/instructor did not affect scores. Students rated TL higher in educational effectiveness than RL or VL.	Lab preferences questionnaire n=208 Students found the group work to be difficult in the RL and VL. Students liked the convenience and flexibility of RL/VL, but preferred the hands-on experience. Students were unhappy with the instructions for the VL – were not able to complete the lab without seeking help from TAs.	Collaborative work patterns (coordination and communication) are significantly different in RL/VL versus TL. Instructors need to take this into account when designing RLs and VLs. It is important to incorporate a mechanism for student-student communication and instructor-student communication in the design of the RL/VL.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Klahr et al. (2007) No adult learners	TL vs VL (virtual car)	Engineering Design Mousetrap Car	7th and 8th grade K-12	Effectiveness Efficiency Engagement United States	Pre-post- Knowledge assessment questionnaire n=56 Investigated use of physical versus virtual manipulatives Participants were shown all of the parts that would be used to construct cars; given 20 minutes to construct/test as many cars as possible. Students using virtual materials were able to construct more cars than physical materials. No significant difference in learning gains.	N/A	Participants were able to learn as well with physical and virtual materials; therefore, factors other than physical/virtual material should be considered when designing laboratories (safety, accessibility, discipline, etc.) Type of material may impact learner attitude, long-term recall, and knowledge transfer – needs to be further investigated.
Stuckey-Mickell & Stuckey-Danner (2007) **Included Adult Learners**	TL vs. VL (simulation)	Biology Human Biology	UG; NM CC	Effectiveness Usability United States	Self-developed survey n=23; age range 18-55 89.6% perceived TL more effective than VL 60.8% perceived VL lab as effective Laboratory and lecture exam scores were nearly identical	Self-developed survey n=23; age range 18-55 Students perceived the VL as useful – but was not a "real" lab. Students preferred the hands-on experiences and collaborative environment (student-student and instructor-student interactions) of the TL.	Instructors should design VLs with frequent opportunities for student-student and student-instructor interaction. Learning outcomes and objectives for VLs should align with TLs and course outcomes.
Zacharia (2007) No adult learners Age range 20-22	TL vs B (TL + VL) and TL vs VL (virtual apparatus)	Physics Physics by Inquiry: Circuits	UG; NM T4Y; I Public RU	Effectiveness Cyprus	Pre-post conceptual tests TL (n=43); B (n=45) TL vs B Both groups experienced learning gains (pre-/post-differences): Mean TL, 22.23; Mean B, 32.56. TL vs VL Both groups experienced learning gains (pre-/post-differences): Mean TL, 9.69; Mean VL, 19.84.	N/A	VL appeared to better promote conceptual understanding of circuits. It may be manipulation rather than physicality that is the key aspect of laboratory instruction. The ultimate goal should be to take advantage of both methods of experimentation to maximize effectiveness.
Zacharia et al. (2008) Age ranges not reported	TL vs VL (virtual apparatus) and TL vs B (TL + VL)	Physics Physics by Inquiry: Heat and Temp.	UG; NM T4Y; I Public RU	Effectiveness Efficiency Cyprus	Pre-post conceptual tests TL (n=31); VL (n=31) Investigated the use of physical manipulatives versus virtual manipulatives. TL vs VL Both groups experienced learning gains (pre-/post-differences): Mean post-test score TL, 40.4; Mean post-test score VL, 48.9. TL vs B Both groups experienced learning gains (pre-/post-differences): Mean post-test score TL, 50.5; Mean post-test score B, 60.9.	N/A	The combination of physical manipulatives and virtual manipulatives improved post-test score more than use of physical manipulatives alone. Virtual manipulatives can be manipulated faster than physical. It is important to design laboratories that take advantage of the possibilities offered by both types of manipulatives.
Reuter (2009) **Included Adult Learners**	TL vs AHHO (lab kit)	Natural Sciences Soils: Sustainable Ecosystems	UG; NM T4Y Public Branch Campus	Effectiveness United States	Pre-/post assessment; assignment grades; final grades TL (n=50); AHHO (n=47) 42% improvement in pre-/post- assessment AHHO; 21% in TL. Student in AHHO format showed greater learning in lab-related knowledge and skills based on individual assessment questions. No significant difference in assignment scores or final grades.	N/A	Students self-selected format. There was a significant difference between mean ages for TL (25) and AHHO (41) formats. Mean age difference could be a confounding factor – adult learners tend to be more self-directed and motivated.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Koretsky et al. (2011) Age ranges not reported.	TL vs. VL (simulation)	Engineering Capstone	UG; M T4Y Public RU	Effectiveness Efficiency Ease of Use Error tolerance United States	Self-developed post-lab surveys (3 @ 2 TL, 1VL) n = 111 (2 years - 45;66) No significant difference in time spent for each of the (3) lab projects. Students perceived greater cognition, experimental design, and critical thinking in VL. Students perceived greater laboratory skills and content in TL.	3 Self-developed post-lab surveys (3 @ 2TL, 1VL) n = 111 (2 years - 45;66) Students indicated greater sense of ambiguity with the VL. However, the VL provided more opportunities for troubleshooting – perceived greater cognition, experimental design, and critical thinking	Instructional design of VL (and TL) critical to effectiveness
Martínez et al. (2011) Age ranges not reported	TL vs VL (simulation)	Physics Optics	UG; M T4Y; I Public RU	Effectiveness Engagement Ease of Use Spain	End-user Questionnaire TL (n=41); VL1 (typical sim, n=41); VL2 (hyper-realistic sim, n=41). Mean % correct answers: TL (66.95); VL1 (70.37); VL2 (76.47) Overall – users enjoyed the VL and had a positive user-experience.	N/A	Supports previous studies - learning in TL and VL1 nearly the same; however, there was a significant difference for VL2. VL (especially VL2) are an effective method to bring the lab to the students; whether to use TL, VL1, or VL2 will depend upon the nature of the lab content.
Olympiou & Zacharia (2011) No adult learners	TL vs VL (simulation) vs B (TL + VL)	Physics Intro. Physics	UG; M + NM T4Y; I Public RU	Effectiveness Cyprus	Pre-post-tests; Pre-/during/post-Conceptual tests TL (n=23), VL (n=23); B (n=24) Investigated use of physical and virtual manipulatives (and blend of both); All conditions measured in the same lab environment. All 3 conditions improved understanding of concepts. No significant difference in TL and VL; blended combination enhanced understanding more then TL and VL alone.	Students in all 3 conditions surpassed their conceptions of light in color; the blended combination had a greater impact on students transition from scientifically non-acceptable conceptions (SNAC) to scientifically acceptable conceptions.	The use of the blended combination of TL and VL seems to be the most conducive to learning through experimentation.
Pyatt & Sims (2011) No adult learners	TL vs. VL (simulation)	Chemistry General Chemistry	HS K-12	Effectiveness Attitude Usability Ease of Use United States	Self-developed survey (Virtual and Physical Experimentation Questionnaire, VPEQ) based on Science Laboratory Environment Inventory (SLEI), Computer Laboratory Environment Inventory (CLEI), and Attitudes towards Computers and Computer Courses (ACC). Experimental Crossover design n=184 (2 years; 96, 88) No significant differences in mean scores. Students perceived VL easier to complete than TL VL useful – helped students learn concepts. No preference in learning environment.	N/A	Virtual manipulation of objects can also be considered to be "Hands-on". The order the labs were performed (TL first vs VL first) did not matter – no significant difference in survey results with respect to order.
West (2012) No adult learners	TL vs VL (simulation)	Biology Physiology	UG; M T4YI I Public RU	Engagement Australia	Survey (n=166) Student preference: 48.8% TL; 3.6% VL; 47.6% both	Survey (n=166) Factors that students liked about VL: could repeat exercises, guaranteed results, could not make a mistake, convenient Factors students liked about TL: hands-on experiences, performing dissections/seeing results, working in groups	Students indicated that the TL provided a more memorable and stimulating learning experience. VL did not feel real; conveyed little sense of the real experiment. A combination of the VL and TL can be advantageous (students have memorable experience, provide access to wider range of experiments).

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Barbeau et.al. (2013) Age ranges not reported.	TL vs. VL (Virtual Slide Box and videos)	Biology Histology	UG; M T4Y; I Public RU	Effectiveness Satisfaction Canada	Scores on assignments, Final Grades, Course evaluations TL (n = 116); VL (n=120) No significant differences in quiz scores, exam scores or final grades. The method of course delivery did not have any significant impacts on satisfaction.	Course evaluations TL (n = 116); VL (n=120) Student comments reflected overall favorable perception of the online format. Students in TL expressed a desire for access to archived lectures and virtual slide box as a supplement to the TL.	The blended laboratory format may be ideal.
Brewer et al. (2013) Age ranges not reported	TL vs AHHO (kits)	Chemistry General Chemistry	UH; M T4Y; I Public RU	Effectiveness Engagement Canada	Scores on assignments (Laboratory Reports, Safety Quiz, Exam); Final Grades TL (n=28); AHHO (n=44) Grade distribution for AHHO was fairly consistent with TL; with the exception of those students receiving Fs (3x more AHHO students, did not complete the course).	N/A	It is possible to provide distance students with meaningful laboratory experiences.
Stefanovic et al. (2013) Age ranges nor reported	RL vs VL (simulation)	Engineering Control Techniques	UG; M T4Y; I Public RU	Effectiveness Serbia	Teachers (n=12) and developers (n=4) ranked the top 14 learning objections; Students evaluated each objective (n=231) Used an algorithm to evaluate which mode was more effective. Students preferred RL to VL; RL better fulfilled the laboratory objectives.	N/A	Both RL and VL are able to meet the laboratory goals and objectives; lab solution should be carefully selected based upon specific purpose and objectives that must be achieved.
Tatli & Ayas (2013) No adult learners	TL vs. VL (simulation)	Chemistry General Chemistry I	HS K-12; I	Effectiveness Self-efficacy Efficiency Turkey	Chemical Changes Unit Achievement Test; Laboratory Equipment Test TL (n=60); VL (n=30) No significant differences in achievement or in students' ability to recognize laboratory equipment	Teacher Interviews n=20 Unstructured laboratory observations Students in both the TL and VL indicated they were able to complete the experiments, felt confident in their results, and were able to associate the experiment with daily life.	VLs can be used to supplement and/or replace costly TLs.
Vogt et al. (2013) **Included adult learners** TL (<20) Blended (25 – 30)	TL vs B (VL + AHHO) VL (image archives, spectral data) + AHHO (home supplies)	Astronomy	UG; NM T4Y Public RU	Effectiveness United States	Exams TL (n=85); blended (n=21) In the blended lab, students completed 4 VLs and 4 AHHO lab activities; adult learners (25+, had children, taken more math); higher % withdrew from course due to lack of time, transferred to TL, and life events. Average exam score: TL: 24.7 ± 3.6 Blended: 28.8 ± 1.8		The blended format as effective as TL; increases accessibility for adult learners.
Chen et al. (2014) No adult learners	TL vs VL (simulation)	Physics Boyle's Law	HS K-12; I	Effectiveness Engagement Attitudes Taiwan	Pre-/post-conceptual tests (n=68) Physical vs Virtual manipulation of laboratory equipment No significant difference in conceptual knowledge. Students that engaged in lab with physical manipulation had better attitudes towards the laboratory experience.	Interviews Physical manipulation: better contribution to inquiry practices – how to improve the experiment, how to interpret and apply the results. Virtual manipulatives lead to "mindless" data collection.	Students had an unrealistic trust in technology – lacked critical view of results generated from computers. Virtual manipulation of objects was not as fun as (or create a lasting impression) labs that included the physical manipulation of objects.
Attardi & Rogers (2015) Age ranges not reported	TL vs. VL (simulation; 3D anatomical models) + videos	Biology Systemic Human Anatomy & Physiology	UG; M T4Y: I Public RU	Effectiveness Canada	Quiz/Exam scores, Grades Students self-selected TL (n=365); VL (n=40) No significant differences in quiz scores and exam scores (with the exception of exam 3; TL outperformed virtual)	N/A	Previous academic performance (in foundational courses) was a better predictor of student performance in anatomy than delivery format. Blended approach may be best.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Crandall et al. (2015) Age ranges not reported	TL vs VL (simulation)	Chemical Engineering Food Chemistry	UG; M T4Y Public RU	Effectiveness Engagement Ease of Use Efficiency United States	Knowledge Assessment Questions; 2 Surveys (1 after each lab experience). n=48 Crossover design No significant difference in performance on knowledge assessments. 52% found hands-on enjoyable <50% indicated VL could be useful in other science classes	2 Surveys (1 after each lab experience) n=48 Most students found the VL entertaining; several indicated it was juvenile or "cheesy". Most student enjoyed the TL; they liked working with others, could ask questions/get answers quickly	Simulations could be used as replacements for TL or as an effective introduction (pre-lab) prior to TL.
Oser & Fraser (2015) No adult learners	TL vs. VL (simulation)	Biology Genetics	HS K-12	Effectiveness Attitudes Engagement United States	Laboratory Assessment in Genetics questionnaire N=322 (n=79 females in TL, n=74 males in TL, n=92 females in VL, n=77 males in VL) No significant difference in perception of learning environment, attitude, or achievement (as a whole).	N/A	Take a deeper look into gender data. Males showed a slight advantage over females in the VL.
Sari Ay & Yilmaz (2015) No adult learners	TL vs VL (simulation)	Physics Electricity for Life	7th grade K-12; I	Effectiveness Engagement Attitude Turkey	Pre-/post- Science Laboratory Attitude Scale (SLAS) TL (n=33); VL (n=36) Scores increased for both conditions; students' attitudes and achievement were more positively affected by VL than TL.	N/A	VL is effective alternative to TL. Recommend the combination of both for best learning environment.
Tekbiyuk & Ercan (2015) No adult learners	TL vs. VL (simulation)	Physics Circuits	5th grade K-12; I	Effectiveness Attitude Turkey	Pre-/Post-test scores for Simple Electric Circuits Achievement Test and Simple Electric Circuit Attitude Scale TL, n=33; VL n=32 No significant differences in pre-/post-test scores for TL and VL. Neither the physical and virtual environments had an effect on attitudes towards subject.	N/A	Students in VL scored slightly higher in Recognizing Circuit Elements and Identifying Variables that Affect Brightness of Bulb; students in TL performed better on Forming Complete Circuits. A blended version – with access to both virtual simulation and actual breadboard – combines the strengths of both environments.
Attardi et.al. (2016) Age ranges not reported	TL vs. VL (simulation; 3D anatomical models) + videos	Biology Systemic Human Anatomy & Physiology	UG; M T4Y; I Public RU	Effectiveness Efficiency Engagement Canada	Survey VL (n=20); TL (n=310) Majority (82.6%) of students preferred the T Most important factors to student success: access to instructor/TA, access to cadavers/specimens, and in-person quizzes (students were stressed by the on-screen timer, inability to ask clarifying questions, and inability to go back to previous questions).	Interviews VL (n=20); TL (n=20) Perceived strength of VL: pace control (students could pause/review lectures at will). Perceived weakness of VL: lack of physicality and instructor-student communication.	Need to improve quality of student-instructor interactions and student-content interactions (3D models) in VL. Blended approach may be the best.
Son & Narguizian (2016) Age ranges not reported	TL vs VL (simulation) vs Blended (TL + VL)	Biology Animal Biology	UG; NM T4Y Public RU	Effectiveness Engagement Attitudes United States	Final course grades; Pre-/Post surveys (attitudes, knowledge) TL (n=186); VL (n=186); blended (n=376) No significant difference between TL and VL; blended significantly higher than TL and VL. There was an overall negative shift in student attitudes towards biology in TL and VL; positive shift in blended. There were no significant differences in knowledge of evolution across all three formats.	N/A	VLs are as effective for student learning as TLs; the combination of VL + TL had a positive impact on students' attitudes toward biology. The addition of carefully designed VLs can be effective (higher grades, positive impact on attitudes) while lowering costs of TL.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Meintzer et. al. (2017) Age ranges not reported	TL vs. RL	Chemistry Analytical Chemistry	UG; M I Public Polytech Institute	Effectiveness Satisfaction Canada	Laboratory scores (n=70); Student perception surveys (n=46); after completion of TL and after completion of RL. Crossover design Laboratory scores were nearly identical for both mode and order labs were performed. Students satisfaction was greater in TL (89%) than in RL (67%) – greatest dissatisfier was perceived lack of student-instructor interaction.	N/A	Increase student-instructor interaction. The appropriate combination of both formats (TL and RL) provides a better learning experience than either format alone.
Reece and Butler (2017) Age ranges not reported	TL vs. VL (simulation)	Biology Biology I	UG; NM + M T4Y Public RU	Effectiveness Motivation Self-efficacy United States	Pre-/post-surveys; pre-/post-tests; grades TL (n=139); VL (n=162) Pre-/post-test for content knowledge and final grades were nearly the same No significant difference in motivation or self-efficacy. Sadly, motivation to learn biology declined from beginning to end of term in both formats.	N/A	Students in VL preferred interacting with virtual TA (available on-demand, just-in-time as needed) to interacting with a real TA. Additional research is needed to gain a deeper understanding of how students collaborate with peers/interact with instructors in both TL and VL.
Ambusaidi et al (2018) No adult learners	TL vs VL (simulation)	Chemistry General Science	HS K-12; I	Effectiveness Engagement Attitudes Self-efficacy Oman	Achievement test, Attitudes towards Science, Attitudes toward Virtual Lab TL (n=35); VL (n=34) No significant difference in academic achievement or attitudes in science. Students had a mostly positive attitude toward virtual lab. Students enjoyed VL but did not see them as a means to develop new lab skills or to work collectively (with lab partners).	Focus Group (n=12) VL generated positive attitude toward learning – it was fun, felt like a game.	VL provide students with an opportunity to perform laboratories that may not otherwise be possible. Best to use a mixed system in which students have access to both physical and virtual equipment.
Javier & Lomuntad (2018) No adult learners	TL vs VL (simulation)	Physics Electro-magnetism	Grade 10 K-12; I	Effectiveness Manila	Pre-/post- knowledge test TL (n=44); VL (n=43) No significant difference in learning gains between conditions.	N/A	The VL provided an effective learning experience; however, students had difficulty applying concepts without traditional instruction. Inclusion of simulations prior to hands-on could provide additional depth to learning.
Miller et al. (2018) Age ranges not reported	TL vs VL (simulation)	Physical Sciences	UG; NM T4Y Public RU	Effectiveness Attitudes United States	Pre-/posttest scores of content knowledge; Attitudes toward Science survey; Preference of Laboratory Methodology survey Self-selected: TL (n=65); VL (n=31) No significant differences in content knowledge gained, attitudes, or preferences.	Open-ended questions (preference survey) Primary reason students self-selected VL: Convenience Primary reasons students self-selected TL: access to instructor (ask questions, get immediate feedback); learn better with hands-on.	VLs are an effective alternative to TLs. It can be beneficial to use both formats of instruction – Consider designing hands-on activities that can be completed at home.
Nolen & Koretsky (2018) Age ranges not reported	TL vs VL (simulation)	Engineering Capstone	UG; M T4Y Public RU	Effectiveness Engagement Motivation Interest United States	Engagement Surveys (after each lab project; 2 physical, 1 virtual); Pre-/post-course survey (motivation and interest) n=118 Students in VL perceived greater engagement, learning gains, interest in problem-solving, interest in engineering career, and task orientation.	N/A	VL was delivered last – order may have had an impact on student perception. Instructional design may have a greater impact on interest and engagement than mode of delivery.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Špernjak & Šorgo (2018) No adult learners	Triplet TL vs VL (simulation) vs B (computer-supported; performed in lab)	Biology Gas Exchange; Activity of Yeast; Heart Rate	6th grade – 9th grade K-12; I	Effectiveness Attitudes Engagement Slovenia	Knowledge pre-test; Lab performance n=552 (TL, n=200; VL, n=110; blended, n=242) No statistical differences in learning gains with regard to delivery, grade, gender, or school. Students preferred blended > TL > VL	N/A	Students achieve learning no matter how/where the laboratory was performed. Students are more engaged and have better attitudes towards learning when hands-on activities are supplemented with technology.
Gunawan et al. (2019) No adult learners	TL vs VL (simulation)	Physics General Physics	HS K-12; I	Effectiveness Indonesia	Student performance appraisal instrument n=58 No significant differences in formulating problems or summarizing (drawing conclusions). There were significant differences (VL >TL) for hypothesizing, practicing, and communicating.	N/A	VL incorporated guided inquiry; TL used conventional techniques. Guided inquiry (through VL) has a significant effect on students' science process skills; especially in formulating a hypothesis, practicing (practical skills and troubleshooting), and communicating (ability to make connections). Science process skills can be acquired through VL.
Kapici et al. (2019) No adult learners	TL vs VL (simulation)	Physics Circuits	7th grade K-12; I	Effectiveness Turkey	Pre-/Post- Conceptual Knowledge Test; Pre-/post-inquiry skills test Students performed 3 Labs; H = TL; V = VL HHH (n=33); VVV (n=34); VHV (n=39); HVH (n=37) Gains in conceptual knowledge (VHV > HVH > HHH > VVV) Gains in inquiry skills (VHV > HVH > VVV > HHH)	N/A	The blended learning environments (VHV and HVH) demonstrated larger gains in conceptual and inquiry skills; use of both hands-on and virtual more effective than either format alone. Incorporating both types of labs (based upon availability of lab materials) provides effective learning environment for students.
Moosvi et al. (2019) Age ranges not reported	TL vs AHHO (home supplies)	Physics Intro Physics	UG; M T4Y; I Public RU	Effectiveness Engagement Ease of Use Canada	Lab Project Scores TL (n=57); AHHO (n=176) No significant difference in overall project scores; however, there were significant differences in 2 grading categories (clearly staring research questions; estimating uncertainty) and not satisfactory in either format.	Focus Group n=12 (AHHO) AHHO labs were easy but a "nuisance" to perform (difficulties finding suitable supplies); would have preferred performing labs on campus (easy access to supplies/instructor) Projects felt a bit childish at times; "real science" requires sophisticated equipment and yields precise data. Several enjoyed the freedom and flexibility of designing/performing the labs at home.	The AHHO lab is a viable alternative to the TL; average project scores were nearly identical in both formats. Access to better equipment may/may not alter perception of "real lab" – "real lab" may also be tied to development of community of practice (collaboration with others).
Reck et al. (2019) Age ranges not reported	TL vs HO (kit, used in laboratory)	Engineering Control Systems	UG; M T4Y Public RU	Effectiveness Efficiency Engagement Ease of Use United States	Exam scores, Laboratory report scores, Concept Inventory Test, Self-developed survey TL, n=37 (2 years; 26, 11) HO, n=27(2 years; 27, 10) No significant differences in exam scores, lab report scores, or concept inventory test. No significant differences in student experiences with equipment, time spent on lab, or perception of learning, with the exception of 1 item (TL > understanding of control systems/components).	Laboratory observations, Reflection, Open-ended questions on satisfaction survey, focus groups Students in both groups enjoyed the laboratory experience overall; were mostly able to complete the laboratory in the allotted timeframe (did not feel stressed about ability to complete). Some students expressed that they were just following directions – did not really understand what/why they were performing each step of the procedure.	Students were able to achieve the same learning objectives using the kit as students using the traditional equipment. Equipment (traditional vs kit) did not appear to have a significant impact on students' experiences. Kits are an acceptable alternative to traditional lab.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Cossovich et al. (2020) No adult learners	TL vs AHHO (kit)	Physics Working with Electrons	UG; NM T4Y; I Private	Effectiveness Engagement Self-efficacy Motivation US/China	Assignments; Questionnaires n=17 (from 11 countries) No significant differences in assignments. Kits were effective for learning; increased motivation and self-efficacy. Kits were fun and engaging to use.	Student feedback n=17 Kits helped students to understand the lecture concepts; hands-on learning gave students practical, hands-on learning experiences, able to discover details that simulations could not reveal; actual experience soldering, building circuits, etc.	AHHO labs are an effective and engaging laboratory experience; support students' research skills; and help motivate students to continue their efforts. The lack of integration between different communication tools was less than ideal – but did not hinder learning.
Kelley (2020) No adult learners	3 formats TL vs AHHO (kit) vs AHHO (home supplies)	Chemistry General Chemistry	HS K-12	Effectiveness Engagement Ease of Use Error tolerance Efficiency Self-efficacy United States	Lab reports, Survey Labs 18, 19, and 20 completed during COVID-19 transition (n=59) 54/59 completed lab 18 (TL) as written (4 requested alternative); 40/59 completed lab 19 AHHO (kit) as written (18 completed alternative lab); 10/59 completed Lab 20 AHHO (home supplies) as written (47 requested alternative lab) Access to materials was a key issue facing students (students had to pick up lab kits from school/purchase home supplies) Students that completed hands-on labs earned grades comparable to previous terms; students that completed alternative labs (videos, given data set) earned significantly lower grades (10-30%, depending on assignment).	Student feedback Positives AHHO: It was fun to do laboratories; doing stuff with my hands helps me learn; you can set your own schedule; breaks up the monotony of so much screen time Negatives AHHO: Cannot ask instructors/peers questions in the moment; cannot model what others are doing; cannot redo if messed up or spill; feels childish; hard to keep younger siblings out of supplies	Students preferred (and were more successful) running experiments; no students reported a preference for videos/alternative assignments, but some expressed that they could complete them more quickly. Hands-on activities seemed to boost morale and interest in the course. Obtaining supplies was a significant challenge for many students.
Rosen & Kelley (2020) Age ranges not reported	TL vs AHHO (kit)	Physics Calculus-based Intro. Physics	UG; M T4Y Public RU	Effectiveness Engagement Self-efficacy United States	Self-developed survey TL (n=747); AHHO (n=251) No significant difference in epistemological beliefs (effectiveness) or help-seeking (self-efficacy); there was a significant difference in socialization (engagement).	N/A	Results consistent with previous research; effectiveness strongly correlates with the way labs are taught, not the location. The main differences were in the value of socialization (or engagement) with peers and the instructor. Students that self-selected the TL placed a higher value on access to instructor/lab partners/TAs for immediate feedback. Social interactions and construction of knowledge may assume different forms when considering the multitude of electronic options for online learners. Design should be based upon specific learners.
Schultz et al. (2020) **Included adult learners** (3.6%)	TL vs VL (videos) vs AHHO (home supplies)	Chemistry General Chemistry	UG; M + NM T4Y; I Public RU	Effectiveness Engagement Efficiency Self-efficacy Australia	Lab reports; proportion submitted; posts/student Significant difference in average lab score TL (n=188) and VL (n=768); average lab score TL consistent with average lab score from previous 2 academic years. No significant difference in proportion of submissions (TL, VL, AHHO). No significant difference in mean lab scores for Lab 1 (TL and VL; n=768) and AHHO (n=935; 895). Students were more engaged in VL/AHHO labs; 0.51 posts per student as compared to TL, 0.13	Student reflections n=13 VL/AHHO activities were engaging and interesting; some expressed that the VL/AHHO labs took more time to complete than TL; seberal indicated working alone at home was difficult and stressful.	Difference in average lab report scores (TL > VL) may be attributed to interactions/support from TAs in lab. Guided support is needed for labs.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
Achuthan et al. (2021) Age ranges not reported	TL vs RL	Engineering Mechanics of Solids	UG; M T4Y; I Private	Effectiveness Ease of Use Efficiency India	Pre-/Post- Conceptual Understanding Questionnaire (CUQ); Effectiveness of Use of Learning Platform survey TL (n=50); RL (n-50) Both groups showed significant difference in Pre-/Post CUQ scores; RL users post- was significantly greater than TL. When compared to TL, RL users conducted the experiment 3x more frequently; completed assignments in 30% less time; had 200+% increase in pre-/post-CUQ scores (TL, 133% increase).	N/A	Significant differences in design may contributed to differences in Conceptual Understanding and Effectiveness. TL: Students worked in groups; set-up their own samples; limited by lab time and availability of Universal Testing Machine (UTM) RL: Students worked individually; samples were pre-set; allowed for more focused time on UTM and for replication.
Johnson & Barr (2021) Age ranges not reported	TL vs VL (videos; data-sets)	Engineering Mechanic Engineering Practice	UG; M T4Y Public RU	Effectiveness Engagement Efficiency United States	Assignment scores (compared to pre-pandemic terms) Averages for 3/7 VL assignments were lower by 6-16%. No significant changes in overall grade distribution. No significant difference in demonstration of conceptual knowledge.	Student Reflections n=400 Emergency transition from TL to VL due to pandemic Student perceived VL to be less than TL; videos are not a substitute for hands-on engagement and interaction with hardware; felt less engaged due to loss of social interactions with peers/instructors Students felt they were forced to develop self-directed learning skills; considered this as a positive, "engineers have to learn to overcome/adapt"; helped them prepare for future (better communication skills, time management, accountability, professionalism)	Videos (of lab being performed) does not provide an adequate learning experience for students; synchronous live-stream, AHHO, or RL are better options. Design considerations should not only focus on replicating hands-on activities in the remote environment, but also social engagement (peer-to-peer, student-instructor interactions).
Sindelar & Witkowski (2021) Age ranges not reported	TL vs AHHO (plants around the house)	Biology Plant Science	UG; NM T4Y Public RU	Effectiveness United States	Final Exam, Self-developed survey (n=55, AHHO; n=104, TL) No significant difference in scores. Students in the TL indicated a better understanding of the implications of the lab than VL.	Self-developed survey Majority of students indicated that the lab was fun, appreciated the hands-on aspect. Strength of the VL: convenience. Weakness of the VL: limited interaction with the instructor	Students prefer to be passive learners, but enjoy the hands-on aspects of both labs. Students value time with the instructor (in-class or synchronous online meeting) to ask questions, get clarification.
Stokes & Silverthorn (2021) Age ranges not reported	TL vs VL (simulation + virtual microscope + data pack)	Biology Anatomy & Physiology	UG; M T4Y Private	Effectiveness Engagement Efficiency Ease of Use United States	Survey; Lab Report grades TL (Lab 1, 2, 3, n=14, pre-COVID closure); VL (Lab 8,9, n=19, post-COVID closure) Students performed better on TL (worked in groups) assessments than VL (worked individually); students in VL that attended online session with instructor performed better than when completed lab on own). Significant difference in student perception survey scores (TL >> VL).	Student feedback Students overwhelmingly enjoyed the hands-on experiences in the TL; found TL easier to perform. Students found the VL complicated, confusing, took a long time; but, did help them understand/retain information.	Students performed better in TL than VL; however, students that interacted with the instructor performed better on the VL than those that did not. Design of VL should include interaction with instructors/students; perform labs synchronously together and/or engage in peer collaboration (in web-conferencing applications like Zoom, Google Meet, etc); social interactions are important.
Chen (2022) Age ranges not reported	TL vs VL (videos)	Bio-technology Lab techniques	UG; M T4Y Public RU	Effectiveness Self-efficacy United States	Lab notebook grades, lab reports, final exam grades, final grades TL (n=9); VL (n=9) No significant difference in final course grades or assignment grades. Students in VL gained self-confidence in the lab procedures- however notebook entries showed more misconceptions in the material when compared to TL.	N/A	Synchronous sessions should be used to supplement the lab videos – discuss theoretical underpinnings, troubleshooting techniques, and pitfalls to avoid. Embed interactive activities to support learning.

Authors	Modes of Delivery*	Discipline Course or Topic	Level** Type***	Constructs Country	Quantitative Results/ Empirical Evidence	Qualitative Results Students' Perception	Implications
DeChenne-Peters et al. (2022) Age ranges not reported	TL vs VL (home supplies + videos + datasets) vs B (TL + VL)	Biology Intro. Cellular and Molecular Biology Course-based UG research (CURE)	UG; M + NM T4Y Public RU	Effectiveness Self-efficacy	Pre-/post- Content knowledge questions; Analysis of Scientific Data questions; Persistence in the Sciences (PITS) survey TL (n=254); VL (n=55); blended (n=147) There were differences in learning gains between modalities: V L> blended > TL No significant difference in modalities in Attitudinal data (self-efficacy, science community values, project ownership, networking).	N/A	Results support that CUREs can be expanded into the blended and online environment. Motivational differences may have contributed to differences between modalities (mix of STEM and non-STEM majors).
Finne et al. (2022) Age ranges not reported	TL vs VL (dataset + videos)	Chemistry Pharm. Analytical Chemistry	UG; M T4Y; I Public RU	Effectiveness Engagement Denmark	N/A	Interviews (n=12) TL (prior to closure); lab reports returned in the laboratory (with extensive written and oral feedback), worked in groups VL (after closure); provided with dataset and explanatory videos; worked on lab reports in groups and submitted work via email; written feedback only Students expressed a feeling of missing out; missed collecting own data and easy access to instructor (for help and feedback); interaction with instructor is scaffolding for student learning.	Students' perceived that their understanding decreased in VL (when compared to TL). Contact and dialogue with instructor plays a significant role in scaffolding students' understanding and scientific judgement. Difficult to connect to make connections between theoretical concepts and data when student does not collect data; need access to lab.
Gnesdilow & Putambekar (2022) No adult learners	TL vs VL (simulation)	Physics Inclined planes	MS K-12	Effectiveness Efficiency United States	Pre-/Posttest scores; post-lab explanations TL then VL, n=60; VL then TL, n=50 Crossover study Both conditions conducted in the classroom No significant difference in posttest scores. Performing the TL was less effective in helping students write explanations (no matter the order).	N/A	VL alone can support student learning; TL then VL may be most beneficial for MS students.
Rayment et al. (2022) Age ranges not reported	TL vs AHHO (lab kit)	Biology Bioscience Lab Skills	UG; M T4Y; I Public RU	Effectiveness Engagement Ease of Use Self-efficacy United Kingdom	Course assignments, final grades TL (previous terms) AHHO (n=450) 98.2% successfully completed practical skills assessment; pass rate in previous terms (TL) were 95.4% and 99.4%.	Student reflections Students found the kit to be a challenging and engaging way to develop/improve practical skills. Activities positively impacted performance on lab assessments.	Students enjoyed the hands-on activities (using the kits); however, only a small percentage of students engaged in the online community building activities (discussion boards, online microscopy activity, celebration event). To increase community with home lab kits requires careful design.
Sithole et al. (2022) Age ranges not reported	VL (simulation) vs AHHO (kit) Vs B (VL/AHHO)	Physics General Physics	UG; M T4Y Public RU	Effectiveness Engagement United States	Modified-version of the Learn Questionnaire (MVLQ) N=108 VL promotes a deep approach to learning; however, ranked lowest in interest and relevance; peer-support, and alignment. May be due to lack of hands-on experience. Students perception of learning, alignment, and interest was ranked highest in AHHO and Combination (VL + AHHO).	N/A	Learning occurred in all three modes of lab delivery. VL alone does not generate the same level of interest/engagement as AHHO or Combination (AHHO + VL). Hands-on activities seem to better support interest/engagement in laboratory activities.

Note. *TL = Traditional Laboratory (hands-on, in-person, on-campus), VL = Virtual Laboratory (as indicated), RL = Remote Laboratory, AHHO = At-Home Hands-On (as indicated), HO = Hands-On; B = blended (as indicated)

[**] UG = Undergraduate, HS = High School, MS = Middle School; NM = Non-majors course, M = Majors course

[***] CC = Community College; T4Y = Traditional 4-year, K-12 (Kindergarten through 12th grade); I = international (not conducted in the United States of America); RU = Research University

Source Table A-II

Summary of Non-Comparison Studies: Virtual, Remote, or Hands-On Laboratory Experiences

Authors	Mode of Delivery*	Discipline Course or Topic	Level** Type***	Constructs	Results/Implications
Mawn et al. (2011) **Adult Learners (avg. age = 36)**	AHHO (field work) (kit) (kit)	Gen. Ed. Cemetery Activity LEGO racer activity Solar Charger activity	UG; NM 4Y Public Online + F2F Commuter campus	Effectiveness Engagement United States	Cemetery (n=17); LEGO (n=25); Solar Charger(n=22) Students utilized the processes of science (predictions, observations, data collection, analysis, and communication of findings); the process of “doing science” can be accomplished outside of the laboratory. Students completed assignments as directed, did not explore beyond the instructions in the manual; instructors may want to provide students with more discovery or inquiry-based activities; open-ended activities may foster student-directed questioning/exploring. Students submitted written reports to instructor; instructors should provide opportunities for communication of findings to their peers. Instructors should focus not only on content, but also process skills development; beneficial for students to refine their scientific process skills (provide opportunity for students to explore their own testable questions).
Lowe et al. (2013) No adult learners	RL	Physics Inclined Planes; Shake tables; hydroelectric apparatus	HS K-12; I	Effectiveness Engagement Ease of Use Australia	Science Laboratory Environment Inventory (SLEI): After completing the RL, 94% of the students (n=112) indicated they were able to lean from the RL; 67% indicated that they developed skills; 61% indicated that the lab was relevant. Students perceive RL as valid practical lab experiences that yield readily obtainable, reliable, authentic, and reproducible data.; positive learning outcomes and skills acquisition. RL are effective laboratory experience.
Erasmus et al. (2014) Age ranges not reported	RL	Biochemistry Atomic Absorption Spectro-photometer (AAS) Lab	UH; M T4Y; I Public RU	Effectiveness Engagement Ease of Use Canada	Students (n=62) completed a survey Working with the AAS remotely was easy, enjoyable, liked working with real samples, deemed the experience valuable (although some thought it was more in alignment with analytical chemistry than biochemistry), provided a positive learning experience, and was engaging. Some indicated they would have preferred to have the AAS in front of them rather than control remotely. RLs are an effective addition to the laboratory experience; can provide students the opportunity to work with specialized equipment they might not otherwise have access to.
Uribe et al. (2016) **Includes adult learners (70% 26-40)**	VL computational simulation	Engineering Thermo-electricity: From Atom to systems	UG + G; M T4Y Public RU	Effectiveness Usefulness Ease of Use United States	Students (n=29) completed pre-/post-test, and perception survey. Overall, students’ average performance significantly increased; however, post-test scores indicated that the students did not achieve mastery-level for the learning objective (moved from low (pre) to acceptable range (post); may indicate that instructors need more training to prepare/guide students through the simulations. On average, the students had positive experience with the simulations (especially in areas of usefulness, ease of use, and future intention to use simulations). Students seem to view the simulations as an easy way to see the effect of changing parameters, testing multiple scenarios, and visualize the equations; valuable precursor/link to real experimentation.

Authors	Mode of Delivery*	Discipline Course or Topic	Level** Type***	Constructs	Results/Implications
Childers & Jones (2017) No adult learners	RL	Electron Microscopy	HS K-12	Effectiveness Self-efficacy Motivation United States	Presence Survey; Motivation Survey; Science Identity Survey n=72 Science learning drive (self-efficacy and motivation), environmental presence (perception of physical interactions with equipment/others, sense of ownership), and inner realism presence (realness of lab, sense self-satisfaction) had largest impact on students' experience. Use of virtual tools can help students' understanding of the interdependence of science, engineering, and technology; enables students to work with research-grade instruments.
Orozco (2017) Age ranges not reported	AHHO (kit)	Forensic Science Introduction to Forensic Science	UG; M T4Y Public RU	Effectiveness Efficiency Ease of Use Self-efficacy United States	The laboratory kit included 13 activities (all materials provided in kit, including instructions for how to use kit and procedures for each activity). Students successfully completed all of the assigned activities (achieved learning objectives; within expected timeframes; kits were easy to use; students were confident in their abilities); however, the kit alone may not provide enough instructional materials for students to fully understand the concepts. Future lab design, kits should be accompanied by recorded lectures and demonstrations.
Sharma & Ahluwalia (2018) Age ranges not reported	VL (simulation)	Physics Millikan's Oil Drop Experiment	UG; M T4Y; I Public RU	Effectiveness Engagement Self-efficacy India	Students (n=41) completed a pre-test (taken after the theory and procedure were explained in the classroom) and a post-test (taken after the completion of the virtual lab experiment); the normalized gain was calculated for each question to determine improvement of conceptual understanding; students also completed a perception survey. Students demonstrated a gain in conceptual understanding for 6/8 questions; students perceived that performing the VL enhanced their conceptual understanding, motivation, and self-learning attitudes. However, student indicated that they did not see the VL as a replacement for the TL; they wanted to touch/feel real apparatus. Suggests the VL will make a good pre-lab preparatory assignment.
Tho & Yeung (2018) No adult learners	RL	Integrated Science Physics (4) Biology (4)	Grades 7-9 K-12; I	Effectiveness Engagement Ease of Use Motivation China	Students (n=32) completed pre-/post-surveys and an interview. The RL system was easy to use and enriched the learning experience (authentic scientific investigation activities with real equipment); there was an increase in interest and motivation; students actively participated in science experiments; internet and language issues (instructions were in English) were largest barriers. The RL system can be used as an effective and engaging tool to facilitate science learning. The tool can bring real science experiments to students in rural or developing regions.
Zidney et al. (2019) No adult learners	AHHO (kit)	Chemistry Gas Laws	HS K-12; I	Effectiveness Engagement Ease of Use Indonesia	Students (n=27) from private HS is rural area used a kit to explore Gay-Lussac's and Charles' Laws. Due to limited access to chemistry labs, students were excited to perform hands-on labs with the kits. The kits were easy to use and stimulated interest in the topic. Both students and teachers rated the kits high in educational (support learning, facilitates understanding; supports development of science process skills), technical (ease of use, safety), and aesthetic aspects (attractive/interesting). Kits are an effective way to bring real science to students in rural regions (with limited access to laboratory facilities).
Al-Soufi et al. (2020) Age ranges not reported	AHHO (home supplies; mobile apps)	Chemistry Applied Thermo-dynamics	UG; M T4Y; I Public RU	Effectiveness Spain	Students (n=23) built a double-beam photometer using their smartphones and household materials; prepared a series of dye solutions, and measured the absorbance-concentration calibration curve. Quality of the assignments and final lab report confirmed that most of the students achieve the learning objectives at level comparable to those completed in conventional lab. Smartphones can be used at home as a viable replacement for some sophisticated pieces of equipment.

Authors	Mode of Delivery*	Discipline Course or Topic	Level** Type***	Constructs	Results/Implications
Andrews et al. (2020) Age ranges not reported	AHHO (home supplies)	Chemistry Gen Chem II	UG; M T4Y Private	Effectiveness Efficiency United States	The home experiments utilized home supplies; students built equipment as needed; researched alternative methods for colorimetric analysis. Did not expect a high degree of precision from the at-home experiments; however, students were able to obtain accurate answers. With the high degree of "discovery", the students indicated that balancing time investment for lab with other coursework challenging; the discovery process facilitated a positive sense of camaraderie among the class (as they worked together to develop "home" measurements). Majors-level chemistry labs can be accomplished at-home without specialized equipment.
Easdon (2020) Age ranges not reported	AHHO (home Supplies)	Chemistry Chemistry of Food, Flavors, Fragrances	UG; M T4Y Private	Effectiveness Engagement United States	Students performed their choice of any two projects (extraction and oxidation of aldehyde, diffusion of candy dyes in different solutions, temperature optimum for enzymatic sucrose hydrolysis) + one developed a lab of their own. Students had a strong interest in developing their own at-home lab; there were difficulties in performing labs at home (lack of counter space, difficulty getting supplies (there was a shortage of yeast during pandemic); troubleshooting/problem-solving skills developed in the at-home labs were valuable.
Hsu & Rowland-Goldsmith (2020) Age ranges not reported	VL (datasets)	Biology Molecular Biology	UG; M T4Y Private	Effectiveness Engagement United States	Students attended synchronous lab sessions; instructor discussed lab techniques/purpose/rationale and were provided with dataset to analyze. Students identified the inability to complete hands-on activities as a major barrier to learning; it was difficult to connect the dataset for analysis to the laboratory task (as they did not perform the task). Students (76.4%) indicated that they preferred exploratory labs over cookbook labs (with pre-determined outcomes); exploratory labs "feel more realistic, get opportunity to critically think through a problem/troubleshoot, more interesting/fun/exciting. Regular interaction with the instructor/students was highly beneficial; instructors should design online lab classes with collaborative sessions.
Nguyen & Keuseman (2020) **Includes adult learners (≤30%)**	AHHO (home supplies)	Chemistry Chemistry in the Kitchen	UG; NM 4Y (T + online) Private	Effectiveness Engagement Self-efficacy United States	3 terms: n=14 (75% T), 12 (8% T), 15 (100% T); Students performed open-inquiry experiments (mixtures, colloids, concentrations, energy, chemical reactions, kinetics) + final project. Some students initially reported anxiety about performing labs at home without instructor guidance; interest and excitement grew with each lab – students were more confident with each lab. Students initially struggled with lab reports (how to convey what they learned in a scientific manner); quality of analysis increased as term progressed. Students reported positive learning experiences and a personal interest in the home-based lab experiments.
Barthet (2021) Age ranges not reported	AHHO (kit)	Molecular Biology Restriction digestion and ligation	UG; M T4Y Public RU	Effectiveness United States	At-home students were paired with in-class students (pairs streamed during lab activities); At-home students were provided all of the necessary equipment, enzymes, chemical necessary to complete the lab. Each pair presented their findings at the end of the lab. The at-home laboratory exercises can be coupled with more complex in-class component to further enhance learning of various molecular techniques.
Doughan & Shahmuradyan (2021) Age ranges not reported	AHHO (home supplies and smartphone)	Chemistry Analytical Chemistry	UG; M T4Y; I Public RU	Effectiveness Self-efficacy Canada	Students (n=7) utilized home supplies to extract starch from a banana and prepare iodine-starch solutions of varying concentrations; they utilized a smartphone to develop a calibration curve for the brightness of the solution (at home calorimetric analysis that typically requires specialized equipment). Students developed troubleshooting skills as the at-home experiments did not always go as planned, practical skills, and learned the importance of good technique.

Authors	Mode of Delivery*	Discipline Course or Topic	Level** Type***	Constructs	Results/Implications
Gutruf et al. (2021) Age ranges not reported **Andragogy**	AHHO (kit)	Engineering Biomedical Engineering	UG: M T4Y Public RU	Effectiveness Engagement United States	Students were able to engage in the full design process with the kits at home; AHHO provided students valuable insights into device design, testing and evaluating the strengths and limitations of designs. Students did find the Live demonstrations of fabrication of student designs interesting or engaging; took a lot of time. Students created a contract for team activities for effective virtual collaboration; aligned with professional work environment.
Howard & Meier (2021) Age ranges not reported	AHHO (kit)	Physics Intro. Physics	UG; M + NM T4Y Private	Effectiveness Engagement United States	Comparison of on-site and AHHO lab exercises for technical performance, % of students with >90% accuracy was equal to (2 labs) or significantly less than (2 labs) the on-site lab activities; clear advantage of equipment available in on-site labs; there is a need for portable, low-cost equipment for AHHO labs. However, lower accuracy may have motivated students with kit to repeat experiment; added benefit, students learned to troubleshoot experimental set ups and data collection. Students perceived benefits from working with AHHO labs; indicated they believed AHHO were same or better than on-site labs (experienced in other courses); enjoyed hands-on and convenience of ah-home.
Klein et al. (2021) Age ranges not reported	VL (simulations + Videos of real labs + data set)	Physics	UG + G T4Y; I Public RU	Effectiveness Germany Austria Croatia	N=578 (5 universities) Students were able to develop experimental skills and that labs reinforced lecture content in all three modes. Own real data > simulated data > given data set High correlation with self-organization (time management, self-regulated learning skills, etc.), communication, and attitudes with learning achievement. First-year students (younger, less-experience) were not as successful as more experienced learners. Gathering own data (even when from a video of a lab being performed or watching instructor perform lab over Zoom, Skype, etc.) both reinforced content and acquisition of experimental skills. Learning may be tied to whether or not the student gathered own data rather than how the experiment was performed (hands-on, virtual, remote, etc.).
Youssef et al. (2021) Age ranges not reported; 2nd year medical students	AHHO (home supplies and mobile apps)	Biology Physiology	G Medical School; I	Effectiveness Efficiency Engagement Ease of Use Attitudes Trinidad & Tobago	N=235 No significant difference in academic performance (compared to previous year). Participants overwhelmingly were satisfied with the AHHO experience; enjoyed the AHHO; some concerned about lack of access to specialized equipment; some indicated that the AHHO labs were time-consuming. Students proposed a combination of AHHO and TL moving forward.
Avci (2022) Age ranges not reported	VL (simulation)	Chemistry Acid-base	UG; NM T4Y; I Public RU	Effectiveness Engagement Turkey	Students (n=36) showed proficiency in learning outcomes (writing hypotheses, testing the accuracy of hypotheses, and reporting results) While the students found the simulations interesting and helpful, students had difficulty examining scientific terminology in the simulation. Instructional material to support concepts in VL should be provided as part of the laboratory learning materials.
Davidson et al. (2022) Age ranges not reported	AHHO (kit) + VL (simulation)	Forensic Science Basic Forensic Science	UG; M T4Y; I Public RU	Effectiveness Engagement Scotland	Survey data (n=32 indicated that their favorite part of the kits was the physicality of the experiments and ability to do the experiments at home; their favorite part of the Laboratory Examination Exercise (LEE) was being able to perform realistic casework and the relaxed supportive environment; held mock court on Zoom (in smaller groups than in the traditional class). Kits (along with the other components) demonstrated that students can develop laboratory skills outside of the laboratory; mock court (Zoom) provided flexibility on timings while still allowing students the opportunity to practice how to present evidence in court; however, for some activities, in-person on-campus teaching is critical. A combination of in-campus and home activities can be an effective way to deliver the class.

Authors	Mode of Delivery*	Discipline	Level**	Constructs	Results/Implications
		Course or Topic	Type***		
Desa et al. (2022) Age ranges not reported.	VL (simulation)	Biochemistry Principles of Biochemistry	UG; M T4Y; I Public RU	Effectiveness Self-efficacy Malaysia	Students (n=24) completed a simulation to explore how metabolism and endurance are related; create a meal that would supply them with enough energy to complete a fitness test; delivered online presentation outlining their data collection, results, conclusions. Student reflections indicated that the "discovery" labs changed the way they learned, made them more confident in their ability to learn lab concepts on their own; Home discovery fosters problem-solving skills, data processing, and interpretation skills; opens a wealth of opportunities to design relevant and adaptive real-world learning experiences.
Honig et al. (2022) Age ranges not reported	AHHO (kits)	Chemical Engineering Heat Exchanger Practical	UG; M T4Y; I Public RU	Effectiveness Engagement Australia	Students (n=41) worked in teams; those in group that utilized the kit were designated as "Process Engineers" (n=11), those that did not (due to lack of space/ability to host practical labs) were designated as "Design Engineer" (n=23); n=11 chose not to identify their role. The groups met via video-conferencing (together and with instructor); students developed their own testing protocol, design configuration, students defended their design, then conducted data collection. Students performed well; however, students that self-selected as Process Engineer outperformed those that self-selected Design Engineer (although difference was not statistically significant). Slight difference may be attributed to hands-on or may be due to stronger students self-selected to be Process Engineers. This offers a promising design that increases the accessibility to those that are not able to go to campus.
Sasmito & Sekarsari (2022) No adult learners	VL (4D model, simulation,)	Chemistry Endothermic & Exothermic Reactions	HS K-12; I	Effectiveness Efficiency Engagement Ease of Use Turkey	n=63 (n=32, control group, watched video tutorial; n=31, experimental group, tested new simulation) There was a significant difference in students' mastery of understanding (test scores); mean score, control (74.31%); mean score, experimental (82.90%). Students indicated (interview) that the simulation was easy to use and an effective, fun, efficient way to complete the experiment; videos were boring. Virtual laboratories appeal to learners, were effective for learning, and make an excellent addition to science instruction.
Sotelo et al. (2022) Age ranges not reported	AHHO (kit)	Engineering Control Engineering	UG; M T4Y; I Private	Effectiveness Engagement Motivation Self-efficacy Mexico	The laboratory kit contained all materials necessary to complete the course objectives (components scaled down in size to be portable and cost-effective). Survey results (n=290) revealed that students overwhelmingly perceived the lab kit to be effective for learning, the kit was fun and engaging to use, provided a tangible experience with real-world equipment, and increased confidence. Final exam scores and numbers of A/B final grades was higher in the course (with kit) than previous terms (without kit). The use of the lab kit overcomes barriers of access (large class-size/limited internet in rural areas for simulations/remote labs) while supporting learning and increasing confidence and engagement.
Youngblood et al. (2022) Age ranges not reported Program: 26% in-person **74% online**	AHHO (kit)	Biology Vertebrate Zoology	UG; M T4Y Public RU	Effectiveness Ease of Use Self-efficacy Attitude	The laboratory kit included all of the materials to complete 8 vertebrate specimens. Pre-/post- course surveys (n=89) were used to evaluate anatomical self-confidence, confidence in laboratory skills, perceptions of support, and concerns about at-home dissections; Pre-/post-course surveys (n=148) to asses student attitudes towards dissections (helpfulness, likelihood to recommend). Students gained anatomical self-efficacy and confidence in practical skills; students indicated the gains were facilitated by hands-on experiences. Perception of support decreased; students indicated that support was unnecessary (instructions provided for each dissection was sufficient – extra support not needed). Students found value (and enjoyed) the hands-on dissections; students indicated the dissections helped in learning outcomes achievement. Some students were concerned about at-home dissections at the beginning of the course; No students were concerned at the end of the course. AHHO dissections provide students with the practical, hands-on skills that is missing from virtual lab simulations; as students confidence increased (and concerns about performing at-home dissections decreased), kits such as in this study can be used to broaden access to the laboratory.

Note. *TL = Traditional Laboratory (hands-on, in-person, on-campus), VL = Virtual Laboratory (as indicated), RL = Remote Laboratory, AHHO = At-Home Hands-On (as indicated), HO = Hands-On; B = blended (as indicated)

**UG = Undergraduate, HS = High School, MS = Middle School; NM = Non-majors course, M = Majors course

***CC = Community College; T4Y = Traditional 4-year, K-12 (Kindergarten through 12th grade); I = international (not conducted in the United States of America); RU = Research University

Source Table A-III

Summary of Non-Comparison Studies (Blended)

Authors	Discipline Course or Topic	Level** Type***	Constructs	Blended Approach	Results/Implications
Dalgarno et al. (2009) Age ranges not reported	Chemistry Intro. Chemistry	UG; NM T4Y; I Public RU	Effectiveness Self-efficacy Australia	Use of VL as Pre-laboratory Preparation Students in online class completed laboratories on campus; in an effort to reduce anxiety (and increase confidence) about the lab experience, students (n=12) were provided with a simulation to become familiar with the apparatus and familiarize themselves with the lab environment. Student that completed the simulation reported the tool helped them prepare for the on-campus lab experience; however, less than half the students utilized the tool. Interviews revealed that Math was the source of students' anxiety, not lack of familiarity of the lab.	The students that utilized the Virtual Laboratory simulation found it was an effective orientation to the layout of the laboratory and laboratory apparatus. The incorporation of simulations that allow student to apply mathematical techniques to chemistry concepts should be considered.
Husmann et al. (2009) Age ranges not reported	Biology Human Anatomy & Physiology	UG; M + NM T4Y Public RU	Effectiveness Efficiency Ease of Use United States	Optical Microscopes replaced with Virtual Microscopes Student performance improved (statistically significant for first 2 exams – greater focus on histology than other exams). Students were able to complete the lab in less time; focused on tissue identification rather than how to use microscope (more time on task); students were able to access the virtual microscope anytime	Use of virtual microscope increased students' comprehension of basic histology. Virtual microscopes are an effective tool; may prove financially desirable for large classes.
Jara et al. (2009) Age ranges not reported	Engineering	UG; M T4Y: I Public RU	Effectiveness Engagement Ease of Use Spain	Java Applets with real-time collaboration Students (n=25) completed synchronous laboratory simulations (with teacher guidance); completed questionnaire. Student performance increased (higher grades) with synchronous VL collaboration. 88% agreed that VLs help them understand concepts; 76% agreed that the collaborative system was easy to use; 64% agreed that the synchronous collaboration was helpful/learned from real-time feedback.	Suggests using VoIP technologies for synchronous collaboration during lab simulations.
Swan & O'Donnell (2009) Age ranges not reported	Biology General Biology	UG; T4Y Public RU	Effectiveness United States	Use of VL as supplemental instruction Students that utilized the VL (n=117) outperformed those that did not (n=666) on Exam 2, Laboratory Practical Exam, and Final Exam. Students that utilized the VL indicated that the VLs reinforced class material and helped to clarify course content.	VLs can provide valuable supplemental instruction; students that utilized VLs outperformed those that did not utilize the VL.

Authors	Discipline Course or Topic	Level** Type***	Constructs	Blended Approach	Results/Implications
Toth et al. (2009) Age ranges not reported	Biology Biological Inquiry: DNA and Gel Electrophoresis	UG; NM T4Y Private	Effectiveness United States	Combination of hands-on and virtual Purpose was to determine which order (Hands-on first or Virtual-first) best supports learning. Comparison of pre-/post-test scores indicated that order did not have a significant impact on end-state knowledge; starting with VL did have a slight benefit. Students' reflections indicate that students recognized the benefits of both formats (ease and speed of experiment, no errors; manual skill of loading the gel, see effect of errors). VL first: 84% found VL before TL to be beneficial prior to hands-on lab. VL second: 72% found no benefit to VL after performing the hands-on lab.	No significant difference in end-state knowledge; order does not impact conceptual understanding. Student feedback indicates that VL prior to TL may be more beneficial than TL prior to VL; VL serves as good preparation/practice for TL.
Saitta et al. (2011) Age ranges not reported	Chemistry General Chemistry	UG; M + NM T4Y Public RU and HS K-12	Effectiveness Engagement United States	Synchronous collaboration: Lab work and Post-lab Discussions UG students (n=21) were paired HS students as part of a Service Learning Project; UG students prepared pre-lab worksheets for HS students, prepared answer key, acted as a virtual lab partner, engaged in post-lab discussions to discuss results/prepare lab report. UG students perceived greater understanding of concepts; quiz scores support students' perceptions (Mean post-lab quiz grades for students that participated in the project, 64.6%; mean score for those that did not, 43.5%).	Used video-conferencing tool to facilitate lab partnership (both the UG and HS students completed the hands-on lab together at their respective campuses) and to facilitate post-lab discussions. Partnerships in which students are at different levels show greater gains in learning (and sense of engagement) for both partners.
Elmer et al. (2016) Age ranges not reported	Biology Exercise Physiology	UG; M T4Y Public RU	Effectiveness Engagement Self-efficacy United States	Use of VL as Pre-lab preparation Crossover study (group 1, n=16; group 2, n=17): Compared pre-/post- surveys and scores on assignments for TL and blended (students watched preparation and demonstration videos prior to completing in-person lab on campus).	No significant differences in assignment scores; students in blended format felt more confident in the lab and perceived learning of key foundational concepts was enhanced by pre-lab videos. The blended format may be an improvement over traditional laboratory instruction.
Bortnik et al. (2017) Age ranges not reported	Chemistry Analytical Chemistry	UG; M T4Y: I Public RU	Effectiveness Russia	Use of VL as Pre-lab preparation Students in the blended lab (n=25) completed VL as pre-laboratory preparation; students in the TL (n=25) had access to traditional instruction only. Students in the blended lab outperformed students in the TL in research skills and practices (scientific validity of goals and objectives, relevant description of methodology/instrumentation, make measurements, perform methodology, validity and clarity of conclusions, practicality and clarity of recommendations).	The use of the VL as pre-lab is an effective way to provide training and preparation; had positive impact on students' research skills and practices. VL that is coordinated with the hands-on practical lab can be a valuable teaching tool.
Davenport et al. (2018) No adult learners	Chemistry General Chemistry	HS K-12	Effectiveness Engagement United States	Use of VL as supplemental instruction Compared pre-/post-test scores of content knowledge (2 assessments developed with questions from California Standards Test in Chemistry, SATII Chemistry Subject Exam, and the New York Regents Examination), and transcripts from teacher interviews, based on: timing of use (introduction to topic, interwoven with lesson, as review) and mode of administration (as homework, individually in the classroom, pair in the classroom).	Post-test scores were significantly greater for students that performed the virtual laboratories as supplements to TL and as review following TL (blended instruction) – the VL labs alone did not perform well. Students working in pairs showed less improvement between pre-/posttest scores. Collaborative learning may be less beneficial when using simulations that provide customized feedback.

Authors	Discipline Course or Topic	Level[**] Type[***]	Constructs	Blended Approach	Results/Implications
Goudsouzian et al. (2018) Are ranges not reported	Biology Cell and Molecular Biology	UG; M T4Y Private	Effectiveness Engagement Self-Efficacy United States	Use of VL as supplemental instruction Compared pre-/post-test scores, perception of learning and self-efficacy survey responses for students: simulation alone (n=25); simulation plus live lab (n=115); and neither lab exercise (n=21).	There were no learning gains for students that did not engage in the laboratory activities (lecture alone is not sufficient to support learning). Students that engaged in simulation alone and simulation plus live lab showed learning gains, decrease in uncertainty, and increase in self-efficacy.; however, students that performed the live lab were more confident in their learning than those that engaged in the simulation alone. While the simulation alone is effective (can be used for students that miss lab), the combination of simulation and live lab was the most effective.
Meagher et al. (2018) Age ranges not reported	Earth Science Geology	UG; NM T4Y Public RU	Effectiveness United States	Use of images as Pre-lab preparation Compared ability to identify correct classification with no training (n=12); training with physical rocks (n=15); and training with images of rocks (n=15). After the training phase, participants were given a physical rock and asked to identify which of the 12-types rock types it falls into.	No significant differences in mean proportion of correct answers for physical training and image training; the mean proposition of correct answers for the group with no training was significantly lower than the groups with training. Image-based training can bus used to replace (less desirable) or supplement (more desirable) physical laboratories.
Davies (2019) Age ranges not reported Adult Learners	Chemistry Titration	UG; T4Y Public RU	Effectiveness United States	Use of Video as Pre-Lab Preparation Used multiperspective video to demonstrate proper use of burette and titration technique. Compared performance of students that trained in-person and trained via video.	No significant difference in performance between students (n=50) that were trained in lab (traditional method) and students (n=85) that were trained using video. Adult learners were able to develop laboratory skills in both formats. Use of videos can save time in lab.
de Toledo Durand et. al. (2019) Age ranges not reported	Biology Animal Physiology	G Medical School	Effectiveness Engagement United States	Animal Experiments Compared the perceptions of medical students that had access to animal laboratory only (n=120); students that had access to virtual classes only (videos) (n=108); and students that had access to both (n=122)	Students preferred combination of hands-on and virtual lab (blended). Students reported that videos alone were not adequate for learning.
Hamad & Aljanazrah (2020) Age ranges not reported	Physics General Physics	UG; M T4Y; I Public RU	Effectiveness Efficiency Palestine	Use of VL as Pre-lab preparation Compared outcomes for students that engaged in face-to-face preparation prior to lab (n=45) and virtual preparation prior to lab (n=45)	No significant difference in learning gains; students with virtual preparation were better prepared to carry out the real lab. Virtual labs as preparation to practical work has the potential to save time in lab.
Paxinou et al. (2020) Age ranges not reported	Education Biology/ Microscopy	UG; NM 4Y; I Public Online	Effectiveness Greece	Use of VL as Pre-lab preparation Compared students' ability to use the microscope when: trained through live demonstration at the beginning of the lab (n=30); trained by watching a video before attending the lab (n=29); trained by using a virtual microscope before attending the lab (n=24).	Students that trained on virtual microscope prior to attending the lab outperformed the other groups. Blending VL and TL promotes an effective learning environment.

Authors	Discipline	Level**	Constructs	Blended Approach	Results/Implications
	Course or Topic	Type***			
Çivril & Özkul (2021) **Adult Learners**	Physics Circuit Analysis	UG 2Y; I Public	Ease of Use Usefulness Turkey	Use of VL as Pre-lab preparation Used Technology Acceptance Model (TAM) to examine students' intention to utilize VL as preparation for lab. The strongest influence on students' intention to use VL was perceived usefulness (PU); learners will complete VL if they think it is useful (to improve laboratory performance, to increase performance on assignments). Perceived ease of use (PEU) was not a strong influence on attitudes or intention.	PU was the key factor in student's intention to use the VL. It is important for course designers/instructors to understand factors that influence learners' attitudes and behaviors.
Chang et al. (2022) No adult learners (all students <25 years old)	Biology Gross Anatomy Systemic Physiology	G Medical School; I	Effectiveness Engagement Taiwan	Use of VL and AHHO as Pre-Lab Preparation Pre-pandemic: Students engaged in live lectures to guide dissection at the beginning of lab time; 14-15 students per group (cadaver) – rotating roster of 4-6 "operators" per lab session, rest observed); groups engaged in peer discussion at the end of lab to review; each lab session 2-10 hours. Post-pandemic: Videos replaced live lectures at the beginning of each lab; 4 students per group (cadaver) – only "operators" present (to reduce the number of students in lab at given time); each lab session 2 hours.	Performance on lecture exams correlated with performance on lab exams in both conditions. Reducing the size of the group and time for peer discussion had a negative impact on performance (especially on lower performing students). The scores on lecture exams can be used as an indicator to lab performance; instructors can identify low performers and help prior to lab exam. The collective intelligence of the larger groups, time to watch other students' dissection techniques, and time for peer discussion at end of lab time seemed to be an important factor in learning; Instructors should take this under advisement, design labs with ample time/opportunity for "large group" peer collaboration.
Heng et al. (2022) Age ranges not reported	Biology Biomedical Science	UG; M T4Y; I Public RU	Effectiveness Efficiency Singapore	Use of VL as Pre-lab preparation Compared plasmid mini-prep yields of TL (pre-pandemic,5-hour lab sessions, n=59); blended (pandemic, 2.5-hour lab sessions, n=54). All students performed the experiments in the lab during the pandemic; students were subdivided into 3 groups; n=20, instructor live demonstration of lab skills (like TL), n=16 video demonstration, (blended), n=18, no demonstration (control).	No significant difference in yield and purity between instructor-live demonstration and video demonstration (when compare to pre-pandemic). However, the yield and purity were much lower for students in groups with no demonstration. Demonstration of laboratory skills prior impacts success of plasmid mini-prep; however, video demonstration is as effective as instructor-led demonstration. Use of video demonstration prior to lab can help with efficiency in the lab (less time spent in lab).

Note. *TL = Traditional Laboratory (hands-on, in-person, on-campus), VL = Virtual Laboratory (as indicated), RL = Remote Laboratory, AHHO = At-Home Hands-On (as indicated), HO = Hands-On; B = blended (as indicated)

**UG = Undergraduate, HS = High School, MS = Middle School; NM = Non-majors course, M = Majors course

***CC = Community College; T4Y = Traditional 4-year, K-12 (Kindergarten through 12th grade); I = international (not conducted in the United States of America); RU = Research University

Appendix B – Surveys, Crossover Phase

Brooke’s (1996) System Usability Scale (SUS): Standard or Classic Version

1. I think I would like to use this system frequently
2. I found this system unnecessarily complex.
3. I though the system was easy to use.
4. I think I would need the support of a technical person to be able to use this system.
5. I found the various functions in the system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found this system very awkward to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Lab 1 Survey and Lab 2 Survey

The Lab 1 Survey and the Lab 2 Survey included the following sections: (a) identification of the laboratory experience; (b) modified version of Brooke’s (1996) SUS questionnaire; (c) reflection questions from Reck et al. (2019), with permission; and (d) demographic information. All of the survey questions were identical, with the exception of Question 1. Both versions of question 1 are shown side-by-side (below).

Laboratory Experience

1. For Lab 1, I completed the experiment that:	For Lab 2, I completed the experiment that
Utilized real coins	Utilized a real metric ruler
Utilized a coin-flip simulator	Utilized a virtual metric ruler

Perceived Usability

2. Please rate the following statements in terms of how much you agree.

   (Strongly Disagree; Somewhat Disagree; Neutral; Somewhat Agree; Strongly Agree)

   I think I would like to do more lab experiments like this one.

   I found this lab experiment unnecessarily complex.

   I thought this lab experiment was easy to do.

   I think I would need more support to be able to do more lab experiments like this one.

   I found the various parts of this lab experiment to be well integrated.

   I thought there was too much inconsistency in this lab experiment.

   I would imagine that most people would be able to do lab experiments like this one very quickly.

   I found this lab experiment very awkward to do.

   I felt very confident doing this lab.

   I needed to learn a lot of things before I could do this lab experiment.

Reflection Questions

3. What aspects of this laboratory experience met your expectations?
4. What aspects of this laboratory experience did not meet your expectations?
5. What aspects of this laboratory did you find enjoyable?
6. How did this laboratory experience frustrate you?
7. What obstacles did you encounter during this laboratory experience?
8. What did you learn from this laboratory exercise?
9. What questions do you still have about this laboratory exercise?

Demographic Data

10. Gender: Male, Female, Other, Prefer not to say
11. Age: <25, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55+, Prefer not to respond
12. Race/Ethnicity: American Indian, Asian, Black or African American, Hawaiian/Pacific Islander, Hispanic, Two or More Races, White, Other, Prefer not to respond
13. Marital Status: Single, Married, Separated, Divorced, Widowed, Other
14. Children Under the age of 18: Yes, No
15. Student status: Full-time student (12 or more credit hours), Part-time student (less than 12 credit hours)
16. Employment Status (Full-time (work for than 36 hours per week), Part-time (work less than 36 hours per week), I am not currently employed

Appendix C – Laboratory Protocols, Crossover Phase

Laboratory 1 Instruction Sheet: Group A

SCIE 211 Lab 1: Probability and Statistics

Instructions

Introduction

The probability of coin flips is a classic problem that is used to illustrate concepts in statistics and sampling. **Probability** refers to the chance of something happening. Assuming you have a fair coin that has no chance of landing on its edge, there is an equal probability of it landing as "heads" (P_H) or "tails" (P_T):

$$P_H = P_T = ½ \text{ or } 50\%$$

Statistics is a branch of study concerned with the collection, analysis, evaluation of data. In statistics, measures of **central tendency** (such as mean, median, and mode) are used to identify a **central** or typical value for a **probability distribution**.

It is important to note P_H is ½ or 50% for each coin toss. If you toss a coin and it lands as "heads", this does not mean that the next coin toss will land as "tails". Or, that a coin is unfair if you flip it two times and each flip comes up "heads".

When one states that there is a 50% chance of a fair coin coming up "heads", this is over the long run. The more coin tosses you perform, the more likely you will see the expected 50% result. In this way, the size of your sample can have a profound effect on your results.

If we look at the probability distribution of 100 Coin tosses coming up "heads" (below), we will see the center of the distribution is 50/100 or 50% as predicted. The probability distribution of two-coin tosses coming up "heads" may or may not center on 50%.

Probability of heads from 100-coin tosses – Image removed for copyright purposes.

The **percent deviation** refers to how much the mean of a set of data differs from the expected value. If the observed data comes out as expected, there percent deviation = 0%. If the percent deviation is small (≤ 10 %), we can say it is due to chance. If the value is large (>10%), other factors may have affected the experiment.

See the sample data chart and formula below to see how to calculate % deviation.

Sample data chart: Single Coin – 10 flips

	Observed	Expected	Difference from Expected |observed – expected|
Heads	8	5	3
Tails	2	5	3
Total occurrences	10	10	Sum of differences = 6

% deviation = $\frac{sum\ of\ differences}{total\ occurances}$ X 100% % deviation = $\frac{6}{10}$ X 100% = 60%

The % deviation in this example is 60%. As this is much larger than 10%, the deviation cannot be attributed to chance. Can you think of any factors that may have affected the results?

In this laboratory, you will perform a coin flip experiment to explore how sample size affects data by looking at the % deviation for each trial.

Objectives:

After completing this laboratory, you should be able to:

- Explain the benefits and importance of the study of statistics.
- Discuss what is meant by the term probability.
- Discuss the importance of sampling to scientific studies.

Materials:

- 2 coins (same)
- Lab worksheet
- Calculator

Procedure – Part I: Single Coin Toss

1. **Toss a single coin 2 times.** Record your results on the Lab 1 Worksheet. Please refer to the Example Chart (above) to see how to record your results and how to calculate % deviation.
2. **Toss a single coin 10 times.** Record the number of heads AND tails that result from the 10 tosses in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in the chart).
3. **Toss the coin 50 times and again record the results**. Record the number of heads AND tails in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in chart).

Procedure – Part II: Double Coin Toss

When two independent events occur simultaneously, their individual "expected" probabilities are multiplied to determine the expected probability of them occurring together. To determine expected values for a double coin toss:

Possible Outcomes (coin 1-coin 2)	Individual Probabilities (coin 1-coin 2)	Probability for Pair
Heads – Heads	½ x ½	¼ or 25%
Heads – Tails	½ x ½	**½ or 50%
Tails – Heads	½ x ½	
Tails – Tails	½ x ½	¼ or 25%

**As heads-tails and tails-heads are the same result (one coin heads, the other coin tails), their probabilities are added together to give us odds of ½ or 50%.

1. **Toss two coins 4 times.** Record the number of heads-heads, heads-tails, and tail-tails in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in the chart). Note: heads-heads means that both coins are heads; heads-tails means that one coin is heads the other coin is tails; and tails-tails means that both coins are tails.
2. **Toss two coins 40 times.** Record the number of heads-heads, heads-tails, and tail-tails in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in the chart). Note: heads-heads means that both coins are heads; heads-tails means that one coin is heads the other coin is tails; and tails-tails means that both coins are tails.

Laboratory 1 Instruction Sheet: Group B

SCIE 211 Lab 1: Probability and Statistics

Instructions

Introduction

The probability of coin flips is a classic problem that is used to illustrate concepts in statistics and sampling. **Probability** refers to the chance of something happening. Assuming you have a fair coin that has no chance of landing on its edge, there is an equal probability of it landing as “heads” (P_H) or “tails” (P_T):

$$P_H = P_T = ½ \text{ or } 50\%$$

Statistics is a branch of study concerned with the collection, analysis, evaluation of data. In statistics, measures of **central tendency** (such as mean, median, and mode) are used to identify a **central** or typical value for a **probability distribution**.

It is important to note P_H is ½ or 50% for each coin toss. If you toss a coin and it lands as “heads”, this does not mean that the next coin toss will land as “tails”. Or, that a coin is unfair if you flip it two times and each flip comes up “heads”.

When one states that there is a 50% chance of a fair coin coming up “heads”, this is over the long run. The more coin tosses you perform, the more likely you will see the expected 50% result. In this way, the size of your sample can have a profound effect on your results.

If we look at the probability distribution of 100 Coin tosses coming up "heads" (below), we will see the center of the distribution is 50/100 or 50% as predicted. The probability distribution of two-coin tosses coming up "heads" may or may not center on 50%.

Probability of heads from 100-coin tosses – Image removed for copyright purposes.

The **percent deviation** refers to how much the mean of a set of data differs from the expected value. If the observed data comes out as expected, there percent deviation = 0%. If the percent deviation is small (≤ 10 %), we can say it is due to chance. If the value is large (>10%), other factors may have affected the experiment.

See the sample data chart and formula below to see how to calculate % deviation.

Sample data chart: Single Coin – 10 flips

	Observed	Expected	Difference from Expected |observed – expected|
Heads	8	5	3
Tails	2	5	3
Total occurrences	10	10	Sum of differences = 6

% deviation = $\frac{sum\ of\ differences}{total\ occurances}$ X 100% % deviation = $\frac{6}{10}$ X 100% = 60%

The % deviation in this example is 60%. As this is much larger than 10%, the deviation cannot be attributed to chance. Can you think of any factors that may have affected the results?

In this laboratory, you will perform a coin flip experiment to explore how sample size affects data by looking at the % deviation for each trial.

Objectives:

After completing this laboratory, you should be able to:

- Explain the benefits and importance of the study of statistics.
- Discuss what is meant by the term probability.
- Discuss the importance of sampling to scientific studies.

Materials:

- Coin simulator (https://flipsimu.com/)
- Lab worksheet
- Calculator

Procedure – Part I: Single Coin Toss

1. Click on the link for the coin-flip simulator: https://flipsimu.com/
2. Choose coin quantity under the settings tab. You should see one coin on the screen.

3. **Toss a single coin 2 times** (by clicking on the coin or the "Flip It" button under the image of the coin). Record your results on the Lab 1 Worksheet. Please refer to the Example Chart (above) to see how to record your results and how to calculate % deviation.
4. **Toss a single coin 10 times.** Record the number of heads AND tails that result from the 10 tosses in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in the chart).
5. **Toss the coin 50 times and again record the results**. Record the number of heads AND tails in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in chart).

Procedure – Part II: Double Coin Toss

When two independent events occur simultaneously, their individual "expected" probabilities are multiplied to determine the expected probability of them occurring together. To determine expected values for a double coin toss:

Possible Outcomes (coin 1-coin 2)	Individual Probabilities (coin 1-coin 2)	Probability for Pair
Heads – Heads	½ x ½	¼ or 25%
Heads – Tails	½ x ½	**½ or 50%
Tails – Heads	½ x ½	
Tails – Tails	½ x ½	¼ or 25%

**As heads-tails and tails-heads are the same result (one coin heads, the other coin tails), their probabilities are added together to give us odds of ½ or 50%.

1. Change the coin quantity to "2". You should see two coins on your screen.
2. **Toss two coins 4 times** (by clicking on each coin individually or the "Flip It" button under the image of the coins). Record the number of heads-heads, heads-tails, and tail-tails in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in the chart). Note: heads-heads means that both coins are heads; heads-tails means that one coin is heads the other coin is tails; and tails-tails means that both coins are tails.
3. **Toss two coins 40 times.** Record the number of heads-heads, heads-tails, and tail-tails in the designated data chart under OBSERVED (keep tally marks on separate sheet of paper and place only the total in the chart). Note: heads-heads means that both coins are heads; heads-tails means that one coin is heads the other coin is tails; and tails-tails means that both coins are tails.

Laboratory 1 Worksheet: Group A and Group B

SCIE 211 Lab 1: Probability and Statistics

Worksheet

Pre-lab Questions:

1. Define probability:

2. Define statistics:

3. Results of tossing a coin 4 times: H, H, H, H
    a. How many times is the coin expected to come up heads? How did you determine this number?
    b. Calculate the % deviation
    c. Can these results be used to conclude that a coin is not fair? Why or why not?

Data Charts:

Single Coin – Two flips

	Observed	Expected	Difference from Expected |observed – expected|
Heads			
Tails			
Total occurrences			Sum of differences =

% deviation =

Single Coin – 10 flips

	Observed	Expected	Difference from Expected |observed – expected|
Heads			
Tails			
Total occurrences			Sum of differences =

% deviation =

Single Coin – 50 flips

	Observed	Expected	Difference from Expected |observed – expected|
Heads			
Tails			
Total occurrences			Sum of differences =

% deviation =

Two Coins – 4 flips

	Observed	Expected	Difference from Expected |observed – expected|
Heads – Heads			
Heads – Tails			
Tails – Tails			
Total occurrences			Sum of differences =

% deviation =

Two Coins – 40 flips

	Observed	Expected	Difference from Expected |observed – expected|
Heads – Heads			
Heads – Tails			
Tails – Tails			
Total occurrences			Sum of differences =

% deviation =

Post-lab Questions:

1. Based upon your results, what effect (if any) does sample size have on the percent deviation for a single coin toss? For a double coin toss?

2. Do you think that the percent deviation would be closer to 0% is we had combined data from everyone in the class? Why or why not?

3. Does a small deviation mean that something was wrong with the experiment? Explain.

4. If three coins are flipped simultaneously, what is the probability that all three will be heads?

5. A penny tossed 120 times results in 62 heads and 58 tails.
   a. Calculate the expected number of heads and tails and determine the percent deviation.

   b. Do you think that the penny was a fair coin? Explain.

6. Applying what you have learned about probabilities: If man and a woman have five children (all girls), is it correct to assume that probability would favor their next child being a boy? Explain.

Laboratory 2 Instruction Sheet: Group A

SCIE 211 Lab 2: Measurements and Significant Figures

Instructions

Introduction

Since all measuring devices are subject to some error, it is impossible to make exact measurements. Scientists record all the digits of a measurement that are known exactly, plus the first one that is **uncertain**. These digits are collectively referred to as **significant digits**. Digital instruments, such as an electronic balance, are designed to limit themselves to the correct number of significant digits, and their readings are properly recorded as given. However, when using instruments such as rulers and thermometers, the experimentalist is responsible for determining the correct number of significant figures. These instruments are properly read to one place beyond the graduations of the scale.

Uncertainty and Significant Figures in Measurements – Image removed for copyright purposes.

In the figure above, notice how Ruler A and Ruler B are marked differently. Ruler A has markings for each whole number; Ruler B has makings for each whole number and the tenths place. A proper measurement contains one digit beyond the gradations of the scale. Therefore, the length of the object would be recorded differently for Ruler A and Ruler B:

Ruler A = 4.8 (tenths place is one place beyond gradations of the scale)

Ruler B = 4.82 (hundredths place is one place beyond the gradations of the scale)

You may consider the length of the object to be 4.81, 4.83 or 4.85. It is acceptable for the last digit in a measurement to be different --- it is understood that the last digit is uncertain (or the experimenter's best guess).

As there is uncertainty in every measurement, it is important for scientists to replicate their measurements. To get a feel for the **precision** of the measurement, scientists typically publish the mean and **standard deviation** for results (such as 4.83 ± 0.02, where 4.83 is the mean, 0.02 is the standard deviation). The standard deviation represents the extent of the "spread" of the measurements ---a large standard deviation shows a lack of precision (or consistency) in the measurements.

To calculate the standard deviation (represented by the symbol σ),

1. Calculate the mean (add up the measurements for each trial, then divide by the total number of trials)
2. Take the difference between each measurement and the mean.
3. Square the differences
4. Calculate the mean of the squared differences.
5. Take the square root.

See the sample data chart to see how to calculate standard deviation.

Trial #	Measurement	(Measurement – mean)	(Measurement – mean)^2
1	4.81	- 0.02	0.0004
2	4.82	- 0.01	0.0001
3	4.85	0.02	0.0004
mean	4.83	X	0.0003
		Standard deviation	0.02

Mean of measurements = $\frac{(4.81 + 4.82 + 4.85)}{3}$ = 4.83

Mean of squared differences = $\frac{(0.0004 + 0.0001 + 0.0004)}{3}$ = 0.0003

Standard deviation = $\sqrt{0.0003}$ = 0.02

We would report this result as 4.83 ± 0.02. This means that given the same measuring tool and object, most experimenters would report values ranging between 4.81 and 4.85 (within one standard deviation of the mean). Values that lie outside of 3 standard deviations are considered to be **outliers**. It is important to identify outliers as they can dramatically impact the mean value. See normal distribution below to see the percentage of data expected to fall within one standard deviation (1 σ), two standard deviations (2 σ), and three standard deviations (3 σ).

Normal Distribution – Image removed for copyright purposes.

Finally, when performing calculations with measured values, it is important to limit your final result to the correct number of significant figures. If we were to calculate the volume of a cube (all sides equal length) using our measurement above, we would get the following:

Volume of object = length x width x height
= 4.83 x 4.83 x 4.83
Calculator answer = 112.678587

As written, the calculator answer implies a level of precision our instrument does not support. Remember, only the last digit is considered to be uncertain. This would mean that our instrument was reliable to the 100,000ths place! This would be an example of the fallacy of false precision.

There are rules for reporting the correct number of significant figures in calculations:

1. Addition/Subtraction: The final answer will have the same number of decimal places as the measurement with the least number of decimal places.

   Example: 4.83 + 4.8 = 9.63 (calculator answer)
   = 9.6 (with correct number of significant figures, as the tenths place is the least number of decimal places).

   Note: It is important to perform the calculation, then round to the correct number of significant figures. You do not want to automatically drop the "extra decimal places" before the calculation. Dropping decimal places introduces calculation error to our reported value.

2. Multiplication/Division: The final answer will have the same number of significant figures as the measurement with the least number of significant figures.

   Example: 4.83 x 4.83 x 4.83 = 112.678587 (calculator answer)

   = 113 (with correct number of significant figures, as each of our measurements has 3 significant figures, our answer can only have 3 significant figures).

   Note: If the measurements have different numbers of significant figures, your answer has the same number as the measurement with the LEAST number of significant figures (SF).

   $$\underset{\text{Example: }}{}\overset{\text{3 SF}}{4.25} \times \overset{\text{2 SF}}{0.75} \times \overset{\text{4 SF}}{3.555} = 11.3315625 \text{ (calculator answer)}$$

   = 11 (with correct number of significant figures)

In this laboratory, you will measure various household objects and find their area and volume.

Objectives:
After completing this laboratory, you should be able to:

- Explain some fallacies of data quality and characterization
- Identify the strengths and weaknesses of the three measures of central tendency: mean, median, and mode.
- Collect and analyze data

Materials:
- Virtual ruler (https://www.ginifab.com/feeds/cm_to_inch/virtual_ruler_on_your_image.html)
- Lab worksheet
- Calculator

Procedure – Part I: Measurements
1. Click on the link for the virtual ruler: (https://www.ginifab.com/feeds/cm_to_inch/virtual_ruler_on_your_image.html)
2. Choose images for three household items.
3. Identify each object on the line provided above the data table provided on the Lab Worksheet.
4. Upload the image for the first object by clicking on choose file button at bottom of screen. You should see the object on the screen with the virtual ruler (an image of a pencil is used for an example).

   Image that shows how to measure object with virtual ruler – Image removed for copyright purposes.

5. Measure the length of each object at least three times (click on the ruler to drag it – be sure to measure objects in cm). Record your measurements in data tables.

   Note: After you make your first measurement, drag the ruler away from the object. Then drag the ruler back to the object for your second measurement. Repeat this for your third measurement.

   Also note: The size of the virtual object may be larger or smaller than the real object. As such, the length of the object may not be the same as the length of the "real" object. It does not matter for the purposes of this laboratory. The object of this experiment is to practice making measurements to the correct number of significant figures.

6. Click on the download button (at the bottom of the screen) to save an image of at least one of your measurements. Insert the image in the space provided on the Lab 2 Worksheet (where you are asked to "insert a picture of your metric measuring tool")
7. Calculate the mean length and standard deviation for each of your objects.

Procedure – Part II: Calculating Volume

1. Locate an image of a rectangular object (box of cereal, book, etc.).
2. Identify the object on the line provided above the data table provided on the Lab Worksheet.
3. Measure the length, width, and height of the object at least three times (each). You can rotate the ruler as needed by clicking on the rotate button at the bottom of the screen.
4. Calculate the mean for each.
5. Using the mean values, calculate the volume. Be sure to report the volume using the correct number of significant figures.

Laboratory 2 Instruction Sheet: Group B

SCIE 211 Lab 2: Measurements and Significant Figures

Instructions

Introduction

Since all measuring devices are subject to some error, it is impossible to make exact measurements. Scientists record all the digits of a measurement that are known exactly, plus the first one that is **uncertain**. These digits are collectively referred to as **significant digits**. Digital instruments, such as an electronic balance, are designed to limit themselves to the correct number of significant digits, and their readings are properly recorded as given. However, when using instruments such as rulers and thermometers, the experimentalist is responsible for determining the correct number of significant figures. These instruments are properly read to one place beyond the graduations of the scale.

Uncertainty and Significant Figures in Measurements – Image removed for copyright purposes.

In the figure above, notice how Ruler A and Ruler B are marked differently. Ruler A has markings for each whole number; Ruler B has makings for each whole number and the tenths place. A proper measurement contains one digit beyond the gradations of the scale. Therefore, the length of the object would be recorded differently for Ruler A and Ruler B:

Ruler A = 4.8 (tenths place is one place beyond gradations of the scale)

Ruler B = 4.82 (hundredths place is one place beyond the gradations of the scale)

You may consider the length of the object to be 4.81, 4.83 or 4.85. It is acceptable for the last digit in a measurement to be different --- it is understood that the last digit is uncertain (or the experimenter's best guess).

As there is uncertainty in every measurement, it is important for scientists to replicate their measurements. To get a feel for the **precision** of the measurement, scientists typically publish the mean and **standard deviation** for results (such as 4.83 ± 0.02, where 4.83 is the mean, 0.02 is the standard deviation). The standard deviation represents the extent of the "spread" of the measurements ---a large standard deviation shows a lack of precision (or consistency) in the measurements.

To calculate the standard deviation (represented by the symbol σ),

1. Calculate the mean (add up the measurements for each trial, then divide by the total number of trials)
2. Take the difference between each measurement and the mean.
3. Square the differences
4. Calculate the mean of the squared differences.
5. Take the square root.

See the sample data chart to see how to calculate standard deviation.

Trial #	Measurement	(Measurement – mean)	(Measurement – mean)^2
1	4.81	- 0.02	0.0004
2	4.82	- 0.01	0.0001
3	4.85	0.02	0.0004
mean	4.83	X	0.0003
		Standard deviation	0.02

Mean of measurements = $\frac{(4.81 + 4.82 + 4.85)}{3}$ = 4.83

Mean of squared differences = $\frac{(0.0004 + 0.0001 + 0.0004)}{3}$ = 0.0003

Standard deviation = $\sqrt{0.0003}$ = 0.02

We would report this result as 4.83 ± 0.02. This means that given the same measuring tool and object, most experimenters would report values ranging between 4.81 and 4.85 (within one standard deviation of the mean). Values that lie outside of 3 standard deviations are considered to be **outliers**. It is important to identify outliers as they can dramatically impact the mean value. See normal distribution below to see the percentage of data expected to fall within one standard deviation (1 σ), two standard deviations (2 σ), and three standard deviations (3 σ).

Normal Distribution – Image removed for copyright purposes.

Finally, when performing calculations with measured values, it is important to limit your final result to the correct number of significant figures. If we were to calculate the volume of a cube (all sides equal length) using our measurement above, we would get the following:

Volume of object = length x width x height

= 4.83 x 4.83 x 4.83

Calculator answer = 112.678587

As written, the calculator answer implies a level of precision our instrument does not support. Remember, only the last digit is considered to be uncertain. This would mean that our instrument was reliable to the 100,000ths place! This would be an example of the fallacy of false precision.

There are rules for reporting the correct number of significant figures in calculations:

1. Addition/Subtraction: The final answer will have the same number of decimal places as the measurement with the least number of decimal places.

   Example: 4.83 + 4.8 = 9.63 (calculator answer)
   = 9.6 (with correct number of significant figures, as the tenths place is the least number of decimal places).

   Note: It is important to perform the calculation, then round to the correct number of significant figures. You do not want to automatically drop the "extra decimal places" before the calculation. Dropping decimal places introduces calculation error to our reported value.

2. Multiplication/Division: The final answer will have the same number of significant figures as the measurement with the least number of significant figures.

   Example: 4.83 x 4.83 x 4.83 = 112.678587 (calculator answer)

   = 113 (with correct number of significant figures, as each of our measurements has 3 significant figures, our answer can only have 3 significant figures).

   Note: If the measurements have different numbers of significant figures, your answer has the same number as the measurement with the LEAST number of significant figures (SF).

   3 SF 2 SF 4 SF

   Example: 4.25 x 0.75 x 3.555 = 11.3315625 (calculator answer)

   = 11 (with correct number of significant figures)

In this laboratory, you will measure various household objects and find their area and volume.

Objectives:

After completing this laboratory, you should be able to:

- Explain some fallacies of data quality and characterization
- Identify the strengths and weaknesses of the three measures of central tendency: mean, median, and mode.
- Collect and analyze data

Materials:

- Metric ruler (measures cm)
- Lab worksheet
- Calculator

Procedure – Part I: Measurements

1. Take a picture of your metric ruler. Insert a picture of your ruler in the space provided on the Lab worksheet.
2. Locate three household items (large, medium, and small).
3. Identify each object on the line provided above the data table provided on the Lab Worksheet. Measure the length of each object at least three times. Record your measurements in data tables. Note: After you make your first measurement, move the ruler away from the object. Then place the ruler back on the object for your second measurement. Repeat this for your third measurement.
4. Calculate the mean and standard deviation for each of your objects.

Procedure – Part II: Calculating Volume

1. Locate a rectangular object (box of cereal, book, etc.).
2. Identify the object on the line provided above the data table provided on the Lab Worksheet.
3. Measure the length, width, and height of the object at least three times (each).
4. Calculate the mean for each.
5. Using the mean values, calculate the volume. Be sure to report the volume using the correct number of significant figures.

Laboratory 2 Worksheet: Group A and Group B

SCIE 211 Lab 2: Measurements and Significant Figures

Worksheet

Pre-lab Questions:

1. Define accuracy and precision:

2. Based upon the gradations on the ruler in the picture below, what is the length of the arrow? How many decimal places should your measurement include?

Students are asked to measure the length of the arrow with image of a ruler -Image removed for copyright purposes.

3. A data set contains the following measurements: 13.22, 13.48 cm, 13.49 cm, 13.49, 13.50

    a. Calculate the mean and standard deviation for the data set.

    b. Based upon the values in this data set, do you think the mean or the median would be a better estimation of the center of the data set? Explain.

4. A wall in your living room is 23.35 ft long and 15.23 ft high.
    a. Calculate the area of the wall (Area = length x height)
    b. Report your answer using the correct number of significant figures.

Data Charts:

Insert a picture of your metric measuring tool. Based upon your measuring device, how many decimal places should each of your measurements include?

Measurements: Length Object 1 – Large (table, couch, etc.)

Identify the object: ____

Trial #	Measurement (cm)	(Measurement – mean) (cm)	(Measurement – mean)^2 (cm^2)
1			
2			
3			
mean		X	
		Standard deviation	

Measurements: Length Object 2 – Medium (book, cookie sheet, etc.)

Identify the object: ____

Trial #	Measurement (cm)	(Measurement – mean) (cm)	(Measurement – mean)^2 (cm^2)
1			
2			
3			
mean		X	
		Standard deviation	

Measurements: Length Object 3 – Small (matchbox car, penny, etc.)

Identify the object: ____

Trial #	Measurement (cm)	(Measurement – mean) (cm)	$(\text{Measurement} - \text{mean})^2$ (cm^2)
1			
2			
3			
mean		X	
		Standard deviation	

Calculating Volume: Identify the object: ______________________

Trial #	Length (cm)	Width (cm)	Height (cm)
1			
2			
3			
mean			

Volume (cm^3): ________________________

Post-lab Questions:

1. Based upon your results, what effect (if any) does the size of the object have on the standard deviation for your measurement?

2. Choosing the proper measuring device for the "job" is important. Was using a standard metric ruler the best measuring device for the large object? If not, what tool would have been a better choice?

3. Did you have any outliers in your data set? How do you know (i.e. how do you determine if a data point is an outlier)?

4. You were easily able to calculate the volume of a regularly shaped object. How might you determine the volume of an object if it has an irregular shape? Would you need a different measuring tool?

Appendix D – Laboratory Protocols, Self-Selection Phase

Laboratory 4 Instruction Sheet: Hand-On

SCIE 211 Lab 4: Correlation and Causation

Instructions

Introduction

Identifying the extent and type of relationships between variables (or phenomena, events, traits, etc.) is an important aspect of the scientific method. The term **correlation is used to refer a mutual relationship that is thought to exist between any two variables, phenomena, events, etc. While** correlation plays an underlying role in establishing causation (i.e. the events must be related in order for A to cause B), the existence of a relationship between variables is not sufficient to imply causation.

The process of successfully attributing causation is difficult; the **causal chain** (or **pathway**) of events is often not clear. Can we be sure that **A** causes **B**, or is it, in fact, the other way around that **B** causes **A**? Or, is there a hidden third, extraneous, or confounding factor **C** that can cause one or the other or both? This hidden or **lurking** third factor (another variable at play/an alternative explanation) is called the **extraneous, spurious,** or **confounding variable**.

There are **three** criteria for figuring out whether or not there is **evidence** for causation:

1) **There exists a strong and consistent correlation. This means** that when the alleged cause **A** is present, the alleged effect **B** tends to be present as well. Also, there should be a plausible **explanatory model (that is consistent with the data and fits with other scientific understanding)** so we can **explain the correlation**.

2) **There is precedence. This** means that both the time-order and direction have been established. In other words, to say that **A causes B**, the cause **A** must **come before** effect **B ---or that B does not happen unless A occurs first.**

3**) All other confounding factors (lurking, spurious, extraneous, or third variables) or alternative explanations have been ruled out. Causation can only be established if A has been shown to directly cause B, without any other intervening variables. This allows us to make predictions in advance, that A** will cause **B. This predicative relationship can be seen, for example, in the dose relationship --- the** *larger* the dosage, the stronger the response.

To demonstrate this principle, you will look at the relationship between the diameter of a balloon and lung volumes.

Tidal volume is the volume of air that you move into and out of your lungs each time you breathe normally. Average tidal volume is roughly 500 mL (Moini, 2020)

The amount of air that can be inspired forcibly beyond tidal volume is called the **inspiratory reserve volume (IRV)**. The average inspiratory reserve volume varies dramatically with sex: 3000 mL in males, 2100 mL in females (Moini, 2020).

The amount of air that can be expelled forcibly beyond tidal volume is called the **expiratory reserve volume (ERV)**. Just like the inspiratory reserve volume, the expiratory reserve volume also varies dramatically with sex: 1100 mL in males, 800 mL in females (Moini, 2020).

Vital capacity represents the maximum amount of exchangeable air that our lungs can move. It is the sum of the tidal volume, inspiratory reserve volume, and expiratory reserve volume, or:

$$\text{Vital capacity (mL)} = \text{Tidal volume (mL)} + \text{IRV (mL)} + \text{ERV (mL)}$$

In this laboratory, you will use a balloon to measure your tidal volume (volume of air moved in normal breath) and vital capacity (maximum volume of air moved). You will consider lurking variables/confounding factors that contributed to any deviation from expected values.

Objectives:

After completing this laboratory, you should be able to:

- Discuss the danger of confusing correlation with causation.
- Identify extraneous or confounding factors.
- Explain how we determine support for causation.
- Evaluate causal claims.

Materials:

- 1 Balloon (12 inch)
- Metric Ruler
- 1 binder clip (or fingers)
- Pencil
- Calculator

Procedure – Part I: Measuring Tidal Volume

1. Blow up and deflate the balloon several times to stretch it out.
2. Sit down and relax. Inhale normally and then exhale only that of a normal breath into the balloon.
3. Immediately twist and clamp the balloon (or pinch with your fingers) so that no air escapes.
4. Place the tip of a pencil vertically onto the zero cm mark on the ruler. Place the balloon on its side next to the pencil. (See figure 1).
5. Holding the balloon in position, move your pencil to the other side of the balloon.
6. Record the diameter (cm) in the table on the Lab 4 Worksheet. Make sure your recorded value has the correct number of significant figures.
7. Repeat the entire process 2 more times.
8. Calculate the average balloon diameter for your three trials on the Lab 4 Worksheet.

Figure 1 - Image shows how to measure diameter of balloon: Image removed for copyright purposes.

Procedure – Part II: Measuring Vital Capacity

1. Use the same balloon as used in Part I.
2. Inhale as deeply as you can and exhale as much air into the balloon as you can.
3. Immediately twist and clamp the balloon (or pinch with your fingers) so that no air escapes.
4. Place the tip of a pencil vertically onto the zero cm mark on the ruler. Place the balloon on its side next to the pencil. (See figure 1).
5. Holding the balloon in position, move your pencil to the other side of the balloon.
6. Record the diameter (cm) in the table on the Lab 4 Worksheet. Make sure your recorded value has the correct number of significant figures.
7. Repeat the entire process 2 more times.
8. Calculate the average balloon diameter for your three trials on the Lab 4 Worksheet.

Analysis

1. Use the graph in Figure 2 (below) to estimate the average diameters of your balloon filled by a normal breath and the balloon filled by a deep breath into cc (cubic centimeters) of lung volume. Record the values in the table on the Lab 4 Worksheet.
2. Using the conversion **1 cc = 1 mL**, convert the volumes in cc to volumes in mL. Record these values in the table on the Lab 4 Worksheet. These represent your observed tidal volume and observed vital capacity.
3. Record the expected values (from pre-lab questions 1 an 2) in the table on the Lab 4 Worksheet.
4. Calculate and record the % difference. Remember, the % difference is:

$$\% \text{ difference} = \frac{|(\text{Observed volume} - \text{Expected volume})}{\text{Expected volume}} \times 100\%$$

Figure 2 – Graph depicting relationship between diameter of balloon and volume of air: Image removed for copyright purposes.

References

Moini, J. (2020). Anatomy and Physiology for Health Professionals, 3rd Edition. Jones and Bartlett Learning.

Measuring Lung Capacity (n.d.). http://www.biologycorner.com/worksheets/lungcapacity.html

Laboratory 4 Worksheet: Hand-On

SCIE 211 Lab 4: Correlation and Causation

Worksheet

Pre-Lab Questions

1. Based upon the information provided in the Lab 4 Introduction, what is your expected value for tidal volume?

   Tidal Volume = ___________ mL

2. Based upon the information provided in the Lab 4 Introduction, calculate the expected value for your vital capacity?

   Vital Capacity = _________ mL + _________ mL + _________ mL = _____________ mL

3. Do you think that the diameter of the balloon will have a positive, negative, or no correlation with lung volumes? Explain.

4. Write a hypothesis for this experiment (What is the relationship between the diameter of the balloon and lung volumes? Which diameter will be larger, tidal volume or vital capacity?).

   Hypothesis:

Data Charts:

Part I: Diameter of balloon with normal breathing – Tidal Volume

	Diameter of balloon (cm)
Trial 1	
Trial 2	
Trial 3	
Average	

Part II: Diameter of balloon with deep breathing – Vital Capacity

	Diameter of balloon (cm)
Trial 1	
Trial 2	
Trial 3	
Average	

Analysis:

Observed Volumes:

	Average Diameter (cm)	Volume from graph (cc)	Observed Volume in mL
Tidal Volume			
Vital Capacity			

% Deviation

	Observed Volume in mL	Expected Volume in mL	% Deviation
Tidal Volume			
Vital Capacity			

Post-lab Questions:

1. Why do you think you were asked to inflate/deflate the balloon several times before starting your trials?
2. Why do you think you were asked to use the same balloon for Part I and Part II of the experiment?
3. Was your proposed hypothesis supported by your evidence? Why or why not?
4. Why do you think that there are differences in expected inspiratory and expiratory reserve volumes for males and females?
5. Are the values for your tidal volume and vital capacities what you expected them to be? If not, please indicate any extraneous or confounding factors that may have caused your observed value to be greater or less than expected (You may want to consider any measurement issues, health issues, etc. that could have contributed to the difference.)
6. Is there sufficient evidence to support that the extraneous/confounding factors you identified are correlated to changes in expected lung volumes? Why or why not?
7. Is there sufficient evidence to support that the extraneous/confounding factors you identified CAUSED your observed value to differ from the expected?

Laboratory 4 Instruction Sheet: Virtual

SCIE 211 Lab 4: Correlation and Causation

Instructions

Introduction

Identifying the extent and type of relationships between variables (or phenomena, events, traits, etc.) is an important aspect of the scientific method. The term **correlation is used to refer a mutual relationship that is thought to exist between any two variables, phenomena, events, etc. While** correlation plays an underlying role in establishing causation (i.e. the events must be related in order for A to cause B), the existence of a relationship between variables is not sufficient to imply causation.

The process of successfully attributing causation is difficult; the **causal chain** (or **pathway**) of events is often not clear. Can we be sure that **A** causes **B**, or is it, in fact, the other way around that **B** causes **A**? Or, is there a hidden third, extraneous, or confounding factor **C** that can cause one or the other or both? This hidden or **lurking** third factor (another variable at play/an alternative explanation) is called the **extraneous, spurious,** or **confounding variable**.

There are **three** criteria for figuring out whether or not there is **evidence** for causation:

1) **There exists a strong and consistent correlation. This means** that when the alleged cause **A** is present, the alleged effect **B** tends to be present as well. Also, there should be a plausible **explanatory model (that is consistent with the data and fits with other scientific understanding)** so we can **explain the correlation**.

2) **There is precedence. This** means that both the time-order and direction have been established. In other words, to say that **A causes B**, the cause **A** must **come before** effect **B ---or that B does not happen unless A occurs first.**

3**) All other confounding factors (lurking, spurious, extraneous, or third variables) or alternative explanations have been ruled out. Causation can only be established if A has been shown to directly cause B, without any other intervening variables. This allows us to make predictions in advance, that A** will cause **B. This predicative relationship can be seen, for example, in the dose relationship --- the** *larger* the dosage, the stronger the response.

To demonstrate this principle, you will look at the relationship between the organic food sales and autism. The connection between the MMR vaccine and autism has been debunked (scientists admitted to falsifying their data for this study). Perhaps the increased prevalence of autism is due to organic food sales????

Objectives:

After completing this laboratory, you should be able to:

- Discuss the danger of confusing correlation with causation.
- Identify extraneous or confounding factors.
- Explain how we determine support for causation.
- Evaluate causal claims.

Materials:

- Internet resources
- Graph (provided in Part III of the experiment)

Procedure – Part I: Organic Foods

1. Research what is meant by the term organic foods.
2. Identify at least two credible sources of information on this topic (videos, articles, informational websites).
3. Provide a copy of the link for each resource and a summary of the key ideas presented in each in Data Table I.
4. Justify why you consider the source to be credible. Provide your justification in the space below Data Table I.

Procedure – Part II: Autism

1. Research what is meant by the term autism.
2. Identify at least two credible sources of information on this topic (videos, articles, informational websites).
3. Provide a copy of the link for each resource and a summary of the key ideas presented in each in Data Table II.
4. Justify why you consider the source to be credible. Provide your justification in the space below Data Table II.

Procedure – Part III: Relationship Between Organic Food Sales and Prevalence of Autism

1. Examine the relationship between these variables in the graph below:

   Graph depicting relationship between autism and organic food sales: Image removed for copyright purposes.

2. Based upon the information from your research and the relationship demonstrated in the graph above, answer the Part III questions in the spaces provided on the Lab 4 Worksheet.

Laboratory 4 Worksheet: Virtual

SCIE 211 Lab 4: Correlation and Causation

Worksheet

Pre-Lab Questions

1. Based upon your research, briefly define what is meant by the term “organic food” and the (supposed) health benefits from consumption of organic food products.

2. Based upon your research, briefly define what is meant by the term “autism” and identify when it is developed.

Data Tables:

Data Table I: Organic Foods

Link for Resource	Summary of Key Ideas
1.	
2.	

Why do you consider these sources to be credible?

Data Table II: Autism

Link for Resource	Summary
1.	
2.	

Why do you consider these sources to be credible?

Part III: Relationship between Organic Food Sales and Prevalence of Autism

Graph depicting relationship between autism and organic food sales: Image removed for copyright purposes.

1. Based upon the graph above,
    a. Is there a positive correlation, negative correlation, or no correlation between organic food sales and prevalence of autism?
    b. How did you determine your answer for part A?
2. Describe at least one conclusion you could draw from the graph.
3. Are you convinced that increased consumption of organic food could explain the increased incidence of autism? Why or why not?

4. Design an experiment that would directly test the hypothesis that increased consumption of organic food will increase the incidence of autism.
5. Identify at least 3 confounding factors that should be considered in your experiment.
6. Even if you performed this experiment, would you be able to establish causation? Why or why not?

Post-lab Questions:

1. Why do you think you were asked to research what is meant by organic foods (and their potential effects on consumer health) and autism (and when it is developed)?
2. Why do you think you were asked to justify the credibility of your resources?
3. Why is it so difficult to establish causation?
4. It is easy to confuse causation with correlation. Find a recent news story (or article) that improperly describes correlation as causation.
    a. Provide the link for the news story (or article)
    b. Briefly summarize the purpose of the story (or study).
    c. Briefly summarize the conclusions of the story (or study).
    d. Briefly summarize how the author incorrectly describes correlation for causation.
    e. Is mistaking correlation for causation be detrimental in this example?

Appendix E – Survey, Self-Selection Phase

Lab 4 Survey

The Lab 4 Survey included the following sections: (a) identification of the laboratory experience; (b) reflection questions; and (c) demographic information.

Laboratory Experience

1. For Lab 4, I completed the experiment that:
   Utilized balloons to measure tidal volume and vital capacity

   Utilized research to examine the relationship between organic food sales and the prevalence of autism

Reflection Questions

2. Why did you choose to complete this laboratory exercise?
3. Was there something about the other laboratory exercise that made it less appealing to you?
4. In your opinion, which of the following factor(s) are the most important for a good laboratory experience? Why? You can choose more than one.

   Effectiveness: the extent to which you are able to achieve the specified goals of the experiment

   Efficiency: the extent to which you are able to complete the experiment within the allotted timeframe

   Engagement: the extent to which you found the experience to be interesting and/or enjoyable

Error tolerance: the extent to which you are able to recover from a mistake without having to start over

Ease of Use: the extent to which you were able complete the laboratory without intervention (support)

Self-efficacy: the extent to which you believe you are capable of performing the tasks required to complete the laboratory

5. In your opinion, which of the following factor(s) are the least important for a good laboratory experience? Why? You can choose more than one.

   Effectiveness: the extent to which you are able to achieve the specified goals of the experiment

   Efficiency: the extent to which you are able to complete the experiment within the allotted timeframe

   Engagement: the extent to which you found the experience to be interesting and/or enjoyable

   Error tolerance: the extent to which you are able to recover from a mistake without having to start over

   Ease of Use: the extent to which you were able complete the laboratory without intervention (support)

   Self-efficacy: the extent to which you believe you are capable of performing the tasks required to complete the laboratoryIn your opinion, How did this laboratory experience frustrate you?
6. What other factors are important for a good laboratory experience? Why?

Demographic Data

7. Gender: Male, Female, Other, Prefer not to say
8. Age: <25, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55+, Prefer not to respond
9. Race/Ethnicity: American Indian, Asian, Black or African American, Hawaiian/Pacific Islander, Hispanic, Two or More Races, White, Other, Prefer not to respond
10. Marital Status: Single, Married, Separated, Divorced, Widowed, Other
11. Children Under the age of 18: Yes, No
12. Student status: Full-time student (12 or more credit hours), Part-time student (less than 12 credit hours)
13. Employment Status (Full-time (work for than 36 hours per week), Part-time (work less than 36 hours per week), I am not currently employed

Appendix F – Permission to Use SUS

Email communication from John Brooke

Hi Amiee

You’re welcome! SUS is free for use (and always has been) as long as the source is acknowledged in any publication (the ‘being free’ bit largely explains why it’s got something like 13500 citations ☺).

If you have any questions, you’re welcome to get back to me any time.

Regards

John Brooke

From: Amiee Wagner [mailto:amiee.wagner@franklin.edu]
Sent: 26 April 2022 19:48
To: john.brooke@poundlane.net
Subject: SUS survey

Dr. Brooke,

I am writing to you to thank you for making the SUS readily available for use in studies. It was such a delight to discover a free tool that has been normed, validated, shown to be reliable even with small sample sizes!

While I will be sure to give you credit in my dissertation, it would be wonderful to receive your blessing for the use of the survey. I am sure that you understand the time and energy savings that come with using a tool that has been rigorously tested with multiple user populations and with multiple products. I will be eternally grateful for your willingness to share such a robust tool that is so quick and easy to use!

Thank you!

Amiee

Appendix G – Permission to Use Reflection Questions

Email communication from Rebecca Reck

Amiee,

Thank you for your interest in my work. It sounds like you are working on a very interesting project. You are welcome to use the same reflection questions. I was amazed at the breadth and depth of the content in those reflections.

Please let me know if you have any other questions.

Best,

Rebecca

REBECCA RECK, PH.D. (she/her)
Teaching Associate Professor

Department of Bioengineering | The Grainger College of Engineering

1406 W. Green Street | 3130 Everitt Lab, MC 278 | Urbana, IL 61801
(P) 217.300.8461 | rreck@illinois.edu

https://bioengineering.illinois.edu/

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

From: Amiee Wagner <amiee.wagner@franklin.edu>
Sent: Wednesday, June 29, 2022 1:51 PM
To: Reck, Rebecca <rreck@illinois.edu>
Subject: A fan of your work....

Dr. Reck,

I am writing to you regarding a paper you published in 2019, Evaluating the Effectiveness of an Affordable and Portable Laboratory Kit for an Introductory Control Systems Course. I very much appreciated the mixed methods design that you employed for this study….and that it is one of the few studies that compares students' laboratory experiences in the same laboratory environment! Your paper is heavily referenced in my research proposal.

I am currently working on my dissertation in practice – a mixed methods comparison study to determine students' preferred mode of at-home laboratory instruction (hands-on or virtual) for an online, non-majors science course at my institution. The reflection questions that you used for your study cut right to the heart of the student experience. With your blessing, I would like to use reflection questions that are very similar to the ones you used in your study. I would, of course, acknowledge your work as the source of inspiration.

Thank you for your excellent work! I look forward to hearing for you (either way).

Amiee Wagner