5-6-2022

Investigating Active Learning through the Lens of Student Engagement

Nicole Naibert
Portland State University

Recommended Citation
Naibert, Nicole, "Investigating Active Learning through the Lens of Student Engagement" (2022). *Dissertations and Theses.* Paper 6012.
https://doi.org/10.15760/etd.7883

Investigating Active Learning through the Lens of Student Engagement

by

Nicole Naibert

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

Doctor of Philosophy
in
Chemistry

Dissertation Committee:
Jack Barbera, Chair
Erin E. Shortlidge
Gwen Shusterman
Dean Atkinson

Portland State University
2022

Abstract

Incorporating active learning into a course has been generally found to lead to improved student learning outcomes; however, not all students benefit from these environments to the same extent. Although active learning environments provide the opportunity for students to interact and engage with the material, whether a student decides to do so is completely up to them. Therefore, the goal of this dissertation was to begin exploring active learning environments through the lens of student engagement and relevant associated variables (i.e., self-efficacy and student perceptions). This was completed through three separate but related projects.

Project I focused on investigating flipped courses at five different institutions, specifically in relation to students' interactions with and perceptions of pre-class materials (PCMs), as well as their self-efficacy. Students' interactions with and perceptions of required pre-class videos for each course were evaluated through student survey responses. A possible trend was found between the amount of peer-to-peer interaction included during the face-to-face (F2F) class time and how many videos students watched and when they watched them. Student responses also included feedback about what they found helpful and not helpful about the videos, such as being able to watch the videos at their own pace but also being unable to ask questions. An additional survey focused around students' self-efficacy was also administered to three of these institutions. The results showed that students' chemistry self-efficacy (CSE) tended to increase over the term. Comparisons of students' CSE at the end of the term between the different institutions indicated that there may be a relation between self-efficacy and the structure of the course.

Project II centered around students' perceptions of active learning environments. A previously developed survey, the Assessing Student Engagement in Class Tool (ASPECT), did not function as expected in the active learning environments at Portland State University (PSU). Therefore, two modified ASPECT (mASPECT) versions were created to address these concerns, as well as to account for two different active learning environments: Deliberative Democracy (DD) activity days and clicker question days. Data collected with the mASPECT versions were analyzed using exploratory factor analysis (EFA) and cognitive student interviews. Data collected after a DD activity resulted in three factors of 'personal effort', 'value of environment', and 'classroom support', whereas data collected after a clicker question day resulted in three similar 'personal effort', 'value of environment', and 'classroom support' factors, in addition to a fourth 'social influence' factor. Although the factors discovered for each version were similar, they were not identical.

The goal of Project III was to develop a survey measure to assess multiple dimensions of student engagement (i.e., behavioral, cognitive, emotional, and social) in worksheet activities and begin to explore the effect of engagement between different groups and in relation to student outcomes. Overall, this project led to the development of the Activity Engagement Survey (AcES), which was informed by both qualitative student interviews and quantitative survey responses. Both qualitative and quantitative results provided evidence that students perceived the two dimensions of behavioral and cognitive engagement to be very similar when considering engagement in worksheet activities, which led to a combined behavioral/cognitive factor. In addition, social engagement themes were discovered throughout the student interviews and items to address social engagement were included in the final survey when students worked on the worksheet with others. Confirmatory factor analysis (CFA) was used to assess possible models of engagement with data collected with the AcES. The most appropriate model was found to be a bifactor model, which includes an overall engagement factor in addition to individual factors of behavioral/cognitive, emotional, and social engagement, with a negative method factor to account for negatively worded survey items. This type of model was found to be most appropriate for data collected from students who worked with others (BC-E-S AcES), as well as with data from students who worked alone and were not asked to respond to social engagement items (BC-E AcES). Data collected with the AcES was then used to explore different comparisons. Although validity evidence was insufficient to allow for comparisons between in-person and remote environments, validity evidence was found to support the comparison of student engagement between students that worked with others and those that worked alone in the remote environment. Results showed that students who worked in a group had higher overall and behavioral/cognitive engagement then those that worked on the activities individually. Additionally, the relation between students' engagement in an activity and their understanding of the material covered in the activity was assessed using multiple linear regression. Overall, only behavioral/cognitive engagement was found to be significantly related to students' understanding of the material.

Acknowledgments

There are so many people that have supported me through different parts of this journey and, although I have tried my best below, words are not quite enough to express how truly grateful I am to each and every person that has helped me reach this goal.

To my advisor, Dr. Jack Barbera, thank you for all that you have done to support me over the past few years. I greatly appreciate your willingness to let me get involved early, as well as your mentorship and encouragement, which have helped me to become a much more confident researcher. I also appreciate all the times you’ve been available whenever I’ve had questions regarding research, as well as when I’ve needed to talk about non-research things.

To my committee, thank you for all the feedback you’ve provided for this work. Dr. Erin Shortlidge – I greatly enjoyed working with you on the ASPECT project and the conversations we had regarding research and kitties. Dr. Gwen Shusterman – thank you for continuing to encourage me to see the ‘big picture’ of my work beyond simply models and numbers. Dr. Dean Atkinson – thank you for providing opportunities and mentorship for teaching higher level labs and for being one of the first people to welcome me into the department.

To all the instructors who allowed me to collect data in their courses over the years, thank you for taking time out of your class to allow for announcements, surveys, etc. Without your support, none of this research would have been possible.

To Dr. Eric Sheagley – in addition to allowing me to collect data in your class each quarter, thank you for also allowing me to take control of teaching half of your

general chemistry class. I learned a lot through doing my teaching practicum with you and greatly value the feedback and support you provided.

To all the people from the DBER community at PSU, thank you for always being willing to provide me with feedback, which has helped me improve during my time here. To the Barbera group – Dr. Regis Komperda, Dr. Katy Hosbein, Dr. Cory Hensen, Dr. Nicole James, Emryse Geye, Safaa El-Mansy, Elizabeth Vaughan, Charlie Bastyr, Gosia Cox, Kylee Brevick – thank you for all of your feedback on my projects, papers, and presentations, as well as providing me with good times and conversations. Thank you especially to those who have been secondary coders and/or researchers for parts of my project – Elizabeth Vaughan, Kylee Brevick, Safaa El-Mansy, and Eric Earle.

Outside of PSU, I want to thank all my family and friends who have supported me through this journey, especially those who I have been able to remain in contact with even though we now live on opposite ends of the country.

Of course, to the person who has provided me with the most support throughout this entire journey, my husband Tyler. You encouraged me when I was unsure of my decision to go back for my PhD and have supported my decision throughout. I never would have completed this program without your support. I have enjoyed all our adventures together (even in the pouring rain) and the many times of good coffee and good beer. Although working remotely was never expected, I have valued all the time I got to work from home with you over the last couple years and our daily walks are always some of the highlights of my day. I am excited to continue on this journey with you and am looking forward to our next life adventure together!

Last but not least, I have to acknowledge the very important roles my three cats have played in this journey. My grumpy ‘old’ lady, Daisy – who has been with me since the beginning of my grad school journey many, many years ago and who has provided me with a constant source of companionship (and meows). Cami – who has provided me with a lot of love and cuddles. And Mac (aka Mr. Flufferkins) – whose playfulness and puppy-dog nature has provided us with a constant source of entertainment, love, and duck watching over the last few years.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Lecturing has long been the traditional teaching format in higher education chemistry courses. In lecture classrooms, students are expected to listen and take notes while the instructor explains the concepts and material to the students. Around the 1990s, in response to a rise in the number of students leaving scientific disciplines (Krieger, 1990), many instructors began to further question how chemistry was being taught at colleges and universities. In 1992, Bodner noted that simply focusing on teaching and curriculum changes were not enough, stating that, "Teaching and learning are not synonymous; we can teach—and teach well—without having the students learn." Thus, to increase student learning in STEM, Bodner (1992) proposed that the learning environment needed to shift away from instructor-centered teaching to student-centered learning, such that students were encouraged to interact with the material, work with others, and be *active* learners.

In the last thirty years, many lesson plans, techniques, and strategies have been developed and/or adapted to encourage this shift from the "sage-on-stage" approach of teaching to more active student-centered learning. Although encouraging more active student learning can be as simple as providing a group worksheet, incorporating think-pair-share exercises, or facilitating whole-class discussions, other strategies, such as Process-Oriented Guided-Inquiry Learning (POGIL), Problem-Based Learning (PBL), and Peer-Led Team-Learning (PLTL), have also been used. Many studies in the literature have documented a positive impact on student learning when active learning strategies are incorporated into the classroom compared to traditional lecture methods (Freeman et al., 2014). Thus, as instructors want their students to succeed in their course, many have incorporated more active learning strategies into their classrooms. However, although higher student grades and lower withdrawal rates are generally seen in these active learning environments (Freeman et al., 2014), there are still students who withdraw from or fail these courses.

One variable which may impact student outcomes in these active learning environments is student engagement. Student engagement is often thought of as being composed of multiple dimensions, including behavioral, cognitive, and emotional components (Fredricks et al., 2004). Although these components have not been measured simultaneously in higher education active learning environments, each component individually has been found to positively influence student learning outcomes in general (Chi & Wylie, 2014; Skinner et al., 2017). As active learning environments are centered around increasing student learning through encouraging students to *engage* with the content through the incorporated activities (Freeman et al., 2014), differences in student engagement may influence the benefits of these activities for each individual student and explain some of the differences seen in student outcomes.

The research presented in this dissertation was comprised of multiple studies focused around investigating different components related to student engagement in active learning environments. These included evaluating student perceptions and self-efficacy, as well as cataloguing how students interacted with required pre-class materials. Additionally, a survey measure was developed to evaluate student engagement in worksheet-based active learning activities.

Statement of Problem

In 2012, a national report from the President's Council of Advisors in Science and Technology found that the economic climate over the next 10 years forecasted the need for one million more STEM graduates than were to be expected. The report noted that many of these additional graduates could be generated by increasing STEM retention from 40% to 50% in college and university. One of their recommendations for achieving this goal was to encourage STEM instructors to incorporate more evidence-based teaching practices, such as active learning strategies, into their classrooms (President's Council of Advisors on Sciences and Technology (PCAST), 2012). It has been found that students generally receive higher grades when they are more engaged in class and/or with a learning activity (Chi & Wylie, 2014; Handelsman et al., 2005; Skinner et al., 2017). Therefore, including active learning strategies into STEM courses can provide more opportunities and encouragement for students to actively engage with the material. However, the level of student engagement is not typically evaluated in active learning environments, especially for individual learning activities.

Courses that incorporate active learning strategies can include multiple components where students are expected to engage. For example, in flipped courses, students are expected to engage with pre-class materials, as well as the activities presented during face-to-face class time. In addition, students may engage with face-to-face activities differently, especially if a course incorporates multiple types of active learning strategies. Students' self-efficacy has been proposed as a precursor to engagement (Zumbrunn et al., 2014), which could be especially important in active learning environments where students are generally expected to complete tasks and problems. Additionally, how students engage with these different components of active learning courses may be linked to their perceptions of the environment (Cavanagh et al., 2018). Therefore, investigation of not only engagement, but other related aspects such as students' self-efficacy and perceptions, is important for better understanding how these environments affect student learning.

Purpose of Study

Simply including active learning in the classroom does not guarantee student success if students do not interact and engage with the active learning activities in a meaningful way. Although research has been conducted on the effects of active learning on overall student performance outcomes in higher education STEM courses (Freeman et al., 2014; Maldonado & Leontyev, 2018; Rahman & Lewis, 2019), little research has been conducted on student engagement during active learning activities in these environments. Therefore, the goal of this research was to begin to evaluate student engagement in active learning environments through measuring engagement and other contributing variables (Figure 1.1).

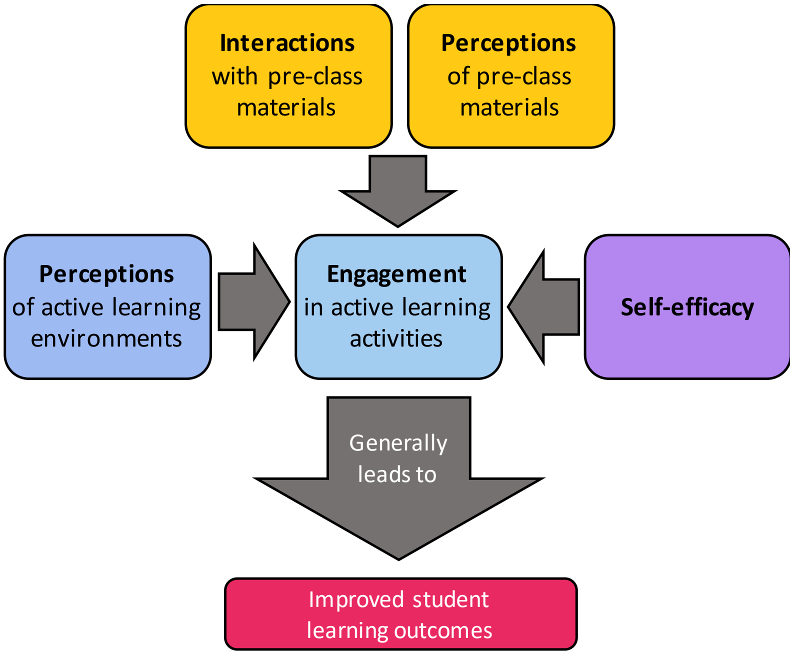

Interactions with pre-class materials
Perceptions of pre-class materials
Perceptions of active learning environments
Engagement in active learning activities
Self-efficacy
Generally leads to
Improved student learning outcomes

Figure 1.1. Overview of the different variables related to students' engagement in active learning courses that were investigated in this dissertation.

The goals of the first and second projects presented in this research were focused around evaluating students' self-efficacy and their perceptions of different active learning environments. The first project focused on quantitatively measuring academic and chemistry self-efficacy of general chemistry students in flipped active learning environments across multiple institutions. As students in flipped courses are expected to engage with pre-class materials before attending the class, their interactions with and perceptions of these materials were also investigated using both qualitative and quantitative methods. The second project focused on evaluating student perceptions in different active learning environments – clicker question days and Deliberative Democracy activities – that were incorporated into a principles of biology course. Data were collected with modified versions of an existing survey measure, the Assessing Student Engagement and Perceptions Tool (ASPECT) (Wiggins et al., 2017). Both qualitative and quantitative data was used to inform modifications to the measure, as well as to evaluate differences in how the survey functioned in the two environments.

The third project included in this dissertation focused on exploring student engagement in worksheet activities in general chemistry. Previous engagement surveys for higher education STEM tend to focus on measuring engagement at the course or module level, and generally only do so for a subset of the three engagement dimensions. Thus, to measure all three fundamental dimensions of engagement in individual active learning activities, a survey measure was first developed by adapting an extant measure of student engagement of middle- and high-school math and science students (Fredricks, Wang, et al., 2016; Wang et al., 2016). Both qualitative and quantitative evidence was gathered to inform this survey measure and ensure it functioned well with these activities and student population. Engagement differences between groups and the relation between engagement and student learning outcomes were investigated with the final survey version.

Research Questions

The overall goal of this dissertation was to investigate different components related to students' engagement in active learning activities and environments. This was completed through three projects, with each focused on a different facet of student engagement. The research questions guiding each project are presented below.

Project I: Students' Self-Efficacy and Interactions with Pre-Class Materials in Flipped Courses

RQ 1.1: What are students' interactions with and perceptions of required pre-class materials in flipped courses?

RQ 1.2: How does student self-efficacy change over the term in flipped courses?

RQ 1.3: How does student self-efficacy compare across flipped courses at different institutions?

Project II: Students' Perceptions of Different Active Learning Environments

RQ 2.1: How well does an existing measure of student perceptions of active learning activities function in different active learning environments?

RQ 2.2: What modifications can be made to an existing measure in order to measure student perceptions of these environments?

RQ 2.3: How well does this modified measure function in different active learning environments?

Project III: Students' Engagement in Worksheet Activities

RQ 3.1: How do students perceive engagement in worksheet activities?

RQ 3.2: What modifications can be made to an existing survey measure in order to measure student engagement in these activities?

RQ 3.3: How well does a modified measure of engagement function in this environment and student population?

RQ 3.4: How does engagement in these activities differ across groups?

RQ 3.5: How does engagement in these activities relate to students' understanding of the material?

Significance of Study

As the push to include more active learning strategies into STEM higher education courses continues, it is important to gather information about what makes these learning environments effective for student learning. There have been many studies that have focused on comparing class outcomes (e.g., grades, withdraw rates, etc.) between courses that incorporated active learning strategies and those that only included traditional lecture. However, with these outcome-based comparisons alone, it can be difficult to obtain information about why there still may be unsuccessful students even when active learning strategies are included. Therefore, there is a need to gather student-level information about elements that may influence student learning in these environments. This research begins to fill this gap by evaluating student engagement, perceptions, and self-efficacy in different active learning environments. Additionally, a large component of this research was the development of a measure of student engagement specifically for active learning environments within higher education STEM. This newly adapted measure can be used by instructors to obtain feedback in their own classes to assist them in designing activities or learning environments that encourage students to meaningfully engage and, subsequently, improve student learning outcomes for all students.

Limitations

There were many limitations to this research that applied to all three projects. Each project relied on self-report survey measures to collect quantitative data. These surveys were part of the research studies and not the courses themselves, which meant that students could not be required to complete them. Additionally, for both the survey measures and the qualitative interviews and focus groups, all data collected were from students who voluntarily consented to participate. Thus, the data may not reflect the perceptions of all students. Additionally, the development and adaptation of the survey measures for Projects II and III were completed at a single institution. Therefore, the results presented may not be generalizable to other institutions, environments, and/or student populations without further data collection and analysis.

As many who are reading this dissertation probably know, the year 2020 saw the emergence of the COVID-19 virus which led to a rapid and extended shift from in-person to remote learning beginning in March 2020 until September 2021. While data collection for Projects I and II was completed prior to March 2020, data were still being collected for Project III after this time. Therefore, the goals and focus of this project had to evolve to adapt to the changing situation. RQs 3.1 - 3.4 were informed by data collected during both in-person and remote learning and the results presented include some comparisons between the two (although this was not an original goal of the project). RQ 3.5 was informed solely by data collected when the course had returned to in-person learning in Fall of 2021.

Chapter 2: Review of the Literature

Introduction

The inclusion of active learning strategies in higher education STEM courses has become more prevalent over the last decade. One of the earliest compilations of the effects of including active learning strategies on student outcomes was completed by Freeman et al. in 2014. In their meta-analysis of 225 studies, they found that the adoption of active learning activities in STEM classrooms increased student performance outcomes (i.e., higher exam grades, lower withdrawal rates, etc.) when compared to traditional classes (Freeman et al., 2014). This same result was also found in later meta-analyses that focused only on studies that included courses specifically within chemistry (Maldonado & Leontyev, 2018; Rahman & Lewis, 2019). However, even though a positive influence was generally seen when active learning strategies were incorporated, some of the studies in these meta-analyses showed little to no change in student performance outcomes when compared to a traditional class (Freeman et al., 2014; Maldonado & Leontyev, 2018; Rahman & Lewis, 2019). As the aim of active learning is to engage students with the material, gathering information about different active learning environments and how students engage with them could provide insight into why these differences in student performance outcomes exist.

Active Learning Strategies

Active learning strategies are usually described as a strategy that encourages students do something besides simply listen to the instructor and take notes during class (Handelsman et al., 2007). The definition of active learning used by Freeman et al. (2014) in their meta-analysis stated that:

> *Active learning engages students in the process of learning through activities and/or discussion in class, as opposed to passively listening to an expert. It emphasizes higher-order thinking and often involves group work.*

These definitions and descriptions of active learning only require that an active learning strategy assists in engaging students in their learning. Therefore, based on the lack of formal requirements, there exist a wide variety of different strategies that are considered active learning, such as Process-Oriented Guided-Inquiry Learning (POGIL), Peer-Led Team-Learning (PLTL), think-pair-share strategies, and more. Descriptions of the different active learning strategies evaluated in this dissertation are given below.

Clickers

The use of a classroom response system (i.e., clickers) has become a fairly popular active learning strategy in higher education classrooms. In chemistry, clickers are most often incorporated into an introductory-level course with a large number of students (Gibbons et al., 2017). Although there are many ways to incorporate clickers, their basic function is to provide a platform where every student in the class can quickly and easily respond to a question posed by the instructor. This allows the instructor to obtain real-time feedback on students' understanding of the question. Additionally, clickers can be used to "break-up" a lecture and allow students a period of time to interact with the material (Caldwell, 2007), as well as decrease students' lapses in attention (Bunce et al., 2010).

Although clickers can be used to obtain individual student feedback, they can also be used along with group discussion and peer learning. Group discussion can occur either before answering the question for the first time or after the students have already answered the question once on their own. Peer discussion with the use of clickers has been found to be more beneficial to students' understanding of the concepts than only including individual responses (Brooks & Koretsky, 2011). Additionally, one study found that students preferred team-based clicker questions more than individual clicker questions (Pearson, 2017).

POGIL

Process-Oriented Guided-Inquiry Learning (POGIL) is a fairly common and flexible active learning strategy. A national survey found that about 11% of postsecondary chemistry instructors of both lower- and upper-level courses regularly (i.e., at least one time per week) incorporated POGIL (Raker et al., 2020). The implementation of POGIL can vary between classes; however, there are generally three common features that are included. These include small groups, student exploration of the material through an activity, and facilitation by the instructor (Moog & Spencer, 2008). Student groups are generally comprised of about 3 – 4 students per group and can include specific student roles (i.e., reader, speaker, recorder, etc.). The activities that are used for POGIL are created specifically to encourage students to explore the concepts and build upon initial information through guiding questions. Additional questions to encourage students to apply the concepts to new information are also commonly included (Moog & Spencer, 2008).

Deliberative Democracy

Deliberative Democracy (DD) is a type of active learning strategy that encourages student deliberation and consensus-building skills. Like POGIL, the exact way that DD is implemented in a classroom can vary between courses. The topics used for DD modules usually focus around a policy or real-world issues (Komperda, Barbera, et al., 2018). For DD modules that span multiple days, students are first introduced to the topic on the first day of the activity, where they are then given time in their groups (usually < 10 students) to assign roles, discuss the topic, and determine what additional resources and information they will need to find before the next activity day. On the final day of the DD module, students bring in the resources they have gathered or read some instructor-provided articles and then discuss the topic to reach a consensus on the issue presented. Facilitation of group discussions, either by the instructor and/or teaching assistants, is used to help students consider additional aspects or ideas they may not have otherwise considered (Komperda, Barbera, et al., 2018).

Flipped Classrooms

One barrier to including more active learning strategies in the classroom is the time it takes to complete many of these types of activities, which inherently reduces the amount of time the instructor has to present new material. “Flipping” the classroom can provide this extra time to incorporate active learning strategies during class. Most flipped classes include two components; pre-class materials (PCMs) that are provided to students before class and face-to-face (F2F) time where students apply the information from the PCMs through participating in active learning activities (Bergmann & Sams, 2012). As these are the two main requirements of a flipped class, the actual structure of the classroom can vary greatly between different flipped classrooms, even within studies conducted in higher education chemistry courses.

PCMs in most flipped classrooms take the form of instructor-made videos (e.g., Casselman et al., 2019; Mooring et al., 2016; Smith, 2013), although other online resources have also been used, such as Khan Academy videos (Liu et al., 2018) and online modules (Gregorius, 2017). Additionally, some flipped classrooms incorporate multiple types of PCMs, which can include videos, animations, podcasts, etc. (e.g., Amaral et al., 2013; Bokosmaty et al., 2019). However, PCMs can also consist of non-online components, such as textbook readings (e.g., Lenczewski, 2016). Incentives to complete the PCMs can also vary, with some flipped courses using note-checking (Hibbard et al., 2016; Parsons, 2019; Shattuck, 2016) or pre-class assignments (Donnelly & Hernández, 2018; Lenczewski, 2016). Additionally, many flipped courses use in-class (e.g., Christiansen, 2014; Fitzgerald & Li, 2015; Woodward & Reid, 2019), or out-of-class quizzes (e.g., Amaral et al., 2013; Bokosmaty et al., 2019; Fautch, 2015) to incentivize students to complete the PCMs.

The F2F environment of a flipped classroom can also vary significantly. Some courses use the F2F time to incorporate known active learning strategies such as PLTL (Liu et al., 2018; Mooring et al., 2016; Rein & Brookes, 2015; Robert et al., 2016) or POGIL (Canelas et al., 2017; Hibbard et al., 2016). Other flipped classrooms use a combination of different active learning strategies, such as peer instruction, problem-based learning, and think-pair-share exercises (Christiansen et al., 2017; Flynn, 2015), while others focus on case studies (Rein & Brookes, 2015) or class discussions (Bernard et al., 2017; Bokosmaty et al., 2019; Ealy, 2013; Ryan & Reid, 2015; Smith, 2013). Even when a specific type of active learning strategy is not mentioned, many studies include a description centered around some type of student groupwork (e.g., Christiansen et al., 2016; Donnelly & Hernández, 2018; Rau et al., 2017; Shattuck, 2016; Weaver & Sturtevant, 2015).

Active Learning Environments

Even when the same active learning strategy is incorporated into different classrooms, the active learning environments that are produced are not necessarily identical. Although each may contain the critical components of the specific strategy, the actual implementation may differ significantly. This can make it difficult to compare the results from different studies, especially since details about implementation are not often reported (Seery, 2015b). Additionally, it has been noted that the adaptation of active learning strategies, while necessary to account for differences between classrooms, can bring the validity of the results into question if these details are not accounted for (Stains & Vickrey, 2017). Variations in implementation may also play a role in the differences seen in student outcomes within studies that incorporated the same type of active learning strategy (Rahman & Lewis, 2019).

Differences in the implementation of active learning strategies can be documented through observational protocols, which are used to categorize the F2F environment of the different classrooms. One such observational protocol created for higher education STEM courses is the Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2013). The COPUS can be used by researchers to record what the instructor and the students are doing for each 2-minute time interval during class time. There are 13 possible student codes including "listening" (L), "working in groups on worksheet activity" (WG), and "student asks question" (SQ), for example. Instructor codes include codes such as "lecturing" (Lec), "asking a clicker question" (CQ), and “moving through the class” (MG), with a total of 12 possible codes (Smith et al., 2013). As the observation is broken into 2-minute time intervals, there can be multiple codes for each interval. A code is recorded whenever an action occurs during an interval, even if it does not take up the entire 2-minutes.

COPUS time-lines can provide fine-grain details about what is occurring in a classroom. The percentages of each code that occurs during class time have also been used to categorize classrooms into different instructional styles, including lecturing, Socratic, peer instruction, and collaborative learning (Lund et al., 2015). In addition, COPUS profiles have been created using cluster analysis of the prevalence of four student codes and four instructor codes from 709 higher education STEM courses. The clusters that were found grouped into three main profiles; didactic, in which “80% or more of class time consists of lecturing”, interactive lecture, where instructors “supplement lecture with more student-centered strategies”, and student-centered (Stains & Harshman; Stains et al., 2018). These profiles and types of instructional styles can assist in comparing classrooms at a larger-grain size.

Assessing Student Outcomes

Student performance related outcomes in active learning studies are typically reported using grades for an exam or the overall course. Exam performance has been assessed through instructor-written exams and concept inventories, as well as ACS exams (Freeman et al., 2014; Maldonado & Leontyev, 2018; Rahman & Lewis, 2019). Studies that used instructor-written exams to assess performance tended to report less of an improvement in active learning classes compared to traditional classes than those that used concept inventories (Freeman et al., 2014). As the difficulty level and/or type of questions on instructor-written exams are usually not reported in studies, these types of exams could include different amounts of lower-level recall questions versus questions that assess higher-level understanding. If lower-level recall questions make up a significant portion of an exam, scores from active learning and traditional environments might not be significantly different, as deeper student understanding is not assessed (Chi & Wylie, 2014). Additionally, while the use of ACS exams provides a broad range of item difficulties, the amount of questions addressing higher-level understanding versus lower-level recall is also unknown and their use for evaluating performance outcomes does not always reveal an improvement in exam scores in active learning classrooms (Seery, 2015b). Thus, it is unknown if the studies that found no or minimal difference in student exam grades between the different environments may have been influenced by the presence of exam questions that inadequately assessed student understanding of the material.

Course grades may also not adequately reflect student learning differences between active learning and traditional classrooms. Retaining students is seen as a benefit of an active learning environment; however, Seery (2015a) has suggested that when students who would have withdrawn in a traditional class remain in a corresponding active learning class, the overall class performance at the end of the term might be lower than if these struggling students had withdrawn. Therefore, this potential difference in student population between traditional and active learning classrooms could make it difficult to meaningfully compare outcomes from the two environments. Additionally, simply focusing on student performance outcomes to evaluate active learning environments disregards potentially valuable student-level data (Seery, 2015b). As students' engagement in class has been shown to be positively related to increased student academic outcomes (Chi & Wylie, 2014; Handelsman et al., 2005; Skinner et al., 2017), focusing instead on how individual students interact and engage with the environment may provide valuable insight into students' learning that is overlooked when comparing exam and course grades.

Definition of Student Engagement

There are two main perspectives of student engagement; behavioral and psychological (Kahu, 2013). The behavioral perspective of engagement focuses solely on student behaviors, where students that spend more time or effort on educational activities have higher engagement (Kahu, 2013). This perspective implies that increasing the amount of time students spend on purposeful activities automatically increases their engagement. The psychological perspective frames engagement as a 'meta-construct' that includes multiple dimensions that are all related and contribute to a student's overall engagement (Kahu, 2013). The three major dimensions, or constructs, are behavioral, emotional, and cognitive, although other dimensions have also been proposed (Kahu, 2013). One major advantage of the psychological perspective is that it does not view engagement simply as behaviors, but incorporates the potentially unseen cognitive and affective dimensions. However, as the dimensions overlap with each other, there can be a lack of distinction between the different constructs. For example, student effort can potentially be categorized as behavioral (e.g., doing multiple examples for practice) or cognitive (e.g., trying hard to make connections between concepts). This limitation can be addressed by using clear operational definitions of the behavioral, emotional, and cognitive dimensions of student engagement (Fredricks, Filsecker, et al., 2016).

Behavioral, Emotional, and Cognitive Engagement

Behavioral student engagement primarily focuses on students' positive conduct and absence of disruptive behaviors in the classroom. According to Fredricks et al. (2004), behavioral engagement can also include "effort, persistence, concentration, attention, asking questions, and contributing to class discussion." Although these behaviors mostly center around the classroom, some definitions of behavioral engagement also include student participation in extra-curricular school-related activities outside of class time (Fredricks et al., 2004).

Emotional student engagement focuses on students' affective reactions to interactions they have in the classroom. These interactions can include how students interact with their peers, the instructor, the class material, or in-class activity. Additionally, many different emotions are usually included; such as interest, boredom, value, students' attitudes towards school, etc. (Fredricks et al., 2004; Sinatra et al., 2015). Although emotional engagement encompasses many different emotions, this broad focus separates emotional engagement from other related constructs that are more narrowly focused on explaining *why* a student displays certain emotions (Fredricks et al., 2004). According to Kahu (2013), emotional constructs (e.g., interest, value, attitude, etc.) represent a "cluster of factors that influence student engagement…whereas the outcome *is* student engagement". Therefore, even though measuring students' overall emotional state may not determine the source of the emotions, it still provides information about the student's emotional engagement at that time.

Cognitive engagement can broadly be defined as students' psychological investment in their learning; however, other definitions have also been used, such as the students' ability for self-regulation or use of deep-level learning strategies (Fredricks et al., 2004). Fredricks et al. (2004) states that the psychological investment perspective of cognitive engagement can include concepts such as "flexibility in problem solving, preference for hard work, and positive coping in the face of failure", in addition to exerting mental effort in order to reach a deeper understanding of concepts (Sinatra et al., 2015). Generally, literature focused on student engagement tends to stress student investment, whereas literature centered around learning and instruction tends to emphasize self-regulation. Similar to the relation between specific emotional constructs and emotional engagement, cognitive engagement definitions also overlap with other constructs, such as intrinsic motivation (Fredricks et al., 2004; Sinatra et al., 2015). However, as was mentioned previously, although other constructs (e.g., motivation) may influence a students' cognitive engagement, measuring engagement itself provides information about engagement as the outcome of interest (Kahu, 2013).

Other Dimensions of Engagement

In addition to the three foundational dimensions, other dimensions have been proposed as part of the engagement 'meta-construct'. Some of these additional dimensions are combinations or re-wordings of behavioral, emotional, and cognitive engagement. For instance, Appleton et al. (2006) describe an 'academic engagement' dimension as "time on task, credit hours toward graduation, [and] homework completion" and a 'psychological engagement' dimension (different from the psychological *perspective* of engagement) as "belonging, identification with school, [and] school membership." These descriptions include components of behavioral and emotional engagement targeted towards a specific environment. However, there are two proposed dimensions of engagement that are not simply rewordings of behavioral, emotional, and cognitive engagement: agentic engagement and social engagement.

Agentic engagement is defined by Reeve and Tseng (2011) as “students’ constructive contribution into the flow of the instruction they receive.” This includes offering input, asking a question, and seeking out personal relevance to the lesson. The purpose of this dimension is to include the concept of agency into student engagement, whereas the three foundational dimensions of engagement are based on measuring student reactions to the structure of the class. Although agentic engagement has been shown to be an independent and related construct to engagement (Reeve & Tseng, 2011), there have been concerns about including it as a dimension of engagement until additional research is completed to provide more evidence on its relationship with engagement overall (Sinatra et al., 2015). Another proposed dimension of engagement is social engagement. Social engagement is defined by Wang et al. (2016) as the “quality of social interactions with peers and adults, as well as the willingness to invest in the formation and maintenance of relationships while learning.” Although this dimension has been studied in middle- and high-school math and science classes (Fredricks, Wang, et al., 2016), it may also be an important dimension in higher education active learning classrooms since collaborative learning is greatly emphasized in these environments (Brint et al., 2008; Chi & Wylie, 2014).

Variables that Influence Engagement

Students’ engagement can be influenced by many different variables, which may occur from both inside and outside the classroom environment. As active learning strategies require students to be involved in their learning, there are some variables that may uniquely influence their engagement in these environments. One such variable is students' self-efficacy, as students in active learning environments are generally expected to complete certain problems or tasks. Additionally, as active learning is not universally adopted in STEM courses, students' perceptions and buy-in to the environment may also impact their engagement in the activities.

Self-Efficacy

Self-efficacy is generally defined as a person's perception of their capability to learn or perform a certain task (Bandura, 1997). The self-system model of classroom support suggests precursors to engagement that are related to a student's self, such as their self-efficacy (Zumbrunn et al., 2014). Additionally, students' with high academic self-efficacy have been shown to have higher cognitive engagement related to using meaningful learning strategies (Walker et al., 2006). Therefore, measuring self-efficacy could provide important information about students' engagement.

Self-efficacy can be measured at different task levels; such as a students' general academic self-efficacy or their self-efficacy for a specific course or task. Within chemistry, chemistry-specific self-efficacy has commonly been measured with slightly modified versions of the Chemistry Attitude and Experience Questionnaire (CAEQ) (Chase et al., 2013; Villafañe et al., 2014; Vishnumolakala et al., 2017) or the College Chemistry Self-Efficacy Scale (CCSS) (Ferrell & Barbera, 2015; Graham et al., 2019; Ramnarain & Ramaila, 2018). These measures include items that are more narrowly focused on the chemistry course and chemistry-specific tasks versus more general academic tasks. Overall, within chemistry, self-efficacy has been found to generally increase over a course term (Ferrell & Barbera, 2015; Graham et al., 2019; Villafañe et al., 2014; Vishnumolakala et al., 2017), although it has been found that this positive change could be dependent on demographic group (Villafañe et al., 2014).

Perceptions and Buy-in

Many university students expect to be spending more class time passively and individually learning compared to actively participating in collaborative learning (Brown et al., 2017). Since active learning strategies can only provide students with the opportunities to engage with the material, it is essential that students' make the decision to interact with and buy-in to the learning environment (Cavanagh et al., 2016). Higher student buy-in to an active learning environment has been linked to higher engagement (Cavanagh et al., 2016). Additionally, student engagement has also been found to be influenced by student perceptions of the environment, such as their trust in the instructor (Cavanagh et al., 2018). These perceptions can vary based on instructor implementation, but are also found to vary from student to student even within the same class and instructor (Cavanagh et al., 2016; Cavanagh et al., 2018). Thus, measuring student perceptions of the learning environment could provide valuable information about student engagement in these environments.

A recent survey for measuring student perceptions of active learning environments in higher education STEM courses is the Assessing Student Perspective of Engagement in Class Tool (ASPECT) (Wiggins et al., 2017). This measure was developed in a large-format biology course to assess student perceptions of their value of the group activity, their personal effort, and the instructor contribution for two types of active learning strategies. These active learning environments included a short-activity day, which incorporated a series of clicker questions during class, and a long-activity day, which consisted of groupwork on a worksheet with some follow-up clicker questions. Overall, their results found that student perceptions of the value of the group activity and instructor contribution were significantly less for the long-activity day compared to the short-activity day, while there was no significant difference found for student perceptions of their personal effort between the two activities (Wiggins et al., 2017).

Previous Student Engagement Measures

Student engagement is often measured through observational protocols or self-report surveys. Although a recent study by McNeal et al. (2020) found that the use of biosensors could be used to measure student engagement via galvanic skin response, these type of direct physical measurements have many limitations, including individual variations in response, as well as practical limitations related to collecting data for every student in a large class size (McNeal et al., 2020). Therefore, observational and self-report student engagement measures may be adaptable to more environments due to their ease of use and the standardization for collecting evidence of data validity and reliability from such measures.

Observational Measures

Observational protocols are generally used as a measure of behavioral engagement. For example, the Behavioral Engagement Related to Instruction (BERI) protocol (Lane & Harris, 2015), categorizes student behaviors into 'engaged' or 'disengaged' during blocks of class time. Although the BERI protocol can give a sense of the behavioral engagement of a class, one major drawback is that each observer can only observe a subset of students (approx. 10) in the classroom. Although the BERI developers found that the overall trends of engaged behaviors were similar, they found that different groups of students showed different amounts of overall engagement. Because of this, the developers noted that this instrument is most useful for determining if the class as a whole showed general ‘off-task’ or ‘on-task’ behaviors. Similar observational protocols have also been developed for other classrooms, such as the Student Engagement Observation (SEO) (Harris & Cox, 2003) instrument that was developed for engineering courses. This instrument also categorizes the general behavior of students as either ‘desirable’ or ‘undesirable’.

Cognitive engagement measures have also been created using observational protocols. Specifically, the ICAP (interactive-constructive-active-passive) framework (Chi & Wylie, 2014), which categorizes different modes (i.e., levels) of cognitive engagement from ‘passive’ to ‘interactive’. These modes are mapped onto things students do in the class such that observations can be used to determine students’ cognitive engagement. For example, if the students are tasked with reading a text, they could *passively* read without doing anything else, *actively* underline or copy notes, *constructively* take notes in their own words, or *interactively* dialogue about the text with a partner. One major concern with this framework is that it measures modes of cognitive engagement through assessing what students do. According to Chi and Wylie (2014), “there can also be a misalignment between the overt display of behavior and the covert processes”, suggesting that students could display low cognitive engagement (e.g., copying text) while actually being highly cognitively engaged (e.g., thinking about connections to previous concepts) or vice versa.

In addition to a possible misalignment between overt behavior and covert (i.e., cognitive) processes, along with only being able to observe a subset of students, there are other concerns with using observations to measure engagement in the classroom. Observations are time-consuming to collect and the reliability is dependent on the training of the observer(s) (Fredricks & McColskey, 2012). Additionally, observations are not recommended for measuring the emotional and cognitive dimensions of engagement (Appleton et al., 2006). Thus, if all three foundational dimensions of engagement are to be measured for every student, self-report surveys provide a more viable option. Self-report surveys are fairly easy to administer and do not take a lot of time or training for the person collecting the data (Fredricks & McColskey, 2012). In addition, data can be collected for every student instead of only a subset of students. Surveys can also be used to ask students about their perceptions, feelings, and thoughts, which can allow emotional and cognitive engagement to be measured in addition to behavioral engagement (Appleton et al., 2006; Fredricks et al., 2004; Fredricks & McColskey, 2012). Self-report surveys also have their limitations, which can include students not answering the items honestly or not interpreting items in the way the researchers intended them to be interpreted. However, these concerns can be addressed through data collection techniques (e.g., by making the surveys anonymous or using an outside researcher to collect data) and by conducting student response process interviews during development and evaluation of the survey. A number of self-report surveys of engagement exist in the literature and vary by the aspect(s) of engagement they are designed to measure.

Self-Report Survey Measures

One of the more widely known engagement survey measures is the National Survey of Student Engagement (NSSE) (National Survey of Student Engagement, 2018), which focuses on the behavioral aspect of engagement. The most recent version of this survey includes items that span four broad engagement-related themes; academic challenge, learning with peers, experiences with faculty, and campus environment. Another survey measure called the Student College Engagement Questionnaire (SCEQ) (Handelsman et al., 2005) also focuses on broad themes of engagement, including skills, participation/interaction, emotional, and performance. One concern with both of these surveys is that they use themes related to the broad construct of engagement instead of incorporating the actual theoretical dimensions of engagement (i.e., behavioral, cognitive, emotional) and there has been some concern that the broad themes included in these surveys may not adequately capture the theoretical complexity of the engagement construct (Campbell & Cabrera, 2011; Kahu, 2013).

One survey measure that uses the theoretical dimensions of engagement is the University Student Engagement Instrument (USEI) (Maroco et al., 2016). Although this survey does use the theoretical dimensions of engagement, it suffers from another concern, also applicable to the NSSE and SCEQ, in that it was created to assess engagement at the institution-level. As a consequence, these measures may not be generalizable across disciplines. In a study conducted by Brint et al. (2008), two different 'cultures of engagement' were discovered within the University of California system. The first type included students who were more focused on "individual assertion, classroom participation, and interest in ideas" and represented engaged students within the arts, humanities, and social sciences. The second 'culture of engagement' was more representative of STEM students and was based on "working toward quantitative competencies through individual study and collaborative effort." This difference between the two is thought to be a contributing factor to the lower NSSE scores found for math and science students compared to students in other disciplines (Ahlfeldt et al., 2005). For example, a type of item that a math or science student may score lower on is one that asks about the amount of assigned readings and writings they complete for class (Kahu, 2013). Thus, engagement measures for science students should be relevant to science courses and focus on collaboration and problem-solving (Brint et al., 2008; Sinatra et al., 2015).

One engagement measure that was created specifically for higher education science classes was developed to assess the use of clickers (i.e., classroom response system) in a freshman chemistry class (Aceti, 2017). This survey measure contains seven engagement items related to “the relationship between the clicker and increased interest in course material,” as well as “participation during lecture and interaction with other students and with the professor.” However, this measure again relies heavily on general engagement-related themes instead of the theoretical dimensions of engagement. For engagement measures that have been developed in higher education science classrooms and measure the theoretical dimensions, most focus on only one or two dimensions instead of all three. For example, a survey measure created by Gasiewski et al. (2012) to assess engagement in introductory STEM courses focuses only on the behavioral dimension of engagement, whereas a study done by Seery (2015a) on student engagement in a flipped classroom used a scale that only assesses students’ cognitive engagement (Rotgans & Schmidt, 2011). A two-dimensional instrument for STEM courses that assesses behavioral and emotional engagement was developed by Skinner et al. (2017). This survey measure includes behavioral items related to “students’ effort and active participation in coursework,” as well as negatively worded items to capture behavioral disaffection such as “lack of attention and effort.” Similarly, emotional items are separated into positive items related to “students’ motivated emotions” and negative items related to “negative emotions about working on science.” Although these measures are based on the theoretical dimensions of engagement and were created for higher education STEM classes, they only assess one or two dimensions of engagement. More recently, a measure was developed that incorporated all three theoretical dimensions of engagement (Smith & Alonso, 2020). However, the measure focuses solely on the chemistry laboratory setting instead of the classroom setting and, as such, the constructs are related to students’ interactions and perceptions of data, lab procedures, and data collection, which are not relevant to the chemistry classroom.

Limitations of Existing Student Engagement Measures

Although measures of student engagement in higher education science classrooms exist in the literature, they tend to assess engagement in the class as a whole instead of focusing on specific activities in the classroom. Additionally, most focus on general engagement-related themes or only include one or two of the theoretical dimensions of engagement. Creating a measure of engagement for active learning environments that includes all three foundational dimensions and is targeted towards the individual activity level would allow instructors and education researchers to gauge which activities are more engaging and for which students. A potential measure that could be adapted for this purpose is one developed by Wang et al. (2016) to assess student engagement of middle- and high-school math and science students. The instrument was developed through qualitative interviews with students where they were asked to describe what they were doing, thinking, and feeling while engaged in class (Fredricks, Wang, et al., 2016). The responses from these interviews were then used to inform the creation of items for the behavioral, cognitive, and emotional dimensions. Additionally, a social dimension was found to be present throughout all responses and was included in the final survey measure (Wang et al., 2016).

Measurement

Self-report surveys generally consist of a series of items that the students directly respond to. These items are related to an unobserved variable, called a latent factor or construct (e.g., engagement), and can be used to determine the students' 'score' on that set of items (i.e., factor). Thus, as self-report surveys are not a direct measure of engagement, it is important that evidence is gathered to provide confidence that the survey measures what it is designed to measure. This is done by collecting evidence of validity and reliability of the data generated by the measure. Validity evidence provides confidence that the instrument measures what it is intended to measure, while evidence of reliability provides information about the consistency of the data.

Validity

Validity refers to the extent that the data produced by a measure actually measures the intended construct (Arjoon et al., 2013). There are multiple sources that provide evidence of validity including test content, response process, internal structure, association with other variables, and consequential validity (Arjoon et al., 2013; Knekta et al., 2019; Wren & Barbera, 2013). The amount and type of validity evidence collected is influenced by the goals of the study as well as the amount and type of data validity evidence that has previously been collected for a measure. For measures that were developed in a similar environment, minimal amounts of validity evidence need to be gathered to determine if the instrument functions similarly; however, additional evidence that students are interpreting the items correctly should be collected for measures that are adapted to a new environment or undergo modifications (Knekta et al., 2019). If a new measure is being created from theory, then evidence also needs to be collected to provide evidence that the measure represents the theoretical construct. Additionally, evidence that a measure similarly assesses different groups needs to be collected before comparisons between those groups can be made.

Test Content Validity

Test content validity is concerned with the relation between the survey measure and the construct it is trying to measure. There are two types of validity evidence within test content validity; content validity and face validity (Wren & Barbera, 2013). Both types of evidence are used to provide confidence that the measure aligns with the theoretical definition of the construct. Content validity specifically is focused on assessing the extent to which the measure is a representation of the domain of interest (Arjoon et al., 2013). This is usually evaluated through interviews or open-ended survey responses with subject-matter experts to ensure that the items are representative of the theoretical definition of the construct (Arjoon et al., 2013; Crocker & Algina, 1986). On the other hand, face validity assesses if the measure is *perceived* by the students to measure the construct of interest (Crocker & Algina, 1986). This is also usually assessed through interviews or open-ended responses, this time with student participants (Crocker & Algina, 1986). If sufficient evidence of both types of test content validity are found during development of a measure, then there can be confidence in the identity of the latent construct that is being measured.

Response Process Validity

Evidence of response process validity can provide confidence that students are interpreting the items in the way that they were intended (Arjoon et al., 2013; Knekta et al., 2019). Cognitive interviews, where the student reads through the items and describes the reasoning behind their responses, is a common method to collect evidence of response process validity. During these interviews, the researcher focuses on how the students from the target population read and respond to the items to determine if any of the items are unclear or irrelevant to the students (Arjoon et al., 2013). Additionally, response process data can be collected through short-answer responses on a survey which allows for a larger sample size. However, although more data can be collected through a survey, this format prevents the researcher from asking any follow-up or clarifying questions. Collecting response process validity is not only important when a measure is being developed, but also when a measure is being adapted or modified to a new environment, as students may not interpret items identically in different environments.

Internal Structure Validity

Validity evidence based on internal structure is focused on how the relation between the items and the construct match the expected hypothetical structure (Arjoon et al., 2013; Knekta et al., 2019). As the latent construct is not measured directly, evidence of structural validity can provide confidence that the indicator items (i.e., survey items) that are being directly measured are related to the latent construct in the expected way. This is modeled through a series of regressions, where the item scores are predicted by the factor score (Brown, 2015). Factor loadings, which are the slopes of the regressions, represent the strength of the relation between the item and the factor. These can be standardized to a value between 0 and 1, such that the square of the standardized value indicates the amount of variance of the item that is explained by the construct. Remaining variance not accounted for by the construct, called unique variance, is usually assumed to be measurement error (Brown, 2015). The evaluation of internal structure through modeling the relations between items and factors is called factor analysis.

There are two types of factor analyses; exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA analyzes the relations between items and factors without a preconceived structure. Thus, it is more 'exploratory' in nature; however, EFA can also be used in a confirmatory way. As no initial structure is assumed, EFA does not produce useful information about the fit of the model. However, the benefit of EFA is that it allows for items to load onto all factors and a pattern of relations between certain items and factors to be discovered. Additionally, the pattern of relations between items and factors for different factor structures (i.e., different number of factors) can be easily explored with EFA. Alternatively, CFA requires an *a priori* model of the relation between items and factors. The data collected with the measure is fit to this model and fit statistics (i.e., indices) are generated. Fit indices that fall within appropriate cutoffs provide validity evidence for the proposed structure. Although usually used for confirmatory purposes, CFA can also be used in an exploratory manner by evaluating modification indices (Brown, 2015; Knekta et al., 2019). Modification indices represent an expected change in the model fit if certain modifications are made; however, care should be taken whenever any modification are made to ensure that changes are also supported by theory. One common modification to factor models is the addition of an item-item error correlation. These types of correlations assume that there is another source of item variance that is unaccounted for by the factors (Brown, 2015). This often occurs with similar item wording (Brown, 2015); however, it can also occur if there is another variable that is unaccounted for by the current model (Hermida, 2015).

Association with Other Variables

The association of the measure with other variables can provide additional evidence of validity (Wren & Barbera, 2013). If there is a variable that is theoretically predicted to be related to the construct of interest, then measuring the correlation between the two measures can be used to assess convergent validity. Additionally, evaluating the relation of the construct with a variable that is predicted to not be associated with the construct assesses discriminant validity (Hancock et al., 2010). The association of the measure with other variables can also be evaluated through concurrent and predictive validity. Concurrent validity assesses the degree to which the measure predicts performance on a theoretically related assessment given at the same time, while predictive validity assesses performance on a theoretically related future assessment (Hancock et al., 2010).

Consequential Validity

When a measure is used to make comparisons between groups, consequential validity through measurement invariance must first be established. Obtaining evidence for the various levels of measurement invariance can provide confidence that the different groups are responding to the measure in a similar way and that comparisons represent true differences between the groups. Before measurement invariance can be evaluated, validity evidence of the internal structure for each comparison group must be established. If acceptable data-model fit is not found at this stage, then group comparisons cannot be made. However, if these baseline models display acceptable data-model fit, then additional evidence of measurement invariance can be established to make comparisons to varying degrees. Establishing a certain level of measurement invariance generally requires evidence for each step while moving to greater constraints (Rocabado et al., 2020).

The first step to evaluate measurement invariance is to establish configural invariance. This level of measurement invariance compares unconstrained factor structures and provides evidence that the factor structures for both groups are similar. If configural invariance is established, there can be confidence that the items and factor structures are behaving similarly in all groups and the next level of metric invariance can be tested. Evidence of metric invariance is found through constraining the factor loadings to be equal across groups. If the data-model fit of this model is acceptable, then there can be confidence that the meaning of the factor and items are similar across groups. Although group comparisons are not supported with configural or metric invariance, evidence of invariance at these levels provides support to test for scalar invariance. This next step is accomplished by additionally setting the item intercepts equal across groups. If good data-model fit is discovered, then group comparisons using latent means is supported and represent true differences between the groups. The final level of measurement invariance is conservative invariance. This includes constraining the error of all items between groups in addition to all the previous constraints and is the highest level of measurement invariance that can be established. Support for conservative invariance provides confidence that observed scale scores (i.e., scale averages) can be compared between groups (Rocabado et al., 2020).

Reliability

Although evidence for validity may be found, finding evidence for reliability is also important before a measure can be used in a specific environment. Reliability is concerned with the consistency of the data and can be evaluated for temporal stability over time or internal consistency within a measure (Arjoon et al., 2013; Hancock et al., 2010; Komperda, Pentecost, et al., 2018). Temporal stability of a measure is evaluated through test-retest reliability. Evidence of test-retest reliability is found by administering the measure at multiple times and calculating the correlation between the scores at both time points. If the measure is consistent over time, then the correlation between the time points will be high (i.e., close to 1) (Arjoon et al., 2013; Hancock et al., 2010; Komperda, Pentecost, et al., 2018). This type of reliability evidence is most relevant when the measuring constructs that should be stable over time. Alternatively, for constructs that are expected to change over time or based on the environment (e.g., engagement), evidence of internal consistency can be used to assesses the reliability of the data by evaluating the consistency of students' responses to items measuring the same construct. As of 2018, the most commonly reported coefficient of reliability in the chemistry education literature was Cronbach's alpha; however, as noted by Komperda et al. (2018), alpha should only be used for models that have equal item loadings (i.e., tau-equivalent models). Since most measures used in education research are not designed to meet this requirement, McDonald's omega (McDonald, 1999) is a better choice as it allows for unequal item loadings (i.e., congeneric models) (Komperda, Pentecost, et al., 2018). McDonald's omega is described as the amount of the observed variance explained by the construct (i.e., common construct variance) divided by the total variance (i.e., the common construct variance and error variance) (Komperda, Pentecost, et al., 2018). Values for omega range from 0 to 1, with 1 indicating that all of the observed variance is from the construct. Therefore, a high omega value (> 0.7) provides evidence of reliability in terms of the internal consistency of the items.

Chapter 3: Methods

All the projects included in this dissertation used a combination of qualitative and quantitative methods to address the research questions. In this chapter, methods are first described generally and then details specific to each project and research question are given.

Qualitative Methods

Qualitative methods are often used to collect evidence of test content and response process validity and can take many different forms. For this research, qualitative data were collected through a combination of written responses, interviews, and focus groups (i.e., group interviews) and then analyzed through coding the student responses. Open-ended written responses and focus groups were generally coded using elements of thematic analysis, while response process analysis was used to analyze students' written and interview responses and explanations to survey items.

Thematic Analysis

Thematic analysis can be used to code open-ended responses to specific prompts that are given to students. The purpose of thematic analysis is to analyze themes, or patterns, that appear in the data (Braun & Clarke, 2006, 2020; Braun et al., 2019). There are multiple ways to approach thematic analysis and the coding of open-ended responses. Reflexive thematic analysis focuses on coding as an iterative process, where the codes and themes are developed throughout the coding process, and emphasizes the researcher's subjectivity and interpretation of the data as an essential component of creating themes. Thematic analysis using a coding reliability approach focuses on collecting evidence for pre-determined themes. This type of analysis uses a defined codebook and multiple coders to ensure that there is agreement on what codes and themes are present in the data. A third type of thematic analysis, called codebook thematic analysis, combines elements of reflexive and coding reliability. Although codebook thematic analysis typically starts with some pre-determined codes or themes, additional codes or themes are developed throughout the process (Braun & Clarke, 2020; Braun et al., 2019).

Intercoder reliability (ICR) can be used to determine the agreement between multiple coders when data are coded by each independently using a codebook (O'Connor & Joffe, 2020). Although coders may ultimately come together, discuss the coding of a data set, and reach a consensus (i.e., coding to consensus), reporting agreement allows for transparency throughout the process. ICR can be assessed through Cohen's kappa (Cohen, 1960), which accounts for agreements due to chance.

Response Process Analysis

Response process analysis is focused on determining whether students respond to forced-response items (e.g., multiple-choice or Likert-type scale) in the way intended by the developer (Arjoon et al., 2013). To gather response process data, students are asked to respond to survey items and then explain why they chose the forced-response option that they did. These responses are then coded based on, 1) if their explanation matched their selected response (e.g., if a student agreed with an item then their explanation of why they chose that response should also agree), 2) if they were unclear on the meaning of the item (e.g., if they ask for further clarification or were unsure on how to respond), 3) if they responded that the item was not relevant to them or the context of the course, or 4) if they found certain items very similar. The results from coding these responses can then be used to provide qualitative evidence for removing items prior to quantitative analysis or to provide support for items that are removed during quantitative analysis.

Quantitative Methods

Quantitative methods are often used to provide evidence of structural and consequential validity, as well as to describe potential item response differences between groups. Quantitative data for this research was collected through surveys given to the students, both online and on paper. Evidence of structural and consequential validity were assessed with factor analysis and measurement invariance, respectively. Comparisons between different groups were evaluated with χ^2 comparisons, structured means modeling, or analysis of variance. Relations between predictor and outcomes variables were assessed through regression analysis.

χ^2Comparisons

When an item response scale includes binary (e.g., yes/no) or ordinal (e.g., never/sometimes/always) dependent variables, responses between groups can be compared using χ^2 tests, which compare the observed frequency of response choices to what would be expected if there were no differences between the groups. When multiple groups are compared, a significant result indicates that there is a difference across groups but does not provide information about which groups are different. Therefore, if a significant difference is found across multiple groups, pairwise posthoc χ^2 tests can be completed to compare each group to each of the other groups. When response counts are low and 25% of the cells of a contingency table have expected counts below 5 and a minimum expected count below 1, Fisher’s exact test is more appropriate for calculating the significance of the group differences (Mayers, 2013). The effect size of the differences detected by either test can be determined using Cohen's *w*, where 0.1, 0.3, and 0.5 are considered small, medium, and large effects, respectively (Cohen, 1992).

Factor Analysis

Factor analysis is used to provide evidence of structural validity of the data collected with different measures through analyzing the relations between measured items and factors (i.e., latent variables) that are not measured directly. This can be done without a preconceived structure through exploratory factor analysis or with an *a priori* model through confirmatory factor analysis.

Exploratory Factor Analysis

In cases where many items are added or modified in a measure, exploratory factor analysis (EFA) can be used to provide evidence of structural validity. EFA facilitates the exploration of different factor structures as it allows all items to load on each factor and does not require an *a priori* model of the relations between items and factors. Factors produced through EFA are evaluated based on whether similar items tend to load strongly on the same factor. The number of factors to include in an EFA can be determined from theory and/or quantitative procedures, such as the Kaiser criterion and the scree test (Brown, 2015). The Kaiser criterion selects the number of factors as the number of eigenvalues derived from the correlation input matrix that are greater than 1.0. The scree test is an observational technique that uses the screeplot produced with the eigenvalues on the y-axis and the number of factors on the x-axis. The number of factors is the number of eigenvalues present before the 'elbow' of the plot (Brown, 2015). Using both of these procedures together can provide a reasonable estimate of the number of factors to include for an EFA.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is used when there is already an *a priori* model of the relations between items and factors. Data-model fit is evaluated through a combination of different fit statistics; which can include comparative indices, absolute fit indices, and parsimony correction indices (Brown, 2015). Comparative indices compare the fit based on the proposed model to the fit of a baseline model. These include the comparative fit index (CFI) and Tucker-Lewis index (TLI). CFI ranges from 0 to 1, with values closer to 1 indicating better fit. The TLI is similar to CFI in that values closer to 1 indicate better fit, however, TLI is not bounded and can have a value greater than 1. Absolute fit indices, such as the standardized root mean square residual (SRMR), evaluate the discrepancy between the correlations predicted from the model and the correlations from the data. Parsimony correction indices, such as the root mean square error of approximation (RMSEA), are similar to absolute fit indices except that they incorporate a penalty for model complexity. Values for SRMR and RMSEA that are closer to 0 indicate better data-model fit.

There are no official cutoffs for determining good data-model fit. Some of the most commonly used recommendations are from Hu and Bentler (1999) who suggested good fit is obtained when values are greater than 0.95 for CFI and TLI, less than 0.08 for SRMR, and less than 0.06 for RMSEA. However, it is widely considered that these recommended values should only be used as guidelines when evaluating fit (Hu & Bentler, 1999; Marsh et al., 2004). For example, values above 0.90 for CFI and TLI have been considered to indicate acceptable data-model fit (Bentler, 1990, 1992). Additionally, McNeish et al. (2018) found that fit cutoffs of greater than 0.775 for CFI and lower than

0.20 and 0.14 for RMSEA and SRMR, respectively, could indicate good data-model fit when models contained high item loadings (e.g., ~0.9). In some cases, joint criteria (e.g., CFI ≥ 0.96 and SRMR ≤ 0.09) may be most appropriate (Mueller & Hancock, 2008).

If there is unacceptable data-model fit, modification indices (MIs) can be examined to determine if there are suggested model specifications (e.g., correlation of error terms or association of an item with a different factor) that can be made to the model to improve the fit. The value given for MIs indicate the expected change in χ^2 fit statistic when the modification is included. In the context of CFA, the χ^2 fit statistic evaluates how well the model reproduces the covariance matrix of the data. Although it is rarely used to evaluate data-model fit due to its strong dependence on sample size, it can be used to evaluate differences in nested models (Brown, 2015). The impact of the expected χ^2 change due to the addition of a suggested modification can be determined using Cohen's *w* to calculate the effect size of the change, where 0.1, 0.3, and 0.5 represent small, medium, and large effects, respectively (Cohen, 1992). In addition to having an expected change in χ^2, any modifications made to the model should also be supported by theory and/or qualitative results.

Reliability

Evidence of single-administration reliability of factors found through EFA and CFA can be estimated through the omega statistic (McDonald, 1999). Unlike the alpha statistic, which requires equal item loadings, omega does not require equal item loadings or errors (Komperda, Pentecost, et al., 2018). Similar to the data-model fit indices, there are no formal cutoffs for good reliability; however, values above 0.7 are generally considered acceptable.

Measurement Invariance

Before scores on latent variables can be compared between groups, evidence of consequential validity must be gathered. This is completed through measurement invariance testing (Rocabado et al., 2020). There are multiple sequential steps to evaluate measurement invariance, with each becoming “stricter” and later steps providing support for different types of group comparisons. After sufficient evidence of structural validity is found for a baseline model with the aggregated data set, the data set can be split into comparison groups for invariance analysis. The first step is to establish configural invariance between the groups. This involves testing the same structural model across both groups while allowing all model parameters to be freely estimated for each group. Although establishing this level of measurement invariance does not support any comparisons between groups, if good data-model fit is found, then the next step, called metric invariance, can be tested. Metric invariance also does not provide support for any group comparisons; however, it is an essential step before moving on to the higher levels of invariance that do provide support. Evidence of metric invariance is found when there is good data-model fit for the structural model when the factor loadings are set equal between the groups. The next step of measurement invariance, scalar invariance, involves the additional constraint of equal item intercepts across the groups. Testing for scalar invariance provides evidence that the factor means are not biased from systematic differences in how groups respond to the items. If sufficient evidence is found for scalar invariance, then comparisons of latent means can be supported through the use of structured means modeling (SMM) (Bunce et al., 2017; Rocabado et al., 2020). Further, evidence of scalar invariance allows for evaluation of the highest level of measurement invariance, conservative invariance. Evaluation of conservative invariance additionally constrains the error variances in the model to be equal across groups. Evidence of this level of measurement invariance provides support to compare observed factor scores between groups (Rocabado et al., 2020) using analysis of variance (ANOVA).

For each step of measurement invariance, good data-model fit can be established by comparing the fit statistics between the current model and the model of the previous invariance step. When the difference in χ^2 between the two models is nonsignificant, then there is evidence that there is no significant difference between the two models. Additionally, the change in other fit statistics can be evaluated. Recommended guidelines for these are ΔCFI ≤ 0.010 along with either ΔRMSEA ≤ 0.015 for all levels of invariance or ΔSRMR ≤ 0.030 for configural and metric invariance and ΔSRMR ≤ 0.010 for scalar and conservative invariance (Chen, 2007).

Structured Means Modeling

When evidence for scalar invariance is found, support for group comparisons using latent (i.e., factor) means is provided. Latent mean comparisons are accomplished through structured means modeling (SMM). The model includes the same constraints used to evaluate scalar invariance (i.e., equal item loadings and intercepts across the groups for continuous variables). SMM only allows for relative differences between latent means to be determined, so a reference group is chosen. The latent mean of the reference group is set to zero with the latent mean of the comparator group allowed to be freely estimated. The value obtained for the latent mean of the comparator group is then representative of the difference between the two groups (Thompson & Green, 2013). The effect size of the difference can be calculated as the absolute difference in the latent factor means divided by the square root of the pooled variance of the factors (Thompson & Green, 2013). This effect size calculation is similar to Cohen's d, where 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes for observed scores (Cohen, 1992). However, as latent means are theoretically free from measurement error, the effect size for these differences should be larger than for differences between measured variables (Thompson & Green, 2013).

Analysis of Variance (ANOVA)

When support for conservative invariance is found through measurement invariance testing, group differences can be analyzed for the presence of a statistically significant difference using analysis of variance (ANOVA) with the observed scale score means. The effect size for differences found with an ANOVA can be evaluated using Cohen's f, where values of 0.10, 0.25, and 0.40 represent small, medium, and large effects, respectively (Cohen, 1992).

Regression

Regression analysis can be used to determine the association between predictor and outcome variables. Multiple regression includes the effect of multiple predictors simultaneously. With multiple regression, the effect size of the association between a single predictor and the outcome can be evaluated through calculating Cohen's f^2 using the semi-partial correlation coefficient of the predictor, where values of 0.02, 0.15, and 0.35 represent small, medium, and large effects, respectively (Cohen, 1992).

Project I Methods

This project was part of a larger research project to evaluate different variables related to flipped general chemistry courses across multiple institutions. The research included in this dissertation focused on only a subset of the larger project and was completed in two phases. The first phase encompassed evaluating student interactions with and perceptions of required pre-class materials (RQ 1.1). The second phase focused on assessing differences in students' academic and chemistry self-efficacy, both throughout the term and between different institutions (RQs 1.2 – 1.3). As the data collection for this project was encompassed within the larger research project and collected by multiple researchers, only brief descriptions of the collection methods are described.

Research Questions

RQ 1.1: What are students' interactions with and perceptions of required pre-class materials in flipped courses?
RQ 1.2: How does student self-efficacy change over the term in flipped courses?
RQ 1.3: How does student self-efficacy compare across flipped courses at different institutions?

Participants

General chemistry students from five different institutions were included in this study. The data collected in flipped courses within these institutions were used in both phases, although not all courses were evaluated in both. The courses and respective institutions included in each phase are presented in Table 3.1.

Table 3.1. The courses and/or institutions evaluated in both phases of the project.

Phase One	Phase Two
Course One	Southeastern Institution
Course Two	---
Course Three	Western Institution
Course Four	Northwestern Institution
Course Five	---

Phase One: Students' interactions with and perceptions of pre-class materials

Data Collection

Responses to a pilot survey were collected from 312 students in Courses One, Two, and Five midway through the term during a non-exam week. Additionally, 15 focus groups with 32 students from Courses Two and Five were conducted by two other researchers and video recorded. The final version of the survey was administered in all five courses midway through a subsequent term with a different population of students.

Survey Development

Results from the focus groups and the pilot survey were used to inform modifications to the final survey to better assess students' interactions with and perceptions of the pre-class materials. The focus groups were centered around the following questions regarding the pre-class material.

1) Do you regularly watch the posted videos?
    a. If yes…
        i. How do you watch them?
        ii. What do you do while watching them?
        iii. When do you typically watch them for the first time?
        iv. Do you ever re-watch them? If so, when?
    b. If no, have you ever accessed them?
        i. If yes, why do you not regularly watch them?
        ii. If no, why do you not access them?
2) Do you find the videos helpful for learning the material? Please explain why or why not.

3) What do you like or dislike about the videos? Please explain.

The initial codebook was created by the primary coder after coding two videos, one from Course Two and one from Course Five. A secondary coder independently coded the same videos using the codebook. The two coders then met to discuss discrepancies and made modifications to the codebook. Both coders used the modified codebook to code two additional videos and Cohen's kappa was used to assess ICR. All remaining videos were coded by the primary coder using the modified codebook.

The pilot survey included questions similar to those asked in the focus groups and contained multiple types of items, including single-, multi-, and open-response. The survey contained logic flow steps based on each student's responses, thus not every student saw every item. Single- and multi-response items were evaluated based on the percentage of students that selected each response for that item in relation to the number of students that were presented with the item. A codebook for the open-response items was created by the primary coder. A secondary coder then used the codebook to independently code responses across each item. The coders came together to discuss and come to a consensus regarding any discrepancies. The response percentages for each code were calculated based on the number of students who were presented with the item.

The results from the pilot survey and focus groups led to modifying some of the wording of the items to better represent how students were interacting with the materials, as well as informed the creation of new or additional response options for some of the items. The logic flow steps were modified such that students who watched at least "some" of the pre-class videos still saw items relating to how they watched them and what they found helpful and not helpful about the videos. Additionally, the items related to what students liked or disliked about the videos were removed, as student responses to these items had significant overlap to their responses about what they found helpful or not helpful. All of the item responses in the final survey were adapted to single- or multi-response options, with open-response boxes only being provided when students selected "other" as their response.

Data Analysis

Data collected with the final version of the survey was cleaned to remove multiple responses from the same student and any responses from students who did not consent. The final survey contained both single- and multi-response items. Data were analyzed slightly differently based on the type of item. For single-response items, the percentage of students who selected each response was calculated using the number of students that were presented with that item based on the logic flow steps. Differences in responses across courses were analyzed with χ^2 tests. If the data contained low response counts in some categories then Fisher's exact test was used to determine the significance. Both tests were calculated using the *stats* package (version 3.6.2) in the statistical software R (version 3.6.2). Effect size was calculated as Cohen's *w* using the *rcompanion* package (version 2.3.25) in R. For multi-response items, the percentage of students that selected each possible response was calculated using the number of students that were presented with the item. As students could select multiple options, the total percentage across responses to these items may total over 100%. Similar to the single-response items, χ^2 tests were performed for the responses to the multi-response items. For these items, each response was analyzed separately and treated as a single yes-no response, with students who selected the response being counted as "yes" and students who did not select the response counted as "no". Pairwise comparisons for both single- and multi-response options were analyzed only when there was a significant overall χ^2 result and the response patterns trended with known differences between the courses. The significance for differences detected in pairwise comparisons was determined using Fisher's exact test with a Bonferroni correction using the *fmsb* package (version 0.7.0) in R.

Phase Two: Students' self-efficacy throughout the term and between institutions

Survey Measures and Data Collection

The measure used to evaluate students' chemistry self-efficacy (CSE) included six items related to specific chemistry concepts (Ferrell & Barbera, 2015). Responses were collected on a five-point scale that included *very poorly, poorly, average, well,* and *very well*. Academic self-efficacy (ASE) was measured using eight slightly adapted items from the Self-Efficacy for Learning and Performance subscale of the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al., 1991). Time management (TMT) and Concentration (CON) measures each contained eight items from the corresponding subscales from the Learning and Study Strategies Inventory (LASSI) (Weinstein et al., 2002). ASE, TMT, and CON measures were given on a five-point Likert-type scale from *strongly disagree* to *strongly agree*.

The survey measures were administered to students at the Southeastern, Western, and Northwestern institutions twice, once at the beginning of the term and once at the end. Only students that responded to both surveys were included in the data analysis. Demographics data were collected through self-response items at the end of both surveys.

Data Analysis

Before group comparisons could be made, evidence of structural and consequential validity needed to be established (Rocabado et al., 2020). This was completed by another researcher as part of the larger project. Briefly, evidence of structural validity of the data collected with the measures was found through good data-model fit using slightly modified models. The final CSE, ASE, TMT, and CON measures included four, six, five, and six items, respectfully, and also showed evidence of good reliability. Consequential validity was evaluated through measurement invariance and each measure was found to have sufficient evidence of longitudinal scalar invariance, scalar invariance between institutions, and scalar invariance between demographic groups (i.e., gender identity and underrepresented minority (URM) status).

All SMM analyses were conducted with matched data sets from each institution using the *lavaan* package (version 0.6-5) in R. Means and variance adjusted weighted least squares (WLSMV) estimation with listwise deletion was used to account for the ordinal scales of the items. Due to the use of ordinal variables, models included equal item loadings and thresholds across groups, with item intercepts set to zero for each group (Kline, 2015). Data collected from the Western institution did not include the use of the full response scale for the CSE measure (i.e., no students selected *strongly disagree* for Items 2, 4, and 5 on the post CSE scale). For comparisons made within this institution (i.e., longitudinal comparisons), thresholds for these response categories could be easily removed. However, this was not the case for between institution comparisons. Instead, a ‘dummy’ response pattern was added to the Western Institution data set. This response pattern included the missing response option for the appropriate items and the mean of the responses for all complete items. The effect of adding the ‘dummy’ response pattern was evaluated through adding a series of similar responses patterns to the institutions with complete data sets. The data-model fit statistics and latent means for these data sets were compared to the data sets that only contained real data. The results between the two data sets were similar and no significant differences were detected, suggesting that adding a single ‘dummy’ response pattern to the Western institution data set as needed would not significantly affect the outcome of the results.

Latent factor means from the pre assessment to the post assessment were compared for each measure at each institution individually. The pre factor mean for each comparison was set to zero as the reference. Thus, the latent mean differences that were calculated represent the change in the factor mean over the course of the term at the institutional level. Additionally, the aggregated data set (i.e., all institutions) was used to analyze pre to post latent mean differences based on demographic groups (i.e., gender and URM status). For these comparisons, male and non-URM (i.e., non-Latino/a White or Asian) were set as the reference groups, with female and URM set as the comparator groups, respectively.

Post latent means for the CSE and ASE factors were compared between institutions. For these analyses, the corresponding pre factor (i.e., pre CSE or pre ASE), and the pre TMT and CON factors were included in the model as covariates. This allowed the differences in the factor means of these pre assessments to be accounted for in the post factor mean comparisons between institutions (Hancock, 2004). All pairwise comparisons between the three institutions were made. Each comparison included a reference institution, whose latent means were set to zero, and a comparator institution whose latent means were freely estimated. Post CSE and ASE latent mean comparisons were also conducted with the aggregated data set using the same demographic groups assessed in the pre to post latent mean comparisons.

Project II Methods

This project was focused on evaluating student perceptions of active learning environments and took place in two phases. In the first phase, an existing measure of student perceptions was evaluated in the active learning environment incorporated into the first term general chemistry course at Portland State University (PSU) (RQ 2.1). Results from phase one led to modifications to the survey measure (RQ 2.2). The modified survey was then evaluated during phase two for use with two different types of active learning activities incorporated into the third term principles of biology course at PSU (RQ 2.3).

Research Questions

RQ 2.1: How well does an existing measure of student perceptions of active learning activities function in different active learning environments?

RQ 2.2: What modifications can be made to an existing measure in order to measure student perceptions of these environments?

RQ 2.3: How well does this modified measure function in different active learning environments?

Course Information

Students from two PSU courses were surveyed over two separate terms within the same academic year. Phase one was completed during a first term general chemistry course that incorporated worksheet activities similar in design to POGIL activities. For these activities, students were expected but not required to work in groups of 2 – 4. The activities were facilitated by the instructor, a graduate teaching assistant (TA), and multiple undergraduate learning assistants (LAs). Although clickers were used throughout the activity to gauge students' current understanding, they were also implemented during lecture days, thus the main difference between a lecture day and active learning day was the presence of the worksheet. Phase two was completed in a third term principles of biology course. This course incorporated two types of active learning strategies; clicker days and Deliberative Democracy (DD) activities (Komperda, Barbera, et al., 2018). During clicker days, questions were incorporated into the lecture given by the instructor and students were encouraged to 'think-pair-share' with nearby students when prompts were given. DD activities were two-day deliberation modules. Students were introduced to a real-world problem and then asked to come to an evidence-based consensus on a policy recommendation through completing required readings, quizzes, and worksheets. All students were required to work on DD activities in a randomly assigned group with 3 – 5 students. The instructor, graduate TA, and undergraduate LAs were present and facilitated the group work for DD activities.

Phase One: Evaluation of the ASPECT for measuring student perceptions

Survey Measure

The Assessing Student Perspective of Engagement in Class Tool (ASPECT) was developed by Wiggins et al. (2017) to measure students' perceptions of their cognitive and affective engagement in active learning environments. The measure contains 19 items on a six-point Likert-type scale from *strongly disagree* to *strongly agree*. During the original development of the measure, 16 of these items were found to assess three factors; personal effort (three items), value of group activity (nine items), and instructor contribution (four items). The remaining three items were not found to contribute to these factors (Wiggins et al., 2017).

Data Collection

The ASPECT survey was administered online through Qualtrics. Students were notified of the survey through an in-class announcement on the day of the activity, as well as through an announcement on the course's learning management site, which included the link to the survey. Students were given 24 hours to access and complete the survey. All students who accessed the survey were given a nominal amount of extra credit.

Data Analysis

Student responses were cleaned before analysis by removing students who did not consent, multiple responses submitted by the same student, and any incomplete responses. Additionally, a check item was included in the survey that asked students to select a specific response (i.e., *somewhat agree*) for that item. Students who did not respond to the check item correctly were assumed to have randomly responded to the survey and their response was removed.

As there was already an *a priori* model for the ASPECT due to the original development of the measure (Wiggins et al., 2017), CFA was used to evaluate internal structure validity. The model included three factors of personal effort, value of group activity, and instructor contribution. The CFA was completed using the statistical program R (version 3.6.2) with the *lavaan* package (version 0.6-5). Maximum likelihood estimation with the Satorra-Bentler adjustment and robust standard errors were used to account for any non-normality of the data (Satorra & Bentler, 1988). Good data-model fit was defined as CFI ≥ 0.95, TLI ≥ 0.95, RMSEA ≤ 0.06, and SRMR ≤ 0.08 based on recommendations from Hu and Bentler (1999). Modification indices (MIs) were evaluated using Cohen's *w* as a guideline for the effect size of the suggested modifications.

Phase Two: Evaluation of the mASPECT in different active learning environments

Modified Survey Measure

Modifications to the original ASPECT were made to create a modified ASPECT (mASPECT). Two versions of mASPECT were created; one for clicker days (mASPECT-C) and one for DD activity days (mASPECT-DD). Minor wording changes to both versions included changing the phrase "group activity" to "this class" and changing "instructor" to "Professor/Teaching Assistant". When possible, the same item wording was used for both versions of the survey. However, some minor wording changes were made to account for differences between the environments. For example, mASPECT-DD contained the item, "I made valuable contributions when *working* with other students during today's class" compared to the same item in the mASPECT-C version, "I made valuable contributions when *having discussions* with other students during today's class." Additionally, since LAs were present during DD activity days but not clicker days, only the mASPECT-DD included LA-worded items that paralleled each of the Professor/TA items.

Some new items were also created during the modifications to allow for the exploration of different factor structures, as results from phase one suggested the possibility of a group-related factor and a lack of items to assess personal effort. Eight new items were created to increase the number of items in the personal effort category. Additionally, four "other-focused" items were created that were similar to the "self-focused" items already included in the original ASPECT.

After item modifications and additions, the mASPECT-DD included 35 items and the mASPECT-C included 31 items. These included the original 19 ASPECT items (Items 1-19, see Table 6.1 in Chapter 6), eight new items related to personal effort (Items 20-27, see Table 6.2 in Chapter 6), and four new items related to group function (Items 28-31, see Table 6.2 in Chapter 6). The mASPECT-DD also included four additional LA-worded items (Items 13B, 14B, 15B, and 16B, see Table 6.1 in Chapter 6).

Data Collection

The surveys were given online through Qualtrics. Students were notified of each survey through an in-class announcement on the day of the activity. The link to the survey was included in an announcement posted on the course's learning management site. Students were given 48 hours to access and complete the survey. All students who accessed the survey were given a nominal amount of extra credit. At the end of the surveys, students were given the option to include their email address if they were interested in participating in a brief in-person interview about the survey items. Interested students were randomly selected and sent an email to determine their availability. Interviews were scheduled at a time that was convenient to the students. All students who participated in an interview were compensated with a $20 gift card.

Interview Protocol and Analysis

Each student was interviewed individually and all interviews were audio recorded. During each interview, students were asked to read each of the items for either mASPECT-C or mASPECT-DD aloud, state which response they selected, and then explain their reasoning for choosing that response. Additional follow-up questions were asked by the researcher to gain additional insight into their explanations as needed. The audio recording for each interview was initially analyzed by two researchers independently for evidence of response process validity. If either researcher thought that the student's response or explanation was not in alignment with the item, then the item was flagged for possible confusion or irrelevance to the active learning environment. The two researchers then came together and discussed the responses to each item and came to a consensus on whether items seemed unclear or irrelevant. A third researcher similarly analyzed the responses for clarity and relevance. Items that all three researchers found unclear or irrelevant were removed prior to quantitative analysis of the survey responses.

Survey Data Analysis

Survey responses were cleaned by removing students who did not consent, additional responses submitted by the same student, incomplete responses, and responses that did not respond to the check item correctly. Additionally, the survey included an item that asked students to select the topic of class on the day of the survey, as well as an item asking if students worked with a group or discussed with other students during class that day. These items were included as additional checks to filter out students who may not have attended class and/or who may not have participated in the activity or discussed with other students during clicker days.

The survey items from both mASPECT-C and mASPECT-DD that corresponded to the 16 items that contributed to the original ASPECT factor structure (Wiggins et al., 2017) were analyzed with CFA using the *a priori* model with personal effort, value of group activity, and instructor contribution factors. This was completed with the same estimation type and fit cutoffs that were used in phase one.

Due to the modifications and additions to mASPECT-C and mASPECT-DD, additional factor structures were explored through EFA. The number of factors were determined through the Kaiser criterion and scree test, which were completed with the *stats* package (version 3.6.2) in R. EFAs were completed using the *psych* package (version 1.9.12.31) in R with principle axis factoring and promax (i.e., oblique) rotation, as this rotation method allows for correlations between factors. All negatively worded items were reversed coded. The data collected with each survey version was analyzed using an iterative process consisting of an EFA, removal of items that did not meet certain criteria, and then a subsequent EFA with the remaining items (Hancock et al., 2010). The criteria used for the removal of items were factor loadings of less than 0.4, high cross-loading on two or more factors, or loading on factors that contained less than three items. Items with cross-loadings between 0.3 and 0.4 on a secondary factor were flagged but not immediately removed. Iterative EFAs were completed until all items met these criteria and the final structure produced well-formed factors. All the items in the final factor structures had cross-loadings of less than 0.35 on the non-primary factors. Omega was used to evaluate the single-administration reliability of each factor based on the final factor structure found through EFA using the *userfriendlyscience* package (version 0.7.2) in R.

Project III Methods

This project was centered around evaluating student engagement during general chemistry worksheet activities and was completed in three phases. Phase one focused on gathering information about how students perceive engagement in the worksheet activities, both when done in-person and remotely (RQ 3.1). The second phase used qualitative and quantitative methods to modify an existing measure of student engagement for use with the worksheet activities (RQ 3.2 and 3.3). Finally, the third phase used this modified measure to assess differences in student engagement across different groups (RQ 3.4) and to explore the relation between engagement and student understanding (RQ 3.5). The phases of this project, along with the respective terms, type of instruction, and instructors, are included in Table 3.2.

Table 3.2. Phases of Project III and respective term, type of instruction, and instructor(s).

<table>
<tr><th colspan="3">Phase (RQ)</th><th>Term</th><th>Type of instruction</th><th>Instructors</th></tr>
<tr><td>▨</td><td rowspan="2">Phase Two – preliminary (RQ 3.2)</td><td rowspan="2">▨</td><td>Fall 2019</td><td rowspan="2">In-person</td><td>Instructor 1
Instructor 2</td></tr>
<tr><td rowspan="2">Phase One (RQ 3.1)</td><td>Winter 2020</td><td>Instructor 1</td></tr>
<tr><td rowspan="3">Phase Two – final (RQs 3.2 – 3.3)</td><td rowspan="4">Phase Three (RQ 3.4 – 3.5)</td><td>Fall 2020</td><td rowspan="3">Remote</td><td>Instructor 1
Instructor 2</td></tr>
<tr><td rowspan="3">▨</td><td>Winter 2021</td><td>Instructor 1
Instructor 2</td></tr>
<tr><td>Spring 2021</td><td>Instructor 1
Instructor 3</td></tr>
<tr><td>▨</td><td>Fall 2021</td><td>In-person</td><td>Instructor 1</td></tr>
</table>

Research Questions

RQ 3.1: How do students perceive engagement in worksheet activities?

RQ 3.2: What modifications can be made to an existing survey measure in order to measure student engagement in these activities?

RQ 3.3: How well does a modified measure of engagement function in this environment and student population?

RQ 3.4: How does engagement in these activities differ across groups?

RQ 3.5: How does engagement in these activities relate to students' understanding of the material?

Course Information

All three terms of the three-term general chemistry sequence at Portland State University (PSU) were included throughout this study. Each term consisted of two sections taught by different instructors, with a total of three different instructors over the time of this project. Each instructor similarly incorporated worksheet activities into their class. The design of these worksheets was influenced by POGIL activities, where the worksheets are built to guide the students through their own exploration of the material. For the in-person classes, students were encouraged to work on the worksheets in groups of 2 – 4 students and generally worked with students that were seated nearby. The instructor, a graduate teaching assistant (TA), and multiple undergraduate learning assistants (LAs) moved around the room to facilitate discussion and to answer questions. For the remote classes, collaborative breakout rooms were created. For some of the remote activities, the breakout rooms were randomly assigned, while others were assigned prior to class based on students responses about the 'type' of room they wanted to be a part of. The options included, 1) collaborative rooms where students were expected to work together with microphones on, either with or without their video camera on, 2) a room where students could work independently but ask an LA questions, or 3) to stay in the main room and work on the worksheet independently without the ability to ask questions. For the collaborative breakout rooms, graduate and undergraduate LAs rotated through to facilitate discussion and answer questions. In both in-person and remote environments, the instructors continuously assessed the understanding of the groups throughout the activity by incorporating clicker/polling questions and/or gathering feedback from the LAs.

Phase One: Students' Perceptions of Engagement in Worksheet Activities

Data Collection

Open-ended responses were collected for two separate terms, one during an in-person term and the other during a remote term. Responses related to in-person instruction were collected through an online survey given at the end of a second term. An announcement about the survey was given during class, as well as posted on the courses' learning management site and included a link to the Qualtrics survey. Students were offered a nominal amount of extra credit for accessing the survey. Responses for remote instruction were collected through interviews (group and individual) completed during a first term course. Students were asked if they would be interested in participating in a group interview at the end of the quantitative survey given during the same term. All students who indicated they were interested were contacted and asked to fill out an online consent form and select the session times that they were available to attend. Based on students' availability, group interviews were then scheduled and a Zoom link was sent to the students. Each interview only contained students from a single class section. Fourteen students participated in a total of eight interviews, four for each section, with 1 – 3 students per group. Although each interview was scheduled with at least two students, some students did not show up. In these cases, to respect the other student's time, the interview proceeded as an individual interview with the one student. All interviews were recorded over Zoom with audio and visual and transcribed before analysis.

Survey Items

The survey contained four open-ended response items. The first item asked students “Overall, how engaged were you in the worksheet activities this quarter?” Students were instructed to select a value on a slider scale from *not engaged* (0) to *very engaged* (100) and then to explain why they chose that value. The slider scale value did not contribute to the analysis of the responses, but was included to prompt the students to think about their engagement in the activities before responding. On a separate page, students were then randomly given one of three definitions of engagement and asked to respond to the remaining questions with that definition in mind. The three definitions of engagement were focused around describing the behavioral, cognitive, and emotional dimensions of engagement. The definitions were created by the primary researcher from descriptions of these dimensions by Fredricks et al. (2004), which were then read over by another researcher and modified for meaning and clarity. The final definitions used for these survey items are given in Table 3.3. Students were asked to rank their engagement in the worksheet activities based on the definition they were given and explain their response. They were then asked to respond to the following items based on the engagement definition.

1) How would you describe a student who is **NOT engaged** in the worksheet activities based on the definition above?
2) How would you describe a student who is **VERY engaged** in the worksheet activities based on the definition above?

Table 3.3. Definitions of the three dimensions of engagement included in the survey.

Dimension of Engagement	Survey Definition
Behavioral	Engagement is the physical participation or involvement in the worksheet activities.
Cognitive	Engagement is exerting mental effort to comprehend ideas or skills presented in the worksheet activities.
Emotional	Engagement is the positive feelings towards the worksheet activities.

Interview Protocol

Interviews were completed over Zoom with 1 – 3 students per group using a semi-structured interview approach. The students were first asked to describe their engagement in the worksheet activities. Follow-up questions were asked by the researcher as needed to guide students to describe aspects of what they did, thought, and felt related to their engagement. The students were then presented with definitions of the three dimensions of engagement one at a time using the chat function in Zoom. These definitions were similar to the ones given during the open-response survey; however, as the word "effort" has been associated with both the behavioral and cognitive dimensions (Fredricks et al., 2004), the cognitive definition was modified to remove "effort" (Table 3.4). One student would volunteer to read the definition aloud and then all students were asked to describe being very engaged and not engaged in the worksheet activities based on each definition. Throughout the interview, students were asked follow-up questions as needed for clarification.

Table 3.4. Definitions of the three dimensions of engagement used in the interviews.

Dimension of Engagement	Interview Definition
Behavioral	Engagement is the physical participation or involvement in the worksheet activities.
Cognitive	Engagement is trying to comprehend ideas or skills presented in the worksheet activities.
Emotional	Engagement is the positive feelings towards the worksheet activities.

Data Analysis

The data collected from the open-ended survey and the group interviews were coded using elements of thematic analysis. As the survey responses were collected first, these were coded using an inductive approach, where the codes appear from the data. As the open-ended response boxes were a required part of the survey, responses were first cleaned to remove any “filler” or illogical responses from the data set (e.g., “n”, “.”, etc.). The responses collected based on the different definitions of engagement were analyzed first. Two coders went through the data and separated student responses into individual statements. These statements were then organized by each coder individually into groupings of similar statements with an overall code associated with them. The coders then came together to discuss these possible codes and came to a consensus on the initial codebook. The responses to all the items were then coded iteratively, starting with those related to the specific dimensions of engagement. For each iteration, a subset of responses were selected. The coders individually coded this subset and then came together to discuss any discrepancies or possible new codes that appeared in the data. Based on the discussion, codes were clarified or added as needed and a new subset of responses were selected for the next iteration of coding. This continued until all responses were coded and no new codes were discovered. Approximately 20% of earlier subsets were then recoded by both coders using the final version of the codebook and ICR was evaluated by calculating Cohen’s kappa using the *irr* package (version 0.84.1) in the statistical software R (version 3.6.2). The remainder of the earlier subsets were then recoded by the primary coder using the final version of the codebook. The prevalence of each code within the responses to each item was calculated as a percentage of student responses with the code compared to the total cleaned responses for the item.

As interviews were conducted in a subsequent term to the survey and completed in a different environment (i.e., remote), a combination of both inductive and deductive coding was used. Two interviews, one from each of the two sections, were analyzed first. Two coders separately read through the transcripts and noted phrases from the students that corresponded to their engagement in the activities. The coders then got together and discussed any discrepancies in highlighted phrases to reach a consensus on what statements related to engagement. The primary researcher then grouped these statements into categories with similar statements to create the initial codebook. Groups of statements were compared to codes that were discovered through the analysis of the open-ended survey responses and similar codes were used when possible. Both coders then individually coded the same two interviews using the initial codebook and came together to discuss any discrepancies or possible new codes and modified the codebook as needed. This codebook was then used to code the remainder of the interview transcripts. For each of these, the coders coded each transcript individually and then came together to discuss any discrepancies. No new codes were discovered during this process and each transcript was coded to consensus, although ICR was evaluated throughout using Cohen’s kappa. The prevalence of the different codes in each section of the transcripts that corresponded to the different engagement definitions were calculated as a percentage based on the total number of participating students.

Phase Two: Developing a Measure of Student Engagement

Preliminary Survey Measure

The preliminary survey was based on an engagement measure for middle- and high-school math and science students developed by Wang et al. (2016). Although this measure was not developed for higher education classrooms or specifically active learning activities, it was developed for use in a science learning environment and was grounded in the theoretical definitions of the three dimensions of engagement (Fredricks et al., 2004). The instrument was initially developed through qualitative interviews with students where they were asked to describe what they were doing, thinking, and feeling while engaged in class (Fredricks, Wang, et al., 2016). The responses from these interviews were then coded into behavioral, emotional, and cognitive categories and items were created within each category. Additionally, a social dimension was found to be present throughout all three dimensions (i.e., social-behavioral, social-emotional, and social-cognitive), so unique items were created to assess social interactions. The original survey contains 33 engagement items; 8 behavioral, 10 emotional, 8 cognitive, and 7 social.

The preliminary survey included re-worded items from the middle- and high-school science engagement survey (Wang et al., 2016). All items were changed from present tense to past tense, as the survey would be given after the students completed the activity, and “today’s activity” was included in each item. Other minor wording changes were made with the consensus from another researcher to make the items more applicable to the active learning activities and higher education student population. All items were administered on a six-point Likert-type scale from *strongly disagree* (1) to *strongly agree* (6).

Preliminary Data Collection and Analysis

Preliminary versions of the survey were administered during two terms. During the first term, the survey was administered to students in both sections after a single activity. Students were notified of the survey during an in-class announcement and through the course's learning management site, which included a link to the online Qualtrics survey. Students were given 48 hours after the activity to access and complete the survey. All students who accessed the survey were given a nominal amount of extra credit. After responding to the survey items, each student was asked to provide a short-answer explanation about why they selected their response to one randomized item from each dimension of engagement. At the end of the survey, students were asked to include their email if they would be interested in participating in an interview. From these students, six were randomly selected and participated in a short in-person cognitive interview about the survey. Students were asked to read each item of the survey aloud, state which response they selected, and then explain why they selected that response. Follow-up questions were asked by the researcher as needed to gain additional clarity or details about their explanation. All interviews were audio recorded.

During the second term, the survey was administered to students in one section after two different activities. An announcement of the survey was given at the beginning of class and a paper version of the survey was given to students to complete at the end of class after they had completed the activity. All students who filled out a survey were given a nominal amount of extra credit regardless of research consent. The in-class announcement also notified students about interviews that would take place and directed them to a respective announcement that was posted on the class's learning management site with the link to sign up through Qualtrics. All students who signed up for an interview were contacted and interviews were scheduled with six students. These interviews took place in-person with an identical protocol to those completed in the previous term.

Preliminary data collected during these terms were analyzed to determine if there were items that were not relevant to the environment and population of students and if any items required re-wording. Before analysis, survey responses were cleaned by removing students who did not consent, multiple responses from the same student, and responses that did respond to the check item(s) correctly. Both terms included a 'response-based' check item that asked students to select a specific response (e.g., *somewhat agree*), such that students who did not respond correctly were assumed to have responded randomly to the survey items. Additionally, the first term survey given through Qualtrics included a 'topic-based' check item where students were prompted to select the topic of the day's activity, as students who responded incorrectly may not have attended class that day or participated in the activity. This check item was not included during the on-paper survey administered during the second term since only students who attended class that day could fill out the survey. The cleaned survey data were analyzed through fit statistics, factor loadings, and MIs from individual factor CFAs for each dimension, as well as a four-factor correlated CFA with all four dimensions of engagement. All CFAs were completed with the *lavaan* package (version 0.6-5) in R. Maximum likelihood estimation with the Satorra-Bentler adjustment and robust standard errors were used to account for any non-normality of the data (Satorra & Bentler, 1988). Additionally, short-answer and interview responses were analyzed for evidence of response process validity and results from these were used to support modifications and/or removal of items for the final survey.

Survey Modifications

The results of the preliminary survey and interview data were used to inform minor re-wording of some items for clarity and relevance. Additionally, certain items were found to be irrelevant to the environment or population of students and were removed. For example, the item *"I completed all the required pre-work for today's activity"* was found to be confusing to students in one section, as the instructor of that class did not assign any pre-work for the activity. Additionally, when asked to explain their response to the item *"I plan to share with others what I learned during today's activity"*, one student responded that, "I am not a very social person, so I want to keep my school life separate." This sentiment was repeated by multiple students, which suggested that this item was not relevant for students in higher education courses. Evidence was also gathered that suggested some items might have assessed aspects of both the behavioral and cognitive dimensions in the in-person environment (i.e., "*I put effort into doing today's activity*"). Thus, additional items focused around assessing behavioral aspects only were created and added to the behavioral scale with the collaboration of a secondary researcher. After the preliminary analysis and modifications, the engagement pilot survey measure contained 26 items; 7 behavioral, 6 cognitive, 7 emotional, and 6 social (see Table E.2 in Appendix E).

Data Collection

The pilot survey was administered in both sections during all three terms of general chemistry during the 2020 – 2021 academic year, with three different activities surveyed each term. Students were notified of the survey through an in-class announcement on the days of the activities, as well as through an announcement posted to the course's learning management site. A link to the Qualtrics survey was given both at the end of class through the chat function in Zoom and included on the posted announcement. Students were given 48 hours to access and complete the survey. All students who accessed the survey were offered a nominal amount of extra credit. If assigned breakout rooms were used during the activity, the survey also included an item at the beginning that asked students to select the type of breakout room they participated in that day.

At the end of select surveys during the first two terms, students were asked if they were interested in participating in an individual interview about the survey and/or a group interview about their engagement (details about group interviews are included in phase one). Students who selected that they were interested in participating in an interview about the survey and provided their email were sent an invitation to fill out the consent form online and select when they were available. For interviews about activities with pre-assigned breakout rooms, students were randomly selected and sent invitations based on their response to the type of breakout room they participated in. A total of 21 students participated in response process interviews over both terms. Interviews were completed over Zoom with audio and visual recording. A copy of the survey was provided to the students and they were first directed to think back to the activity and fill out the survey again. They were then directed to read each item of the survey aloud, state which response they selected, and explain why they selected that response. The researcher asked follow-up questions as needed to gain more details about their understanding of the items and/or response reasoning. Students who indicated they did not work on the activity in a collaborative breakout room were not asked to respond to the items related to the social dimension.

Data Analysis

Response process interviews from both terms were analyzed to determine if items were clearly written and relevant to the students in the remote environment. Two researchers individually listened to the recordings and noted if the student's response or explanation was not in alignment with the item. The two researchers then came together and discussed any flagged items and came to a consensus on whether the items seemed unclear or irrelevant. Results from this analysis were used to remove unclear items before CFA and/or to provide qualitative support for the removal of poorly functioning items during CFA.

Survey responses from students who did not consent, multiple responses from the same student (for each activity surveyed), and responses that selected the incorrect check items were removed prior to analysis. All surveys included both a 'response-based' and a 'topic-based' check item. Additionally, each survey either asked students if they worked with others on the activity or, in the case of assigned breakout rooms, which breakout room they participated in. Students who selected that they did work with others on the activity or selected that they participated in a collaborative breakout room were additionally presented with social engagement items.

Data collected with the final survey for all sections and activities were combined into an *aggregated* data set that included responses to the behavioral, cognitive, and emotional engagement items from students who worked alone and those that worked with others on an activity. Although it is expected that students would be represented multiple times in these aggregated data sets, the data were treated as independent due to the specificity of the survey items being directed towards each individual activity and engagement being theorized as a malleable state that changes based on the context (Fredricks et al., 2004; Furlong & Christenson, 2008). Thus, although students were likely repeated within the data sets, each of their responses was unique in that it was in respect to their engagement on a different activity day. The *aggregated* data set was analyzed through CFA. All CFAs were completed using *lavaan* with maximum likelihood estimation with the Satorra-Bentler adjustment and robust standard errors (Satorra & Bentler, 1988) and listwise deletion. Single-factor CFAs for each dimension (e.g., behavioral engagement) were analyzed first using fit statistics, factor loadings, and MIs to assess how well each functioned. Items with suggested item-item error correlations due to a medium to high MI based on Cohen's *w* guidelines (i.e., $w > 0.3$) were flagged for redundancy and qualitative results were used to select one of the items to be removed, as general guidelines for scale development recommend removing or modifying problematic items (Boateng et al., 2018). Following similar guidelines, the resulting scales from the single-factor models were used in the analysis of a three-factor correlated CFA (see model 1 in Figure 3.1). Items were flagged and removed at this stage if there were suggested cross-loadings onto other factors with medium to high MIs. Additionally, a *social* subset of the *aggregated* data set was created that only included students who indicated they worked on the activity with others. The *social* data set included responses to behavioral, cognitive, emotional, and social items. These responses were analyzed in a similar fashion to the overall *aggregated* data set, although the social factor was also included in the analysis of single-factor CFAs, as well as any of the combined models (e.g., model 2 in Figure 3.1).

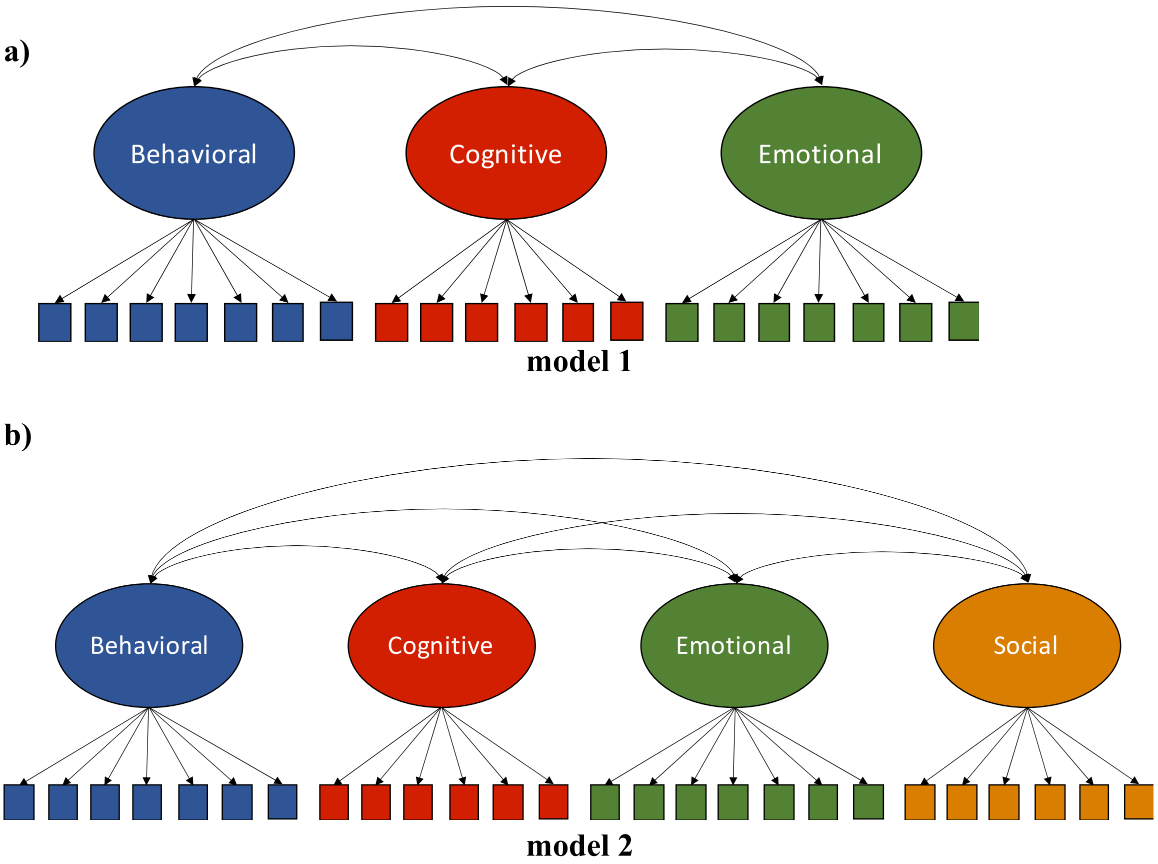


Figure 3.1. Expected correlated a) three-factor, and b) four-factor, models. Models shown assume all items were retained for the individual scales.

Data collected with well-functioning items, as found through single-factor and correlated CFAs, were further analyzed by examining alternative models. Although the three- and four-factor correlated models include the associations between the different dimensions, there may be an alternative model that better describes student engagement in these activities. Three types of model modifications were considered when generating the alternative models.

The first type of modification was focused on evaluating if any of the individual dimensions were better described by a combined factor (e.g., behavioral/cognitive engagement). The theoretical definitions for each dimension describe a different aspect of engagement; however, there exists an inherent overlap between the dimensions as they are all parts of engagement (Fredricks et al., 2004; Sinatra et al., 2015). For a correlated factor model, high factor correlations can indicate that the set of items are measuring very similar constructs (McDonald, 1999). Thus, when evaluating models 1 and 2, any factor correlations greater than 0.8 were noted and an additional CFA was conducted using an alternative model where the two highly correlated factors were combined into a single factor (e.g., see model 3 in Figure 3.2).

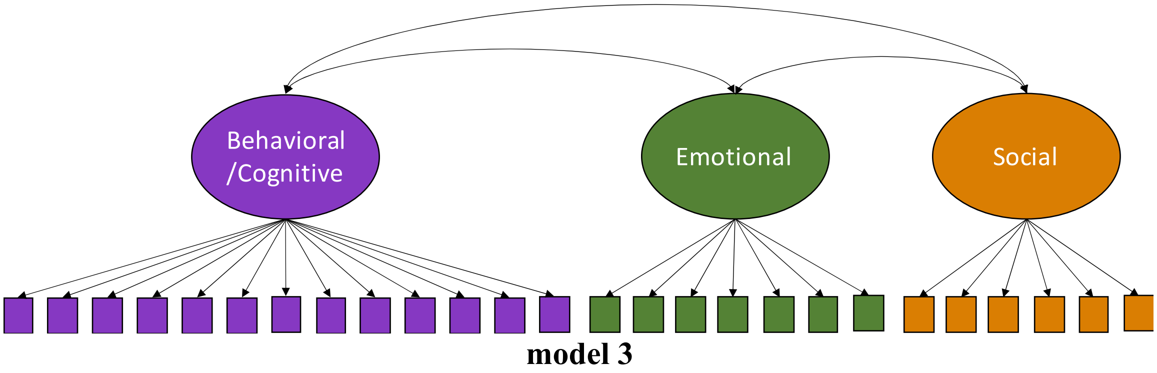


Figure 3.2. Example of a combined-factor model assuming highly correlating behavioral and cognitive factors for model 2.

The second type of model modification was concerned with the relation between overall engagement and the individual dimensions of engagement. Although the correlated model accounts for the relation between the dimensions, all of the dimensions are theoretically related through the overarching construct of engagement. This relation can be evaluated through a second-order type of model, where the commonality between the factors of the individual dimensions is accounted for through the presence of a higher-order factor of overall engagement (see model 4 in Figure 3.3) (Chen et al., 2006). Another model which accounts for the relation between individual dimensions and overall engagement is the bifactor model. This type of model includes overall engagement as a separate factor directly related to each item included in the measure (see model 5 in Figure 3.3). This allows for individual engagement dimensions and overall engagement to be assessed at the same time. Thus, students' overall engagement as well as information about the distinct contributions of the individual dimensions beyond the contribution of overall engagement can be assessed (Chen et al., 2012). The bifactor model type has been suggested as the appropriate model to use when the theoretical framework supports multiple individual factors connected through an overarching factor (Chen et al., 2006). Additionally, support for the bifactor model has been found during the development of other engagement measures that included individual engagement dimensions (Ben-Eliyahu et al., 2018; Wang et al., 2016).

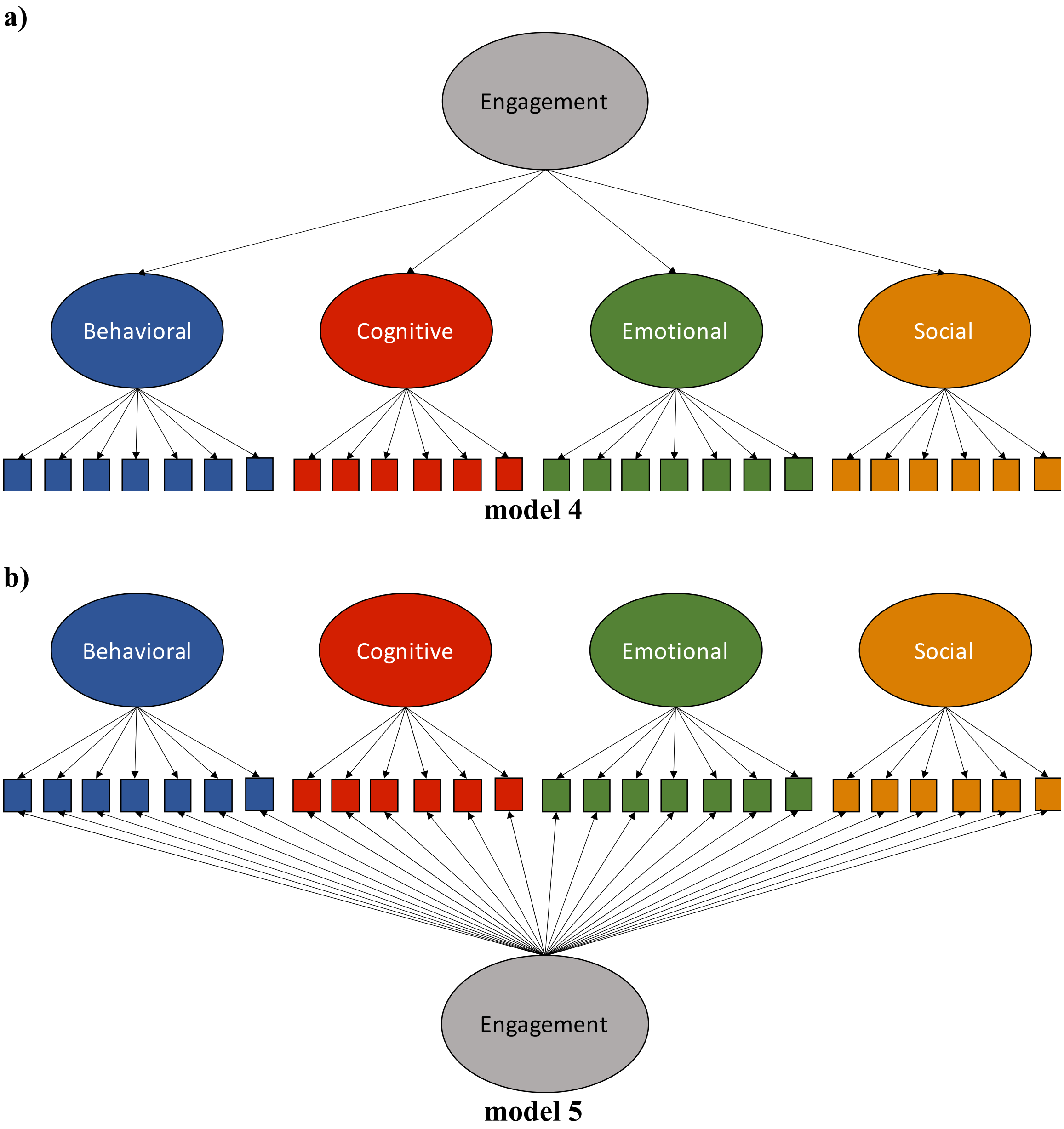
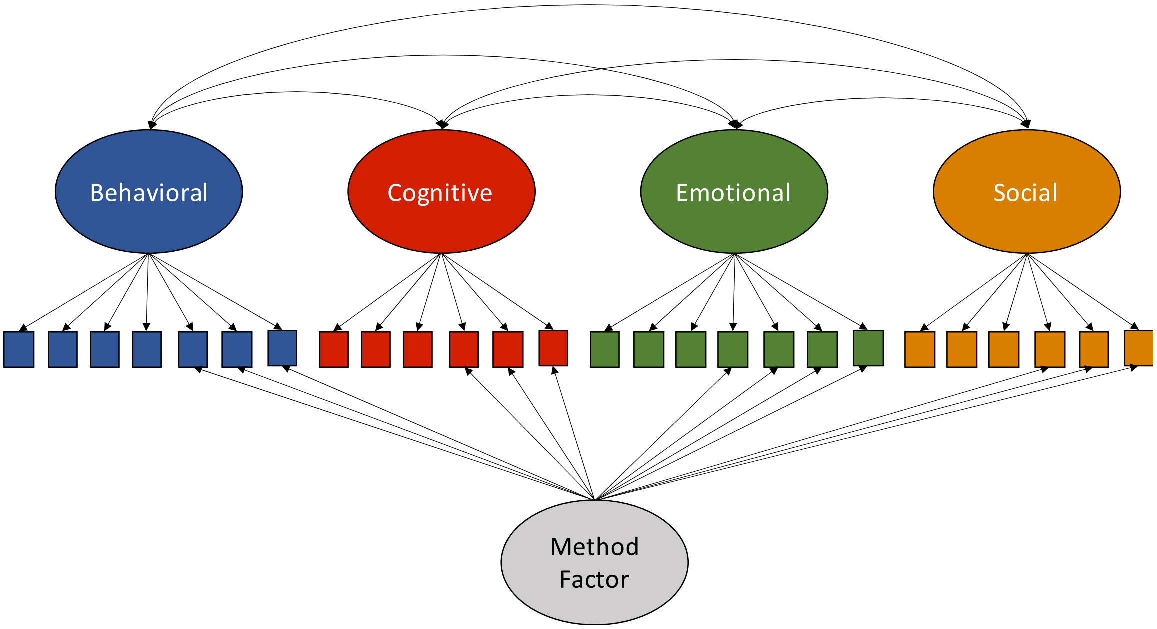


Figure 3.3. Possible a) second-order, and b) bifactor, models of engagement. Models shown assume all items were retained for the individual scales.

The final type of alternative model evaluated was the inclusion of a ‘method factor’ to account for possible differences in student responses to positively- and negatively-worded items. Although including both positive and negative items in a single measure is thought to encourage students to think and respond to each item instead of simply agreeing with everything, it is possible that students may respond these types of items differently simply based on the direction of the wording (e.g., more likely to agree with a positively-worded item than to disagree with the same negatively-worded item) (Wang et al., 2015). Even after negatively-worded items are reverse coded, the addition of a positive or negative method factor can account for this possible bias and improve data-model fit (Wang et al., 2016; Wang et al., 2015; Ye & Wallace, 2013). Therefore, once a 'most appropriate' model was selected based on the evaluation of the original and alternative models, a model with a positive and/or negative method factor was also examined and an overall 'best fitting' model was selected (see model 6 in Figure 3.4 for an example of a model with a method factor).



model 6

Figure 3.4. Example of a model that includes either a positive or negative method factor.

The results from the CFAs for both the *aggregated* and *social* datasets were analyzed based on the resulting fit statistics associated with each model. Fit statistics were evaluated based on the recommendations from Hu and Bentler (1999) for good data-model fit (CFI ≥ 0.95, TLI ≥ 0.95, RMSEA ≤ 0.06, and SRMR ≤ 0.08), as well as acceptable fit if CFI and TLI were above 0.9 (Hinkin, 1998). If the RMSEA value was close to 0.06, the 90% confidence interval for this statistic was evaluated to determine if it fell within the appropriate range. Joint criteria recommended by Mueller and Hancock (Mueller & Hancock, 2008) (i.e., CFI ≥ 0.96 and SRMR ≤ 0.09) were used where appropriate. If good data-model fit was also found for alternative models based on these guidelines then substantive information about the theoretical framework and qualitative evidence were used to guide the decision of selecting the most appropriate model (Morgan et al., 2015). Single-administration reliability of the final scales were evaluated through calculating omega for each of the final individual factors using the *userfriendlyscience* package (version 0.7.2) in R.

Phase Three: Student Engagement Differences and Relation to Student Understanding

Data Collection and Cleaning

The results from phase two informed the creation of the final survey, the Activity Engagement Survey (AcES), that included a 19-item version for students that worked with others (BC-E-S AcES) and a 15-item version for students that worked alone and therefore were not presented with the social engagement items (BC-E AcES). Both surveys included 10 items related to behavioral/cognitive engagement and 5 items related to emotional engagement. The BC-E-S AcES also included 4 items related to social engagement. The most appropriate model for both versions of the AcES was found to be a bifactor model that included an overall engagement factor in addition to the individual dimensions, along with a negative method factor to account for negatively worded items.

The data for phase three came partially from the *aggregated* and *social* data sets from phase two, as well as additional data collected from three activities during Fall 2021 from the in-person section. During Fall 2021, students were notified of the surveys through in-class announcements and announcements posted on the course's learning management site that included the link to the Qualtrics survey. All announcements were given at the end of class during the day of an activity and students had 48 hours to access and complete each survey. All students were required to complete the surveys as part of their course grade; however, only responses from students who consented to participate in this research project were included in the final data set. Students were also asked for consent to allow their course grades to be retained as part of this study.

All students from Fall 2021 were presented with the 15 items from the BC-E AcES and students who indicated they worked with others on the activity were additionally presented with the four social engagement items. Both a topic-based and response-based check item were included on each survey and used to remove responses from students who did not respond to either correctly. Survey responses were cleaned in a similar manner to responses from the 2020 – 2021 academic year.

Group Comparisons

The data collected from phase two and phase three were used to explore two different group comparisons. The first comparison only used the *aggregated* data set from phase two and was focused on comparing engagement between students that worked by themselves (i.e., *independent* group) and those that worked with others (i.e., *social* group). As the independent group did not include responses to the social items, comparisons were only made between behavioral/cognitive, emotional, and overall engagement using the BC-E AcES. The goal of the second comparison was to use the

BC-E-S AcES to compare engagement between students from Instructor A's remote section from Fall 2020 and students from Instructor A's in-person section from Fall 2021.

For both comparisons, evidence of measurement invariance was first collected before any differences in engagement were compared. Internal structure validity of each group's data set was initially evaluated using CFA with the bifactor model with a negative method factor to provide confidence that the relation between the items and the construct matched the hypothetical structure for that group. If evidence of internal structure validity was supported for each, then consequential validity was assessed through measurement invariance testing. This was completed sequentially starting from configural invariance, to metric invariance, then scalar invariance, and finally conservative invariance. Support for each level of invariance was gathered through evaluating the change in fit statistics between the models based on guidelines from Chen (2007). If any step was not supported, measurement invariance testing did not continue and the "stricter" levels were not tested. Group comparisons were only made if they were supported through measurement invariance testing (i.e., evidence of scalar and/or conservative invariance). These analyses were completed using maximum likelihood estimation with Satorra-Bentler adjustment and robust standard errors with the *lavaan* package in R.

If evidence of scalar invariance was supported for a group comparison, SMM was used to evaluate the differences between the groups using *lavaan*. As SMM compares the difference in latent means and not the absolute scale scores, one group was selected as the reference group and the latent means for the factors of that group were constrained to zero. The resulting latent means for the comparison group represent the difference in the factor means when compared to the reference group. The effect sizes of any differences were calculated as the absolute difference in the latent means divided by the square root of the pooled variance of the factors, similar to Cohen's d for observed variables.

If evidence of consequential invariance was supported for a group comparison, the unweighted mean scale scores of the individual engagement dimensions were compared between the groups using ANOVA. This was completed using the *lessR* package (version 3.9.2) in R. The effect size of any differences was evaluated using Cohen's f.

Relation between Engagement and Understanding

Only data collected during Fall 2021 were used to explore the relation between students' engagement in an activity and their understanding of the material. The course included three instructor-written exams throughout the term; one after each surveyed activity. Therefore, students' grades on relevant exam items were used as a proxy for their understanding of the material. For each activity, only the related content on the subsequent exam was considered in order to reduce the amount of time between a students' engagement in an activity and assessment of their understanding of the content covered on that activity. Relevant exam items were first selected by the primary researcher based on their alignment to the content covered in each activity. These selections were independently confirmed by two other researchers. Four exam items were found to relate to the first activity, which covered the topic of *Solutions and Dilutions*. The other two surveyed activities were not included in this analysis due to a combination of low sample sizes and lack of overlap between the content on the activity and the exam questions.

Students' scores on the four exam items related to *Solutions and Dilutions* were first converted into a z-score using mean and standard deviation data from the entire section. The exam items z-score was used as the outcome variable with the mean scale scores of the individual engagement dimensions (i.e., behavioral/cognitive, emotional, and social) as predictors. The multiple linear regression analysis of these relations was completed using the *lm* function in R. The effect sizes of the individual predictors were evaluated using Cohen's f^2 with the semi-partial correlation coefficients of each predictor as found using the *ppcor* package (version 1.1) in R.

Chapter 4: A Multicourse Comparative Study of the Core Aspects for Flipped Learning: Investigating In-Class Structure and Student Use of Video Resources

Reprinted with permission from Naibert, N., Geye, E., Phillips, M. M., & Barbera, J. (2020). A Multicourse Comparative Study of the Core Aspects for Flipped Learning: Investigating In-Class Structure and Student Use of Video Resources. *Journal of Chemical Education, 97*(10), 3490-3505. Copyright 2020 American Chemical Society.

Abstract

Since 2013, the number of publications on flipped learning within chemistry have steadily increased. However, most of these studies focus on flipped course reforms within individual institutions, while the outcomes of any learning environment are dependent on how the environment is structured and the degree to which students interact with its elements. In this study, we apply a coordinated set of assessment practices to investigate similarities among flipped chemistry courses at five institutions in the United States. All courses in the study followed the two basic tenets of flipped learning: 1) foundational information was delivered through pre-class materials (PCMs) and 2) the face-to-face (F2F) environment applied or extended the content through active learning. Each F2F environment was characterized using video recordings analyzed with the Classroom Observation Protocol in Undergraduate STEM (COPUS) tool. Each individual course showed consistent use of F2F time across each session recording, however, there were significant differences in the predominant student behaviors between courses. Student behavior in two of the courses (Courses Four and Five) was dominated by work in small-groups on problem solving worksheets, in contrast to another course (Course One) where responding to whole-class questioning posed by the instructor dominated students' behaviors. While behaviors in the two remaining courses (Courses Two and Three) included a mix of responding to clicker and whole-class questions, one of them (Course Three) also included large episodes of students simply listening during instructor presentation of material. A mid-semester survey was administered in each course to characterize students' interactions with, and perceptions of, the PCMs. Of particular note, student self-reports of the number of videos viewed and the timing of viewing trended with the amount of peer-to-peer interaction during F2F sessions. That is, students in courses with more consistent groupwork reported watching more of the video content and doing so before the F2F session. These results demonstrate that flipped classrooms can take many forms and suggest that F2F structure may create non-grade-based incentives for PCM utilization.

Graphical Abstract

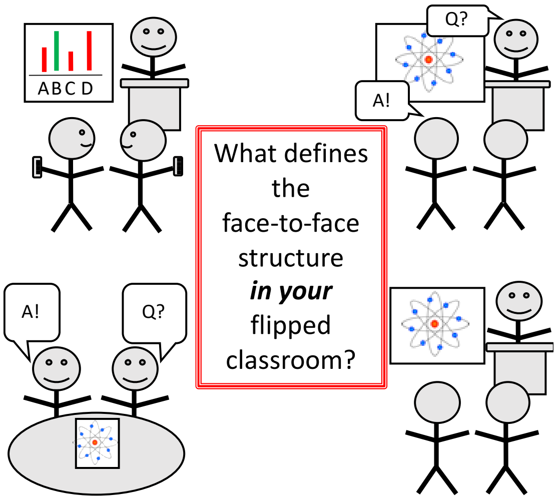

ABC D
Q?
A!
What defines the face-to-face structure ***in your*** flipped classroom?
A!
Q?

Figure 4.1. Graphical abstract for Chapter 4.

Keywords

First-Year Undergraduate / General, Chemical Education Research, Collaborative / Cooperative Learning, Student-Centered Learning

Introduction

A national call from the President's Council in 2012 (President's Council of Advisors on Sciences and Technology (PCAST), 2012) to increase the number of science, technology, engineering, and mathematics (STEM) degrees encouraged instructors to take a closer look at how higher-education STEM classrooms supported student learning. A subsequent report from the National Research Council on Discipline-Based Education Research (DBER) (National Research Council, 2012) and the Freeman et. al. (2014) meta-analysis, further promoted the inclusion of more active learning activities in higher-education STEM classrooms. However, one barrier to including more active learning is the amount of time it takes to include these activities in the classroom (Shadle et al., 2017). To address this barrier, some instructors "flip" their classroom, allowing significant flexibility in how to structure the course. Most flipped classrooms follow two basic tenets: 1) foundational information is delivered to students through pre-class materials (PCMs), and 2) the face-to-face (F2F) environment is used to apply or extend the information through active learning (Bergmann & Sams, 2012; He et al., 2016). Therefore, a "flipped" classroom provides the instructor with flexibility in the type of PCMs provided to the students, as well as the type of in-class active learning conducted in the F2F environment.

Structure of flipped learning environments

Higher-education flipped chemistry classrooms include a variety of different F2F active learning environments, such as Peer-Led Team-Learning (PLTL) (Liu et al., 2018; Mooring et al., 2016; Rein & Brookes, 2015; Robert et al., 2016), Process-Oriented Guided-Inquiry Learning (POGIL) (Canelas et al., 2017; Hibbard et al., 2016), or other combinations of peer instruction, problem-based learning, and/or think-pair-share type exercises (Christiansen et al., 2017; Flynn, 2015). In addition, some classes incorporate case studies (Hill et al., 2019; Rein & Brookes, 2015) or whole-class discussions (Bernard et al., 2017; Blackburn, 2017; Bokosmaty et al., 2019; Ealy, 2013; Ryan & Reid, 2015; Smith, 2013) into their class time. Most incorporate some type of groupwork (Amaral et al., 2013; Bancroft et al., 2019; Christiansen, 2014; Christiansen et al., 2016; Donnelly & Hernández, 2018; Fautch, 2015; Fitzgerald & Li, 2015; Gregorius, 2017; Lenczewski, 2016; Parsons, 2019; Rau et al., 2017; Shattuck, 2016; Weaver & Sturtevant, 2015) and several use mini-lectures or Just-in-Time Teaching to provide clarification on concepts when needed (Amaral et al., 2013; Bancroft et al., 2019; Christiansen, 2014; Christiansen et al., 2016; Christiansen et al., 2017; Eichler & Peeples, 2016; Fautch, 2015; Lenczewski, 2016; Parsons, 2019; Rau et al., 2017; Robert et al., 2016; Ryan & Reid, 2015; Shattuck, 2016). Some provide additional time for these activities by adopting a hybrid structure where students participate in active learning during a specific block of time or day of class (Bokosmaty et al., 2019; Lenczewski, 2016; Liu et al., 2018; Mooring et al., 2016; Rein & Brookes, 2015; Robert et al., 2016).

Regardless of the type of F2F learning environment, flipped classes are structured such that students enter the classroom with some level of understanding, which they build upon during class time (He et al., 2016; Kavanagh et al., 2017; Pienta, 2019). Often the assumption is that this understanding comes from PCMs that students complete before class. Abeysekera and Dawson (2014) noted that a benefit of this tenet of flipped learning is the ability for students to self-pace their learning and therefore manage cognitive load (Clark et al., 2005). Stemming from this early acknowledgement of cognitive load as an underpinning framework to understand the benefits of flipped learning, cognitive load theory (Sweller, 1994) has been used to provide direct support for the design of a flipped course (Mooring et al., 2016). By allowing students the opportunity to use PCMs on their own time and at their own pace, it tends to reduce the information overload burden (Sweller, 1994). In addition, students have the potential to re-watch PCMs, if needed (Casselman et al., 2019). The design of PCMs (Casselman et al., 2019) have also been supported through the cognitive theory of multimedia learning (Moreno & Mayer, 1999). PCMs in most flipped chemistry courses entail students use of some type of online component with dual aspects targeted to both auditory and visual information, with instructor-made videos (e.g., screencasts) being the most common (Bancroft et al., 2019; Casselman et al., 2019; Christiansen, 2014; Christiansen et al., 2016; Christiansen et al., 2017; Ealy, 2013; Fautch, 2015; Flynn, 2015; Hibbard et al., 2016; Mooring et al., 2016; Parsons, 2019; Ranga, 2017; Reid, 2016; Rein & Brookes, 2015; Robert et al., 2016; Ryan & Reid, 2015; Shattuck, 2016; Smith, 2013; Weaver & Sturtevant, 2015; Webber & Flynn, 2018; Woodward & Reid, 2019). However, other online resources, such as Khan Academy videos (Liu et al., 2018) or interactive online modules (Gregorius, 2017) have also been provided as PCMs. Further, some flipped classes include multiple formats of PCMs, such as videos, screencasts, interactive modules, animations, podcasts, etc. (Amaral et al., 2013; Bernard et al., 2017; Bokosmaty et al., 2019; Canelas et al., 2017; Eichler & Peeples, 2016; Fitzgerald & Li, 2015; Pilcher, 2017).

Student utilization of pre-class materials (PCMs)

Instructors may provide incentives to encourage students to complete PCMs on time; for example, checking that students have taken required notes on the PCMs

(Hibbard et al., 2016; Parsons, 2019; Shattuck, 2016). Many utilize quizzes administered either in-class (Christiansen, 2014; Christiansen et al., 2016; Christiansen et al., 2017; Fitzgerald & Li, 2015; Smith, 2013; Woodward & Reid, 2019) or out-of-class as part of the required PCMs (Amaral et al., 2013; Bancroft et al., 2019; Bokosmaty et al., 2019; Christiansen et al., 2016; Ealy, 2013; Eichler & Peeples, 2016; Fautch, 2015; Flynn, 2015; Mooring et al., 2016; Pilcher, 2017; Robert et al., 2016; Weaver & Sturtevant, 2015). The purpose and difficulty of quizzes varies across studies, with some descriptions emphasizing that quizzes are scored primarily for completion points and/or formative assessment (Bancroft et al., 2019; Bokosmaty et al., 2019; Christiansen, 2014; Fautch, 2015; Weaver & Sturtevant, 2015; Woodward & Reid, 2019) or that students are given an opportunity to discuss answers before submitting (Smith, 2013). Regardless of the type and level of incentive, PCMs are provided to students to acquire initial content understanding, which can be built upon during F2F activities. Thus, determining "if", "how", and "when" students are using the assigned PCMs may be important for characterizing the success of a flipped learning environment. However, not all studies on flipped chemistry courses include this type of information. Some have determined "if" students completed the PCMs by keeping track of who turned in problem-solving activities (Lenczewski, 2016), by asking students if they watched the required videos for that day (Woodward & Reid, 2019), or by assessing students' performance on quiz questions based on non-conceptual topics embedded in the videos (Christiansen, 2014). Some studies include self-reported student estimates of time spent studying outside of class each week (Gregorius, 2017) or specifically watching the videos (Shattuck, 2016). To gain information about "how" and "when" students interact with PCMs, more detailed survey questions have been administered to students asking about the frequency and nature of video use and of the videos' perceived usefulness, length, quality, and impact on the class (Smith, 2013). Details about "if", "how", and "when" students interact with PCMs has also been reported through the use of analytics data collected from student access tracking (Bancroft et al., 2019; Mooring et al., 2016; Parsons, 2019; Ranga, 2017; Shattuck, 2016; Weaver & Sturtevant, 2015) or other video tracking data (Mooring et al., 2016; Ranga, 2017; Shattuck, 2016). In addition, analytics data has been used to provide information about "if" and "when" students were re-watching videos (Mooring et al., 2016; Shattuck, 2016).

Students' use of PCMs can vary greatly from class to class. For example, Lenczewski (2016) noted that "*about 33% of the class completes 80% or more of the assignments per semester*" and Bancroft et al. (2019) reported that although about 98% of students watched the assigned videos before class at the beginning of the term, this number dropped to around 85% by the end. Mooring et al. (2016) reported that 80 – 95% of students watched the assigned videos and completed the incentive quiz, with Woodward and Reid (2019) noting that they only obtained 79% viewership without regular email reminders. Thus, students make use of the PCMs to varying degrees. Therefore, depending on the structure of the F2F environment, students may enter into learning activities with insufficient understanding to fully participate.

Purpose and Rational

This project is part of a larger study on flipped learning environments and their impact on student motivation and performance. The first phase of this coordinated, multi-institution study involves evaluating 1) the structure of each F2F learning environment and 2) students' use and perceptions of the PCMs that support these environments.

In a recent meta-analysis, Rahman and Lewis (2019) reported on the effectiveness of a range of evidence-based instructional practices (EBIPs), including flipped learning. Their analysis of fifteen flipped learning studies revealed 'trivial to medium' effect sizes. In addition, they determined how the effect sizes for an EBIP were moderated by other factors, such as the type of assessment used and course size, and noted that an additional moderator for flipped learning may be variation in how the environment is structured. In acknowledgement that flipped structures vary, Seery (2015b) noted that in some studies "*there is a vagueness about what happens during class time, and a more robust framework needs to be developed so that there is a basis for what happens in class time and how it builds on pre-lecture work.*" The call for more robust frameworks of EBIPs, in general, extends beyond Seery's specific call. In 2016, Stains and Vickrey noted that the findings of EBIP studies are compromised if factors related to implementation cannot be accounted for. To address this, they proposed the use of a fidelity of implementation (FOI) framework. FOI studies have taken many forms within discipline-based education research (see examples in Stains and Vickrey (2017)), perhaps the most salient form for flipped learning is the investigation of how course structure impacts outcomes. While many flipped chemistry course studies provide descriptions of the F2F environment and the type of active learning activities employed during class time, few report observational data for consistent cataloguing of the structure. Of those that have, Canelas et. al. (2017) categorized the percentage of classroom time that students spent participating in active learning and Donnelly and Hernández (2018) used the Behavioral Engagement Related to

Instruction (BERI) protocol to determine the percentage of students who were behaviorally engaged throughout class time. As there is no single prescribed structure for the F2F portion of a flipped classroom and the positive impacts of active learning (e.g., higher exam scores, lower failure rate, etc.) have been linked to students being more engaged with course activities (Wieman, 2014), it is important to understand the degree to which a given F2F structure is student-centered and the nuanced differences that exist. Therefore, we conducted consistent evaluations of each flipped learning environment such that the structure of their F2F settings can be accounted for when comparing course outcomes.

Recent studies on active learning have noted the association of student buy-in with self-regulated learning and course performance (Cavanagh et al., 2016), and how students' perceptions may influence their buy-in to the learning environment (Brazeal & Couch, 2017; Cavanagh et al., 2016). Cavanagh (2016) notes that "*active learning provides students with opportunities to engage in the learning process, and students may decide to participate based on a series of judgements,*" one of which is whether they view an activity as valuable to their learning. These judgements have been shown to be influenced by the classroom climate (Eddy & Hogan, 2014; Freeman et al., 2007), and students may change their initial expectations within a course based on teaching practices and course demands (Brown et al., 2017). Within flipped chemistry courses, lack of engagement with PCMs may undermine the potential benefits of the learning environment (He et al., 2016; He et al., 2019). Incentivizing timely interaction with the PCMs have been noted to influence students' utilization of them (Christiansen et al., 2016; Woodward & Reid, 2019). In addition, the degree of a course's structure has been attributed to increased utilization of PCMs (Eddy & Hogan, 2014). In defining "low-", "moderate-" and "high-structure" courses, one parameter used by Eddy et. al. (2014) was "in-class engagement". This parameter was broken down by the amount of student-talk occurring during an F2F session through the use of activities such as clicker questions, worksheets, or case studies. By this parameter, a "low" course was defined as having student-talk for <15% of F2F time, "moderate" courses as 15 – 40%, and "high" courses as >40% of F2F time. Other parameters used when defining the degree of a course's structure included the frequency of the preparatory and review assignments. While their study did not include any high-structure courses, when comparing students' behaviors and perceptions between courses with low- and moderate-structure, Eddy et. al. (2014) found that students in the moderate-structure course reported studying for more hours per week, completing recommended reading before class, and perceiving the preparatory work to be more important. The authors believe that the increased course structure led to more "accountability" on the students' part for their learning, thereby improving course performance in this as well as other studies (Freeman et al., 2011). These outcomes are in line with the ability for classroom environments to stimulate the development of self-regulated learning (Zimmerman, 1995) and that self-regulated learners are empowered to create goals, use strategies, and implement actions to meet their goals (Ridley et al., 1992).

Given the many ways in which a flipped learning chemistry course can be structured, and to move beyond a more general description of flipped learning as being defined as providing PCMs and more active learning during F2F sessions, this study employed a coordinated set of evaluation techniques (i.e., classroom observations and a survey of students' use and perceptions of the PCMs) across multiple flipped chemistry courses to answer the following research questions:

1) What are the predominant student behaviors and instructional styles for the F2F settings within these environments?
2) What are the students' self-reported use and perceptions of the PCMs in these environments?
3) What associations exist between instructional style in the F2F setting and PCM utilization?

Methods

Course Descriptions

Flipped chemistry courses from five institutions across the United States were involved in this study, none of which were the authors' home institutions. Institutions varied by size, type, acceptance rate, and demographic profile (Table 4.1). Four of the institutions were four-year public research universities and the fifth was a two-year community college. These data collection sites were selected based on the corresponding author's knowledge of who the instructors were and that they were not new to course flipping. Each invited instructor had a minimum of two years of experience in flipping their course and was the primary person involved in developing the course materials. The general structure of each course followed the two basic tenets of flipping: 1) foundational information was delivered to students through pre-class materials (PCMs), and 2) the face-to-face (F2F) environment was utilized for the application or expansion of the information through active learning (Bergmann & Sams, 2012; He et al., 2016).

Both General and Introductory Chemistry courses were included in the study. Multiple sections of each course, taught by the same instructor, were combined in the datasets (Table 4.1). Course schedules and settings varied, the most notable of which is

Course Five, which took place only once per week (for 75 min) in a fully collaborative space. The remainder of the courses were held on more traditional 2- or 3-day per-week schedules in either standard fixed seating lecture halls or ones designed with rotating chairs to promote collaborative work.

Table 4.1. Course and institution details.

Course	**One**	**Two**	**Three**	**Four**	**Five**
Type	General I	Introductory I	General II	General I	General I
Cycle	On sequence	On sequence	On sequence	Off sequence	On sequence
Sections	1	2	1	2	2
[a]Enrollment	200	72	281	360	171
Schedule	75 min, 3 times per week, morning	50 min, 3 times per week, morning	80 min, 2 times per week, evening	80 min, 2 times per week, morning	75 min, once per week, afternoon
Setting	Auditorium style – rotating chairs	Auditorium style – fixed chairs	Auditorium style – rotating chairs	Auditorium style – rotating chairs	Collaborative space – circular tables of 9
Institution					
Region	Southeast	Southwest	Southwest	Northwest	Midwest
Size (Approx.)	55,000	10,000	35,000	30,000	50,000
Type	Four-year, Public, Doctoral – Very High Research Activity	Two-year, Public, Associate’s College – Mixed Transfer/ Career & Technical	Four-year, Public, Doctoral – Very High Research Activity	Four-year, Public, Doctoral – Very High Research Activity	Four-year, Public, Doctoral – Very High Research Activity
Acceptance	50%	100%	30%	78%	45%
[b]Demographics	Asian – 4% Black – 13% Hispanic – 60% White – 13% Other – 7%	Asian – 9% Black – 5% Hispanic – 25% White – 54% Other – 7%	Asian – 27% Black – 4% Hispanic – 12% White – 39% Other – 18%	Asian – 7% Black – 1% Hispanic – 9% White – 61% Other – 22%	Asian – 9% Black – 4% Hispanic – 3% White – 65% Other – 19%

[a]Total enrollment across all sections at the start of the term. [b]’Other’ category can include designations of International, Pacific Islander, 2+ ethnicities, and/or other designations inconsistently reported across institutions.

Pre-Class Material (PCM) Description

Within each course, instructors assigned videos that corresponded to the content of each F2F class day. With the exception of Course One, in which the instructor curated relevant online videos from different sources, courses used instructor-created content videos (Table 4.2). Each instructor noted that their number and length of videos varied by topic across each term. Students completed an online quiz that covered the related video content prior to each F2F day in Courses One, Three, and Four. In Course Two, each F2F day started with a clicker quiz based on the video content. No grade-based viewing incentive was utilized in Course Five.

Table 4.2. Topics observed and video content details for each participating course.

Course	Topics observed	Video type	[a]Number of videos assigned	Length range (minutes)	Majority length range (minutes)	Viewing Incentive
One	Periodic trends	Instructor curated online videos	2	6 – 11	6 – 11	Online quiz
Two	Oxidation numbers, activity series, and types of reactions	Instructor created screencasts	11	3 – 12	5 – 10	Clicker quiz
Three	Intermolecular forces, phase diagrams, and solid structures	Instructor created screencasts	10	2 – 17	5 – 10	Online quiz
Four	Lewis structures and formal charge	Instructor-created screencasts and recorded chalk-talks	4	11 – 25	15 – 20	Online quiz
Five	Electron configuration, periodic trends, and bonding	Instructor-created screencasts and recorded chalk-talks	9	15 – 44	20 – 25	None

[a]Each course was observed on two consecutive class periods, these numbers correspond to the total across the observation days.

Within each participating course, two modes of data collection were employed: 1) observations of the F2F learning environments, and 2) online surveys administered to enrolled students. As the two data sources have different data collection protocols, the remainder of the methods section, as well as the results section, will be organized by those two sources. All data collected within this study was approved by the Institutional

Review Board (IRB) at Portland State University and appropriate consent was acquired from instructors and students as required by the IRB.

Part 1: Structure of Face-to-Face (F2F) Learning Environment

Course Observations

Within each course, F2F sessions were recorded by a member of the research team on two consecutive class meetings. Recordings took place midway through a term on a non-exam week. Recordings were captured from the back of the room in a location where a large swath of the students could be seen as well as any primary location for the instructor (e.g., lectern or board). Throughout a recording, the researcher could zoom in-out and pan the camera to capture both instructor and student behavior as needed. The associated audio captured the majority of the instructor's talk, with the exception of their close-contact conversations with individuals or groups. Additionally, students' whole-class questions were captured but close-contact group conversations were not. Due to the logistics of planning on-site data collection, instructors were aware of which days they would be observed. With the variety of courses and the timing of site-visits, we were unable to observe the same content coverage across courses.

Coding Protocol

The Classroom Observation Protocol in Undergraduate STEM (COPUS) (Smith et al., 2013) was developed for use in higher education STEM, making it an appropriate choice for our needs. The protocol consists of codes used to document the real-time behaviors of both instructors and students during an F2F classroom session. Analysis of COPUS codes has been used to determine different instructional styles from lecturing and Socratic to peer instruction and collaborative learning (Lund et al., 2015) and to establish instructional profiles called COPUS Profiles (Stains et al., 2018). These profiles range from “didactic”, where more than 80% of the F2F time is used for lecturing, to “interactive lecture”, where the use of student-centered strategies emerges, and finally to “student-centered”, where the use of groupwork dominants the F2F time (Stains et al., 2018).

Each video was coded by two researchers using the COPUS. Each researcher was trained following the recommendations provided by the COPUS developers (Smith et al., 2013). After independently coding the first video for a given course, the coders met to compare their consistency in code use. This initial meeting provided clarity to the coders on how the various COPUS categories were defined respective to the course being observed. If observed behaviors in a course did not clearly align with code definitions (Table A.1), the coders discussed and reached consensus on if or how the definition applied. Details on the code descriptions can be found in Appendix A. After these discussions, the coders recoded the first video before moving to the second. This coding process produced Cohen’s kappa scores >0.85 for each video, indicating high inter-rater reliability (Cohen, 1960).

In addition to reviewing the individual COPUS categories, the data from each course were entered into the COPUS Analyzer (Stains & Harshman). This tool was developed during a large national study of STEM teaching practices (Stains et al., 2018) and provides COPUS Profiles for a given course based on a reduced set of observation data (4 instructor and 4 student codes). The COPUS Analyzer matches individual course data to one of seven clusters generated from a latent profile analysis conducted during the national study. Clusters are labeled as representing a various instructional style, either didactic lecture, interactive lecture, or student-centered. As described by the analyzer developers (Stains et al., 2018), a didactic lecture style *"depicts classrooms in which 80% or more of class time consists of lecturing,"* whereas an interactive lecture style represents *"instructors who supplement lecture with more [compared to didactic instruction] student-centered strategies…such as clicker questions with groupwork.*" The student-centered style "*depicts instructors who incorporate student-centered strategies into large portions of their classes.*" In this study, clusters for a given course did not vary across observation days.

Part 2: Students' Perceptions of Pre-Class Materials (PCMs)

Survey Development

The survey was developed during previous semesters through an iterative process that included two rounds of focus groups and one round of a pilot survey that included open-ended responses. This process informed the wording of items and response options. The final version of the survey was created such that each item contained multiple response options where students could choose a single option or multiple options, depending on the item type. Open response boxes were provided only if an 'Other' response was selected for an item. Brief descriptions of the development process are outlined in Appendix A.

Survey Administration

Survey participants were students recruited from the flipped courses to participate in the study. Survey deployment in each course was coordinated to take place midway through the term during a non-exam week. The instructor was provided a brief script to make an initial in-class announcement regarding the survey. A note similar to the script was posted on the classroom management platform of each course. Students who were interested in participating clicked on a link to the Qualtrics survey that was part of the announcement note. Some instructors offered a nominal amount of extra-credit points for accessing the survey.

Analysis

Data collected from the final version of the survey were cleaned to remove duplicate entries and any students who did not consent. For single-response items, the percentage of students selecting a given response was determined with responses to these items totaling to 100%, although there was some variability due to rounding. For multi-response items, the percentage of students choosing each response was determined. As students could select multiple options, responses are not mutually exclusive and may not total to 100%. As skip logic steps were built into the survey flow based on a student's response, not all students were presented each item. Therefore, the total number of student responses does not remain constant across each item within a course.

Differences in responses across courses for single-response items were analyzed using chi-square tests. However, as some of the possible options had a low number of student responses, Fisher's exact test was often used to determine significance. Fisher's exact test is considered more appropriate than a chi-square test when 25% of the cells of a contingency table have expected counts below 5 and a minimum expected count below 1 (Mayers, 2013). Both chi-square and Fisher's exact tests were calculated using the stats package in the statistical software R (version 3.6.2). Cohen's *w* (Cohen, 1992), a measure of effect size, was calculated using the rcompanion package in R. General guidelines for Cohen's *w* suggest a small effect size for values around 0.1, medium around 0.3, and large around 0.5 (Cohen, 1992). Multi-response items were also analyzed with chi-square and Fisher's exact tests. Each response option in a multi-response item was treated as a single yes-no item to create the contingency table for that option, where students who selected it were counted as 'yes' and students who did not select it were counted as 'no'. Therefore, each option within each multi-response item was analyzed as a separate contingency table to detect differences across courses. Post-hoc pairwise comparisons were only conducted when response patterns matched observed differences between course structures or features of their supporting elements (i.e., all possible pairwise comparisons were not run to hunt for significant differences). When pairwise comparisons were conducted, Fisher's exact test with a Bonferroni correction was calculated using the fmsb package in R. A significance cutoff of $p < 0.05$ for all pairwise comparisons was used unless otherwise noted.

Results and Discussion

Part 1: Structure of Face-to-Face (F2F) Learning Environment

To answer the first research question (*What are the predominant student behaviors and instructional styles for the F2F settings within these environments?*), we examined the in-class learning resources put in place by the instructors and what those meant in relation to what students were asked to "do" in the F2F sessions. The predominant student behaviors and instructional styles identified at this stage were compared to those from the COPUS analyzer to discuss the course-to-course trends. The observations for each course, which included the COPUS codes and details of the learning activities, are compiled in Appendix A. The COPUS codes were compiled into a timeline that allows for the visualization of when a code was observed and for how long during each observed day of instruction, providing insight to the dynamic structure of each F2F session (Figures A.1 – A.5).

Based on the observation of the five courses, three primary in-class resources were noted. The instructors in Courses One, Four, and Five, employed worksheets that contained problems and guiding information. The sheets were used to facilitate the majority of the F2F time and students documented their responses on the sheets. In contrast, the Course Two instructor framed the majority of the F2F time around series of clicker questions. These were used at the start of the F2F session to gauge student understanding of prior material and then throughout the session to provide real-time feedback as new content was introduced. The Course Three instructor used a combination of resources including prepared lecture slides, clicker questions, and, during one of the days, a 'game'. Across both observation days, the majority of the F2F time was direct instruction from the prepared slides with blocks of clicker questions at the end of a module to provide formative feedback. While one of the observation days in Course Three included a group activity in the form of a game, the instructor noted that this was a deviation from the typical F2F practice. Given the flexibility in how the F2F portion of a flipped class can be supported, the range of resources utilized is not surprising. The next step was to look at how these resources were implemented and what student behaviors resulted.

A timeline of all student and instructor COPUS codes for each observation day in each course is presented in Figures A.1 – A.5. The student codes can be grouped into behaviors where students are 'receiving' information, conducting 'groupwork', doing 'individual work', engaged in 'questioning', or doing 'non-work' (see Table A.1 for code groupings). To compare and contrast what students were *doing* across these F2F sessions, we focused on the average amount of groupwork and questioning observed (Figure 4.2). With respect to groupwork, on average, more than 75% of the F2F time in Courses Four and Five included observations of students working with their peers on the instructional worksheets (WG code). In both courses, students sat in working groups of 4 – 5 from the start of class. In contrast, the students in Course One were observed to be engaged with their peers on the worksheets in less than 35% of the F2F time, forming groups of 2 – 3 each time they were directed to answer worksheet questions. Therefore, use of the same resource looked very different with regard to students' peer-to-peer interactions over an entire F2F session. Clicker (CG) and other (OG) groupwork in Courses Two and Three was observed during roughly 50% and 40% of the F2F time respectively. Similar to the peer-to-peer interactions in Course One, clicker question discussion typically included only 2 – 3 students. The game in Course Three, coded as OG, involved groups of 3.

The questioning category in Figure 4.2 includes observations of students answering questions posed by the instructor (AnQ) or asking a question (SQ) to the instructor. These codes apply to whole-class questions not to any questions or answers during instructor facilitation of groupwork. These types of student behaviors were observed in over 80% of the F2F time in Course One and over 50% in Course Two. In both courses, the instructor used the materials (worksheet or clicker questions respectively) to introduce content and then ask whole-class questions. In Course Two, there was a near equal balance of students responding to as well as asking whole-class questions, often times overlapping within a 2-minute time-block (displayed as AnQ & SQ in Figure 4.2). Student questions were infrequently observed in all other courses. In

Courses Three and Four student responses to whole-class questions were observed during less than 20% of the F2F time and less than 10% in Course Five. These student behaviors were spread out in Courses Three and Five and only present during activity wrap-up sessions in Course Four.

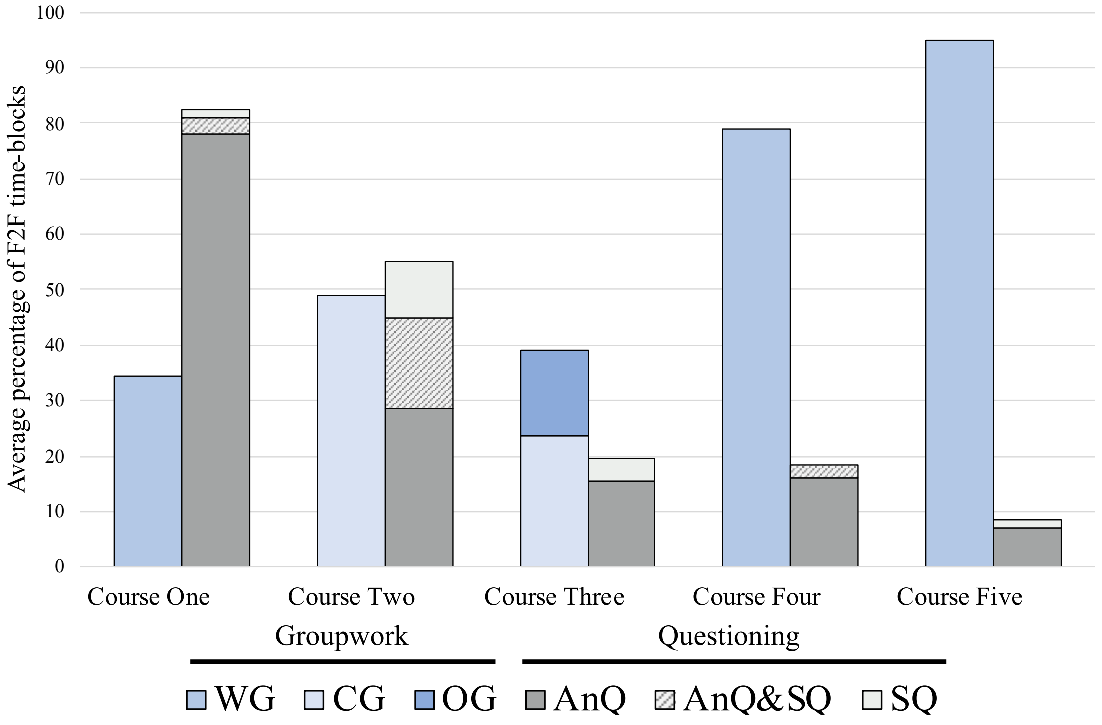


Figure 4.2. Average percentage of F2F time that students were observed doing ‘groupwork’ (blue bars) or ‘questioning’ (grey bars). The groupwork COPUS code category includes independent observations of ‘worksheet’ (WG), ‘clicker’ (CG), and ‘other’ (OG) activities. The questioning category includes independent observations of students ‘answering’ (AnQ) or ‘asking’ (SQ). The combined category (AnQ & SQ) notes that these two codes overlap within the time-blocks. See Table A.1 in Appendix A for the details of each COPUS code.

The observed student behaviors in Figure 4.2, averaged across observation days, indicate that the predominate teaching strategy in Courses One and Two was Socratic lecturing. As noted by Freeman et. al. (2011), “*Socratic lecturing involves the frequent use of questions posed to the class*” with intent to “*engage student attention and provide feedback to the instructor.*” Course Two was balanced by the use of clicker-questions.

The intent of this type of learning strategy is to develop student thinking and the application of knowledge (Freeman et al., 2011). Based on the Figure 4.2 groupings, the predominant teaching strategy employed in Courses Four and Five was small-groupwork on worksheets. The intent of this learning strategy is to provide students with hands-on practice with problem solving and conceptual understanding (Freeman et al., 2011). As indicated by the Course Three data in Figure 4.2, neither of these student behaviors dominated the F2F time. In this course, student listening, while the instructor delivered course content from prepared slides, was also a frequently observed behavior.

COPUS Analyzer Classification of F2F Structures

The COPUS Analyzer (Stains & Harshman) utilizes the WG, CG, OG, and SQ student codes and the lecture (Lec), posing questions (PQ), clicker questions (CG), and one-on-one discussion (1o1) instructor codes (Table A.1 and Figures A.1 – A.5) to generate clusters and associated instructional styles. When entered into the COPUS Analyzer, the observation data for each course produced three different clusters representing two different instructional styles (Table 4.3). The clusters for each course were consistent across the observation days.

Table 4.3. COPUS analyzer (Stains & Harshman) clusters and related instructional styles.

Course	One	Two	Three	Four	Five
Cluster	6	4	4	6	5
Style	student-centered	interactive lecture	interactive lecture	student-centered	student-centered

Taken together, the COPUS timelines (Figures A.1 – A.5) and analyzer clusters (Table 4.3) provide different levels of resolution for classifying each course’s F2F structure (Reisner et al., 2020). At a lower level of resolution, the identified clusters indicate the category to which each structure belongs (i.e., interactive lecture or student-centered). At a higher level, the timelines (Figures A.1 – A.5) and aggregated student codes (Figure 4.2) provide additional insights to each F2F environment. In reviewing the timelines from Courses One and Four (Figures A.1 and A.4), in the light of the analyzer output (Table 4.3), it is noted that courses within the same cluster (i.e., cluster 6) can vary greatly in students' behaviors. For example, students in Course One consistently answered questions (AnQ) posed by the instructor to the whole-class and less frequently worked in groups (WG) (Figure 4.2). Whereas in Course Four, students worked consistently in groups (WG) and only answered questions (AnQ) at the end of the F2F session, during the activity wrap-up (Figure A.4). Therefore, despite their similar cluster groupings and defined style, students in these courses were observed to be interacting with their peers to very different levels. In contrast, the Course Five analyzer data also resulted in the student-centered instructional style but was associated with cluster 5. In considering their peer-to-peer interactions during F2F time, students in Course Five were most similar to those in Course Four. Therefore, while the COPUS analyzer and our aggregated student behaviors utilize the same data, they provide complementary information at different levels of detail for the study. It is important to note that these observations were from a two-day snapshot of each course within the overall term, thus conclusions are drawn based upon these observed days.

When considering peer-to-peer interactions during F2F time, groupwork dominated what students were *doing* in Courses Four and Five compared to the other three courses. This difference in the amount of peer-to-peer interactions is reflected upon when discussing some of the survey responses in the next section.

Part 2: Students' Use and Perceptions of Pre-Class Materials (PCMs)

The final version of the survey was given in all five courses. The student response rate ranged from 19% to 85% (Table 4.4). Although Course Five had a low response rate, it was still included in the survey analyses as support for trends seen in Courses One through Four. Response rates were determined based on week-1 course enrollments (Table 4.1) and therefore may not be reflective of the true rates at the time of survey administration. Students saw certain survey items based on their previous responses to items using skip logic and, as such, not all students were presented each item. Therefore, the number of students in each course that responded to each item is provided in each results table.

Table 4.4. Survey participation by course.

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Survey responses, n (%)	65 (36)	34 (57)	240 (85)	278 (84)	59 (19)

Do Students Watch (and Re-Watch) the Videos?

As each instructor regularly assigned videos related to the course material for each F2F session, the first survey item presented to students was, "**How many of the assigned videos have you watched?**" As can be seen in Figure 4.3, response distributions varied by course, from a majority responding that they watched 'Most' or 'Some' of the videos (Courses One and Two), to a majority of respondents noting that they watched 'All' the videos (Courses Four and Five). A chi-square test was used to determine whether there was a significant difference between the proportions of student responses by course. The result showed there was a statistically significant difference ($\chi^2(12) = 144.169$, $p < 0.001$); the Fisher's exact test also showed a statistically significant difference (p < 0.001). The value of Cohen's *w* (*w* = 0.46), suggested that this difference represents a large effect.

As course structure has been noted to influence student behavior (Cavanagh et al., 2016; Eddy & Hogan, 2014) and the dominant student behavior varied across courses (Figure 4.2), post-hoc pairwise comparisons were conducted to detect where course-to-course differences were significant. These results showed that there was no significant difference between Courses One and Two or between Courses Three and Four. However, all other pairwise comparisons showed a significant difference, with effect sizes ranging from small (*w* = 0.20 between Courses Four and Five) to large (*w* = 0.74 between Courses One and Five). Values for all significant pairwise effect sizes are included in the description of Figure 4.3. The difference between courses in the percentage of students who selected that they watched 'All' the videos suggests that students' viewing behaviors may trend with the structure of the in-class environment. That is, students in courses where the predominant student behavior was responding to instructor-posed whole-class questions were less likely to report watching all of the videos provided as PCMs compared to those in courses dominated by peer-to-peer interactions during groupwork. Viewing differences are also noted to align with the frequency of F2F sessions (Table 4.1), where Courses One and Two meet for F2F sessions three-times-per-week compared to meeting twice-weekly (Courses Three and Four) and one time only (Course Five). Viewing differences are not explained by differences in point-based incentives, as Course Five is the only one that does not provide such an incentive (Table 4.2).

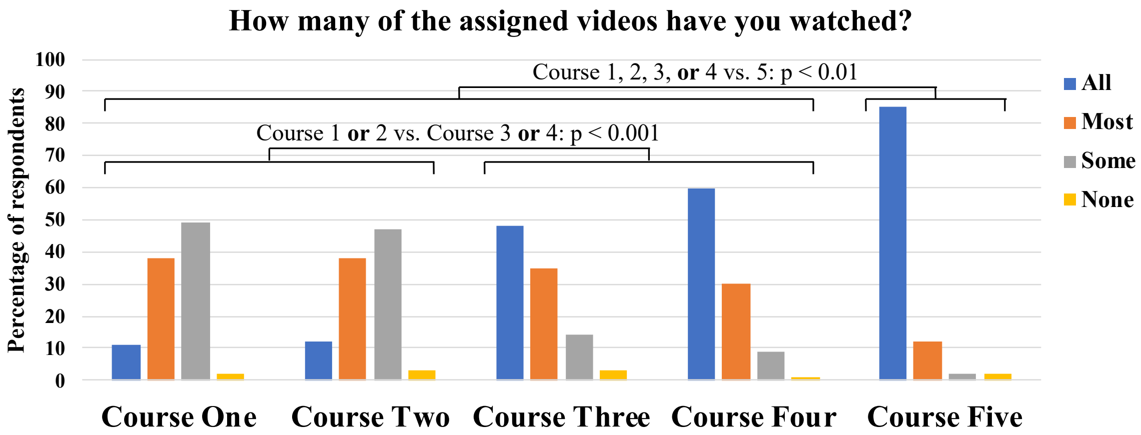


Figure 4.3. Percentage of respondents in each course categorized by viewing frequency. [a]Significant pairwise comparisons between Course Five and Courses One (w = 0.74), Two (w = 0.71), Three (w = 0.29), or Four (w = 0.20). [b]Significant pairwise comparisons between Courses One or Two and Courses Three (w = 0.31 and w = 0.24, respectively) or Four (w = 0.39 and w = 0.30, respectively).

Students who reported watching ‘Most’ or ‘Some’ of the videos were directed to respond to the item, “**Why have you not watched all of them?**” The responses from each course are provided in Table A.5, found in Appendix A. Response options to this item were categorized into ‘General excuses’, ‘Not helpful’, and ‘I prefer other [types of resources]’. Overall, student responses to many of the options were fairly consistent across courses with a few that differed significantly. Post-hoc pairwise comparisons did not reveal any notable trends based on course structure. One potential trend of interest was based on the properties of the videos themselves. For example, Courses Four and Five had videos with longer individual run times (15+ minutes) than the other courses (Table 4.2). Student responses to the option ‘they [the videos] are too long’ (within the General excuses category of Table A.5) differed significantly ($p < 0.001$, $w = 0.24$) with Courses Four and Five having higher percentages of students who selected this option. However, pairwise comparisons only supported a response difference for Course Four compared to Courses Two and Three, as few students in Course Five were presented with this item based on skip logic.

Because students in each course had access to the videos for the entire term after they were posted, those that watched at least 'Some' of them were also asked to respond to the item, "**Have you ever watched a video (or part of a video) more than once?**" Overall, a large percentage of students from each course responded that they did watch a video (or part of a video) more than once (Figure 4.4). These results are similar to a previous study that found students reported watching each pre-lecture video approximately three times on average (Smith, 2013). When a chi-square test was conducted, there was no statistically significant difference across courses ($\chi^2(4) = 7.8755$, p = 0.096).

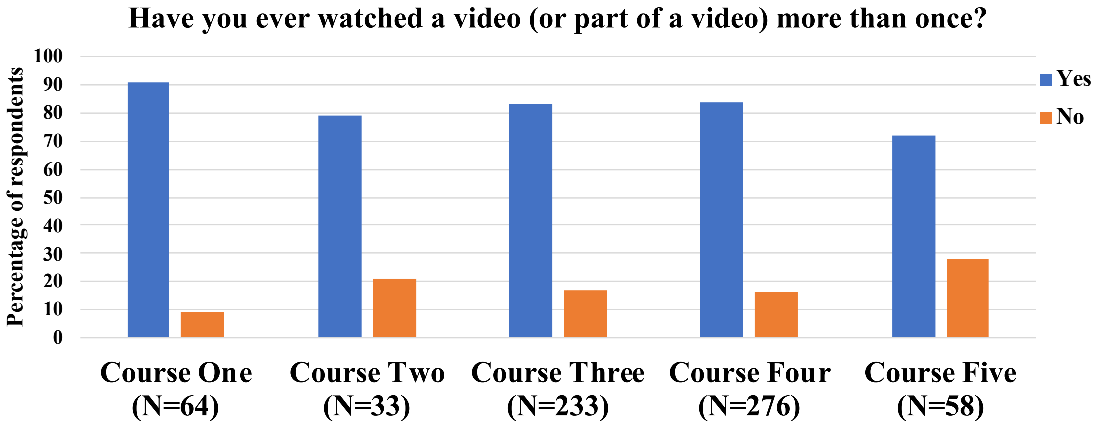


Figure 4.4. Percentage of respondents in each course categorized by re-watching.

Students who responded that they had never re-watched a video (Figure 4.4) were asked to select reasons why they had not (Table A.6, Appendix A). While the numbers of students who were presented with this follow-up item was low, the majority of respondents in most courses selected the response 'I refer to the notes I take the first time I watch'. The exception to this were the majority respondents from Course One who selected 'I watch other videos to get a different perspective than the ones posted', which may be a result of the video properties itself, since Course One was the only course not supported by instructor-created screencasts (Table 4.2).

When Do Students Watch the Videos?

All students who responded that they watched at least 'Some' of the videos (Figure 4.3) were asked, "**When do you typically watch the videos for the first time?**" Students could only select the one option that best represented when their first viewing occurred. The percentage of student responses to this item for each course is presented in Table 4.5. Both chi-square ($\chi^2(16) = 171.410$, $p < 0.001$, $w = 0.51$) and Fisher's exact tests revealed a statistically significant difference by course ($p < 0.001$). Post-hoc pairwise comparisons between the number of students who responded that they watched the videos for the first time 'BEFORE the material is covered in class or on the homework' showed statistically significant differences with varying effect sizes (Table 4.5). As a result, courses dominated by instructor-peer interaction during the F2F time (i.e., Courses One and Two), as identified by the COPUS data (Figure 4.2), had lower percentages of students who reported watching the videos for the first time 'BEFORE the material was covered in class or on the homework' when compared individually to courses with more F2F time spent in peer-to-peer interactions (Courses Four and Five). Student responses on initial viewing in Courses One and Two was split between before and after material coverage and when struggling on homework. While, to our knowledge, no earlier studies have explicitly asked students about when they viewed the videos, some have tracked behaviors through their hosting platform (Bancroft et al., 2019; Mooring et al., 2016; Parsons, 2019; Ranga, 2017; Seery, 2015a; Shattuck, 2016; Weaver & Sturtevant, 2015). For example, with this type of tracking information, Seery (2015a) reports viewings that were higher during the evening prior to an F2F activity and Ranga (2017) reports relatively low views across the semester with spikes just prior to exams. Additionally, Bancroft et. al. (2019) report a steady decline in PCM completion over a 14-week term.

Table 4.5. Response percentages, by course, to survey item "When do you typically watch the videos for the first time?"

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	64	32	233	276	58
[a]**When do you typically watch the videos for the first time? Choose the BEST option.**					
Response options	Percentage of student responses to item				
BEFORE the material is covered in class or on the homework	36[b]	28[b]	74[c]	90	84
AFTER the material is covered in class or on the homework	20	34	12	4	5
When I don't understand something on a homework problem	33	28	7	2	0
When I start studying for an exam	11	9	7	3	10

[a]Survey item presented to everyone except those who selected 'None' for initial question, shown in Figure 4.3. [b]Significant pairwise comparisons ($p < 0.05$) between Courses One or Two and Courses Three (w = 0.33 and w = 0.32, respectively), Four (w = 0.54 and w = 0.49, respectively), and Five (w = 0.52 and w = 0.56, respectively). [c]Significant pairwise comparison ($p < 0.05$, w = 0.22) to Course Four.

To further understand the timing of students' viewing, those who had responded to re-watching the videos (Figure 4.4) were also directed to the item, "**When do you re-watch videos?**" Response options to this item and student selections are presented in Table A.7 (Appendix A). Generally, more students responded that they would re-watch parts of a video than those that responded they would re-watch an entire video. The majority of responses across all courses for re-watching only part of a video were, 'when I have missed something the first time' and 'when I need clarification at a later time (e.g., for homework or when completing a lab)'. Lower percentages reported re-watching 'when studying for an exam'. Although some of the response options had significantly different percentages of students that selected them, no obvious trends were seen in the responses based on the F2F structure.

How Do Students Interact with the Videos?

Students who regularly watch the videos can interact with them in different ways. To assess students' interactions with the videos, those that watched at least 'Some' of them (Figure 4.3) were asked, "**When you watch the videos, how do you watch them?**" Percentages of student responses to this item for each course are included in Table A.8 (Appendix A). Response options were categorized into 'Pacing of viewing' (e.g., pausing and/or rewinding) and 'Blocking of viewing' (e.g., all in one sitting). Overall, although there were some significant and non-significant differences between courses for the 'Pacing of viewing' options, no notable trends were present. The 'Blocking of viewing' options showed significant differences between courses. Specifically, Courses Four and Five had larger percentages of students who selected 'I watch the assigned videos in one sitting' instead of 'I spread out watching the assigned videos throughout the day or week.' Post-hoc pairwise comparisons showed a significant difference in the number of students who selected this option from Courses One, Two, and Three when compared to Courses Four and Five, ($p < 0.001$) with effect sizes ranging from small to large (details provided in Table A.8). Although these differences could be influenced by the increased peer-to-peer interaction in Courses Four and Five, it could also be affected by the longer video lengths (Table 4.2) or the lower meeting schedule (1 – 2 times per week vs. 3 times per week) of those courses (Table 4.1).

Since the videos are meant to provide information in lieu of traditional lecture, simply watching a video without doing anything else could be considered akin to simply sitting in lecture without taking notes. Therefore, students were asked, "**When you watch the videos, what do you do while watching?**" One way to categorize students' reported interactions with the videos is through the ICAP framework (Chi & Wylie, 2014), which defines certain student behaviors as indicative of Interactive, Constructive, Active, or Passive engagement. It has been found that students who interact with an activity (including content delivery) at a higher mode of engagement score higher on knowledge assessments than those who interact with the same activity at a lower mode of engagement (Chi & Wylie, 2014). Framed with regard to content delivery (see Table 1 in Chi et al. (2014)), the modes of the ICAP framework start with Passive as the lowest mode, where students are simply receiving information (e.g., listening to a lecture and not taking notes). Active is the second mode and describes students who are repeating the delivered information (e.g., copying problem solutions or taking verbatim notes). The third mode is Constructive and includes generating new information based on what is presented (e.g., solving problems, comparing and contrasting ideas, or drawing trends that were not presented). The highest mode is Interactive, which is when students participate in constructive dialoguing regarding the information. Thus, what students are doing when they watch the videos can provide some information about how they are engaging with them. The response options provided to the students that were generated from the qualitative focus group data mirror behaviors that are described in the ICAP framework (Chi & Wylie, 2014). Using this framework, the response options presented in Table 4.6 were categorized into ‘Passive’, ‘Active’, and ‘Constructive’ engagement behaviors. Because a flipped course structure has students watch the videos outside of class, there were likely no opportunities that allowed the students to participate in ‘Interactive’ engagement. One of the response options, ‘I work the problems as they are presented in the video’ could have been categorized as ‘Active’ or ‘Constructive’ engagement depending on if students were simply copying down the problems or working them out on their own. This option was categorized as ‘Constructive’ for these survey results as qualitative data from short-answer responses and focus groups indicated students would generally do the latter. This can be seen in one of the short-answer student responses collected during survey development, *“if the video contains practice problems, I pause the video and do the problem myself first.”*

Table 4.6. Response percentages, by course, to survey item “When you watch the videos, what do you do while watching?”

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	64	32	232	273	58
[a]When you watch the videos, what do you do while watching? (Select all that apply)					
Response *categories* and options	Percentage of student responses to item				
Passive behaviors					
I just focus on the video itself (i.e., just listen or watch doing nothing else)***	44	41	28	36	5
Active behaviors					
I take notes on the material presented***	73	44	72	79	98
Constructive behaviors					
I work the problems as they are presented in the video***	73	47	63	43	67
I work on the homework problems***	27	19	21	8	5
I work on my chemistry lab assignments***	2	22[b]	1	3	2
Distracted behaviors					
I do other (non-chemistry-related) activities	2	6	7	5	5

[a]Survey item presented to everyone except those who selected ‘None’ for initial question, shown in Figure 4.3. ***$p < 0.001$. [b]Significant pairwise comparisons ($p < 0.05$) between Course Two and Courses One ($w = 0.35$), Three ($w = 0.35$), Four ($w = 0.27$), and Five ($w = 0.34$).

Although the percentages of students who selected the ‘Passive’ engagement response is relatively high in some courses (Table 4.6), it is not the majority response in any course and does not necessarily mean students were not engaging with the material at some point. As was noted earlier, a majority of students report re-watching the videos (Figure 4.4), usually focusing on parts that they missed or did not understand (Table A.7, Appendix A). Since students were instructed to select all the options that apply to what they do while watching, it is possible that they passively engaged the first time they watched and then actively or constructively engaged during a later time. The ability to rewind and re-watch allows students to interact with the videos with multiple modes of engagement, such that students who responded that they 'just focus on the video itself' could also have interacted with the video at a higher mode of engagement as well. This process was described by students in some of the focus groups that were conducted. For instance, one student stated that, "*I'll watch it once. Not once, but I'll watch one part through. Like if he's doing an example, he explains a lot, I'll just sit there and watch him explain the whole example. And then I'll go back and pause it and just write everything down. I'm not writing and listening…that's too much. I'll just watch it and then write it down*."

The difference in percentage of student responses for each response option (Table 4.6) when evaluated across courses were statistically significant ($p < 0.001$) with small to medium effect sizes. Only one potentially course-specific trend was noted. Course Two had a higher percentage of students who responded to the option, 'I work on my chemistry lab assignments', which was found to be statistically significant from the other courses with medium effect sizes when pairwise comparisons were completed. This was the only course that had the laboratory component incorporated as a part of the overall class grade, and as such, this option may only be relevant in courses where the lab is well-aligned with the course material. Lastly, course responses were low and found to not differ on the 'Distracted behavior' option, 'I do other (non-chemistry-related) activities'. Overall, taking into account that students could (and often did) select more than one response option, it is difficult to say that the different percentages indicated that students in some courses were more or less engaged in the videos than in other courses.

Why Do Students Find the Videos Helpful or Not Helpful for Learning?

All students who responded that they at least watched 'Some' of the videos (Figure 4.3) were asked to respond to two survey items about what they thought made the videos "**Helpful**" or "**Not Helpful**" for their learning. The response options for the "**Helpful**" items were categorized into 'Control of learning' or 'Perceived usefulness' (Table 4.7). Across all courses, the majority of students responded that they found the videos helpful for their learning due to their ability to control the pace and timing/location of watching them. This matches what has been reported in other studies that collected data on what students liked best about the flipped course, specifically comments related to control of learning (Christiansen, 2014; Ealy, 2013; Gregorius, 2017; Mooring et al., 2016; Parsons, 2019; Ranga, 2017; Reid, 2016; Shattuck, 2016; Weaver & Sturtevant, 2015). The students' perceived usefulness of the videos varied across courses, with no apparent trend present by course, although perceived usefulness has also been found to be part of what students would comment they liked about a flipped course in other studies (Mooring et al., 2016; Parsons, 2019; Ranga, 2017; Smith, 2013). While some response options showed significant differences by course, no course-specific trends were noted.

Table 4.7. Response percentages, by course, to survey item "Were the videos helpful to your learning? If so, in which ways were they helpful?"

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	64	32	231	263	57
[a]Were the videos helpful to your learning? If so, in which ways were they helpful? (Select all that apply)					
Response *categories* and options	Percentage of student responses to item				
Control of learning					
I can watch them at my own pace (e.g., rewinding, pausing, fast-forwarding)	80	78	87	81	93
I can watch them where and when I want***	72	44	75	68	88
Perceived usefulness					
They are easy to understand or include useful explanations and/or practice problems***	58	44	60	35	75
They help to reinforce the material***	72	44	61	33	44
They show other perspectives and ways of solving problems***	47	25	28	17	14
They often contain visual representations to understand the content*	47	41	50	49	72

[a]Survey item presented to everyone except those who selected 'None' for initial question, shown in Figure 4.3. **p < 0.01. ***p < 0.001.

When asked what was "**Not helpful**" about the videos with regard to their learning (Table A.9, Appendix A), the highest response percentages fell into the category of 'Do not meet learning expectations'. In most courses, the majority selected that they were unable to ask questions, with the next highest (and majority in Course Three) was that the videos did not contain enough practice problems. Students' dislike of not being able to ask questions while doing PCMs has been noted in other flipped studies as well (Ealy, 2013; Shattuck, 2016; Weaver & Sturtevant, 2015). Students also provided lower percentage responses across the categories of 'Not relevant to course', 'Don't hold attention', and 'Poor quality/disorganized' with no discernable patterns by course. One notable course-based response was in Course One, which used non-instructor made videos. Only 12% of students in Course One selected that '[the videos] have a different focus than the class material', which was one of the lowest among the courses. This reflects well for the use of curated online videos from different sources.

Conclusions

The conclusions from this multi-course investigation are framed by our research questions.

What are the predominant student behaviors and instructional styles for the F2F settings within these environments?

Timelines of COPUS codes (Figures A.1 – A.5), aggregated student behavior codes (Figure 4.2), and outputs from the COPUS Analyzer (Table 4.3) were used to categorize the student behaviors and instructional style of each F2F session. These sources provide complementary means (Reisner et al., 2020) for categorizing the predominant student behavior and instructional style of each course for the two days in which courses were observed. Courses One, Four, and Five were categorized as "student-centered" and Courses Two and Three as "interactive lecture" by the COPUS Analyzer. However, the timelines provide higher resolution and show greater variability in styles. The predominant student behavior observed in Courses One and Two were instructor-student interactions through whole-class questioning. This practice was supplemented by brief rounds of groupwork on worksheet problems in Course One and clicker-based groupwork in Course Two. Within Course Three, student behaviors included large blocks of listening during lecture delivery of material, interspersed with responses to whole-class questioning. Additional behaviors included clicker-based groupwork on both days and a group game on day one. In Courses Four and Five, the dominant student behaviors observed were group discussions on worksheets. In Course Four, groupwork was supplemented by instructor-led wrap-up sessions where students both listened to lecture and responded to whole-class questioning. In Course Five, groupwork was supplemented by intermittent questions posed by the instructor, with students responding individually. This is the first study to our knowledge to coordinate a systematic comparison of how the F2F time is used across a variety of flipped learning classrooms at different institutions. In discussing flipped classrooms more broadly, it is important to be aware of possible heterogeneity in how this learning format is implemented, especially with regard to the degree of peer-to-peer interactions.

Within flipped classroom studies, the types of observations and comparisons conducted in this study could help address two of the critical components (structural and instructional) in a fidelity of implementation framework (Stains & Vickrey, 2017). Structural critical components include the "expected elements related to the design and organization of the program, curriculum, or practice" (Stains & Vickrey, 2017). Therefore, details of how an overall course is organized along with high-resolution insights to how an F2F session is structured allows educators to identify where potential key features deviate from one another when comparing outcomes. Additionally, instructional critical components include the "expected participants' behaviors during implementation of the program, curriculum, or practice" (Stains & Vickrey, 2017). Details in this category include aspects of both instructor and student behaviors and engagement. In studies assessing the impact of active learning or other evidence-based practices (Freeman et al., 2014; Rahman & Lewis, 2019), the type of activity and degree of student engagement are often cited as confounding variables that can impact learning outcomes. Therefore, flipped courses with different F2F structures likely engage students to different degrees which may lead to different outcomes. As different F2F structures may engage students to different degrees, they have the potential to lead to different course outcomes (Chi & Wylie, 2014). Therefore, characterizing F2F time, such as we report here using the COPUS protocol, is vital to triangulating course outcomes and providing insight if difference are or are not found between implementation of the same instructional practice. The course structures compiled in this study will be incorporated into subsequent stages of our larger multi-institution study to explore the flipped learning environment's impacts on various aspects of student motivation and performance outcomes.

What are the students' self-reported use and perceptions of the PCMs in these environments? and What associations exist between instructional style in the F2F setting and PCM utilization?

The most salient outcome from the PCM survey was the variation students reported in their degree and timing of watching the videos. Students in Courses Four and Five more often reported watching all of the assigned videos, and doing so before the related material was addressed in the F2F session, despite the fact that these videos were generally longer than those in the other courses. The predominant student behavior in the F2F portion of these two courses consisted of groupwork on problem solving worksheets. Students in courses where the predominant behavior was responding to instructor-posed whole-class questioning, and therefore engaged in less student talk, were more likely to report watching fewer of the assigned videos and doing so after the related content was covered in an F2F session. In their study on active learning classrooms, Eddy and Hogan (2014) found higher reports of study time, completion of readings before class, and higher perceived importance of preparatory work for students enrolled in courses with increased structure. They explain that their observed outcome was likely due to more student accountability built into the increased structure (e.g., more student talk during F2F time and/or increased frequency of preparatory and review assignments). While neither our study nor Eddy and Hogan's were designed to uncover why these trends existed, it is recognized that classroom environments can stimulate the development of self-regulated learning (Zimmerman, 1995) and that self-regulated learners are empowered to create goals, use strategies, and implement actions to meet their goals (Ridley et al., 1992). These trends and their underlying mechanisms are worth further exploration as He et. al. (2016) found that non-compliance with recommended PCM utilization partially explained the small treatment effect of their flipped course outcomes. While other studies have looked at the impact of point-based incentives on students' PCM utilization in flipped courses (Christiansen et al., 2016; Woodward & Reid, 2019), these data provide a potential link between students' video viewing habits and how an F2F session is structured.

While initial viewing habits differed by course (Figure 4.3), more consistent responses across courses were reported with regard to re-watching habits and overall viewing behaviors. A universally high percentage of students reported re-watching the content videos in each course. The ability to watch and re-watch content videos whenever a student wants has long been touted as a benefit of flipped learning (Ealy, 2013; Smith, 2013). These results directly support Abeysekera and Dawson's (2014) note that PCMs help students to self-pace their learning and therefore manage cognitive load (Clark et al., 2005). When watching (either initially or upon re-watching), the majority of students reported active and constructive engagement behaviors (Chi & Wylie, 2014), with only a few reporting distracted behaviors. The importance of the PCMs (Casselman et al., 2019; Eichler & Peeples, 2016; Rau et al., 2017) and efforts to increase engagement with them (Eichler & Peeples, 2016) have been the focus of some recent reports on flipped learning.

Across courses, students' perception of how the PCMs were helpful was relatively consistent. The majority of students in each course (≳70%) noted that the control of learning was a helpful aspect. This included having control over the pace of learning and well as the location and timing of learning. In addition, aspects of the students' perceived usefulness of the PCMs were consistently selected across courses. Having control over and finding usefulness in learning resources are important aspects of achievement motivation (Deci & Ryan, 2000; Niemiec & Ryan, 2009; Ryan & Deci, 2000). Therefore, student reports of control and usefulness of PCMs are important for their motivation toward the flipped learning environment, which is in turn key for their learning (Anderman & Dawson, 2011).

Relatively consistent responses were found regarding why the PCMs were not helpful. The highest response category was that the videos 'Do not meet learning expectations', which encompassed being unable to ask questions or interact with the instructor, and/or that the videos did not contain enough practice problems. The inability to ask questions or interact with the instructor has been a concern since the earliest chemistry-based manuscripts on flipped learning (Ealy, 2013; Smith, 2013). Therefore, when the idea of not being able to ask questions came up in focus groups for this study, students were asked if they *typically* asked questions during their more lecture-based courses. Most of the students did not, but felt that *the opportunity* was there if they needed to. This could indicate a perceived loss of control for students in a flipped learning environment. Because we show that the inability to ask questions over the PCMs in real-time is a student concern across multiple, distinct flipped learning styles, it is compelling to look for alternatives to these in-the-moment questions. Devoting the start of each F2F session to answering remaining questions has been reported to help in some cases (Smith, 2013), but student frustration has still been reported in others (Ealy, 2013). As a flipped learning environment likely introduces this new aspect into the learning process for many students (i.e., self-regulated preparation using PCMs), it might be beneficial to consider their expectations for this, as well as other aspects, with regard to obtaining buy-in to the environment (Cavanagh et al., 2016).

Limitations

This nonexperimental research has several limitations that should be considered when interpreting the outcomes presented. Despite the coordinated processes across multiple courses, this study only includes select introductory and general chemistry courses. Therefore, outcomes may not be generalizable to flipped chemistry courses with different student populations (e.g., by level, demographics, etc.). Additionally, this study set out simply to document information about the two main tenets of the flipped courses (i.e., pre-class content delivery and use of active learning during in-class sessions). Student performance was not collected; therefore, we cannot comment on how noted differences in F2F structure or PCM use may have impacted student outcomes. This limitation is being addressed in a subsequent phase of the larger project, where a variety of motivational and performance outcomes are also being collected within each course.

With specific regard to the classroom observations, on-site visits were scheduled with each instructor. Therefore, each knew when the consecutive observations would be conducted, potentially impacting the F2F environment structure on those two days. On-site visit timing and course variations precluded the ability to observe similar topic coverage across courses. While day-to-day consistency was seen within each course, the variability of instruction between courses could not be examined at the topic level. However, during instructor interviews none noted adjusting their practices based on content. Finally, video recordings were focused on the class as a whole, capturing a majority of the class and most instructor motion. Therefore, at the student-level, COPUS codes were applied with regard to the majority behaviors observed, not to individual or group-level behaviors. While this is typical for use of the COPUS (Lund et al., 2015; Smith et al., 2013; Stains & Harshman), it does not capture these finer-grained variations that could impact individual student course outcomes.

Data on the pre-class materials (PCMs) only represents the perceptions and opinions of the students who self-selected to participate in the focus groups and surveys conducted. Therefore, these self-reported behaviors may not reflect those of other students, especially within courses with low participation. Future studies are encouraged to utilize the same questions and response options to better gauge the noted response trends across more environments and student populations. Within any self-response study, students' responses could be influenced by social desirability; that is, students might respond based on what would make them "look best". However, as no data from this study was collected within the authors' institutions and none of the instructors were involved in the data collection process, the influence was potentially diminished as the research team had no connections to the students. Lastly, the PCM survey was administered midway through each course, therefore students should have been calibrated to the structure of the course and had several forms of feedback regarding their abilities and performance. This one-time survey does not account for any changes in students' use of the videos or differences in their perceptions over time. Each would be informative to capture a fuller picture of students' engagement with this aspect of a flipped learning environment.

Finally, while PCM use was observed to trend with the structure of the F2F environments, we cannot rule out other influences that may have contributed. As noted earlier, student buy-in may impact the degree to which students engage with course elements (Cavanagh et al., 2016). Additionally, the consistency of student reminders has been shown to impact student use of PCMs (Woodward & Reid, 2019). Therefore, as these aspects were not explicitly measured within this study, they cannot be ruled out. As many flipped learning environments incentivize student use of the PCMs through on-line or in-class quizzes, further investigation of the influence of both F2F structure and incentivization may be warranted.

Implications

For practice

Although the variability between the courses included in this study does not allow for a definitive link to be made between the F2F structure and students' use of PCMs, the results provide some insight into the possible connection between the two tenets of flipped learning. When students were expected to spend a larger percentage of F2F time working in a group instead of participating in Socratic dialogue with the instructor, more of the students reported watching all of the videos and did so before the F2F session. Thus, if an instructor's expectation for students is that they watch all the videos before coming to class, it may be helpful to not only make these expectations clear to the students, but also to include more peer-to-peer groupwork in the F2F environment. If student "accountability" is a driver of increased PCM engagement, instructors may consider other ways this could be instilled, noting that the use of viewing incentive quizzes was not found to make a difference in this study. Likely unrelated to accountability, the use of regular email reminders combined with online homework has been shown to promote high levels of student engagement with PCM videos (Woodward & Reid, 2019).

It can often be difficult for instructors to gauge the amount of time that students are actually spending doing things like engaging in peer-to-peer interactions, responding to whole-class questions, or simply listening to a presentation, which is why collecting observational data using a protocol like COPUS is essential for understanding a learning environment (Reisner et al., 2020). If COPUS results of student behaviors do not align with an instructor's expectations for the environment, then F2F time could be adjusted to account for discrepancies. In addition to documenting student behaviors, measuring students' cognitive engagement could provide another level of detail into what students are doing during F2F time. Use of the ICAP framework (Chi & Wylie, 2014) of cognitive engagement has found that students who "passively" engage with course material generally perform worse on knowledge assessments than students who interact with the material at a higher mode of engagement (i.e., "active", "constructive", or "interactive"). Even though timelines of students' behaviors (e.g., COPUS) may indicate that students are participating in peer-to-peer interactions, these do not necessarily mean that they are cognitively engaging with them at a higher mode. Therefore, evaluating cognitive engagement can provide a deeper understanding of what students are doing in the F2F environment of a flipped course. This level of informative feedback can further influence how changes to instructional practice are implemented.

Many previous studies on flipped courses have presented results about what students thought were and were not helpful about the PCMs and the results of this study are similar to what was found previously (Christiansen, 2014; Ealy, 2013; Gregorius, 2017; Mooring et al., 2016; Parsons, 2019; Ranga, 2017; Reid, 2016; Shattuck, 2016; Smith, 2013; Weaver & Sturtevant, 2015). Specifically, that students generally found that the videos were helpful based on inherent properties, such as being able to pause and rewind or being able to watch when it was best for their schedules. Students also pointed out that they found the inability to ask questions, as well as a ‘lack’ of practice problems, made the videos less helpful for their learning. As such, instructors may want to consider how best to address these aspects when implementing a flipped course. A potentially novel result found in this study was that students’ responses to what was and was not helpful about the videos did not appear to be affected by the source of the video, as Course One used non-instructor made videos curated from online sources. Although the source of the videos was not the focus of this study and further research should be done on the effect of non-instructor made videos on students’ perceptions and use of the videos, this result suggests that thoughtfully selected online videos that align with the class material are perceived as just as helpful to students as instructor-made videos.

For research

The goal of this study was to employ a coordinated set of assessment practices to evaluate the F2F environment and students' perceptions and use of PCMs across multiple flipped chemistry courses. Subsequent phases of this project will utilize these data to triangulate course outcomes. Although previously published studies on flipped courses have been completed, they have typically focused on single courses or institutions and outcome comparisons across these studies can be limited by inconsistent assessment practices. When data is collected using coordinated assessments from multiple courses, comparisons between the courses can be directly evaluated, allowing for general trends and features to be detected and explored. We therefore encourage other researchers studying flipped classrooms, or any evidence-based instructional practice, to begin to design larger coordinated studies that may bring novel insights to our understanding of how these practices are adapted and what impacts an adaptation may have on student and course outcomes.

As variability exists in how instructors' structure and support the two tenets of flipped learning, it is important to provide information about PCM use and the F2F environment structure when presenting results. Without implementation details, the validity of findings from a flipped learning environment may be compromised (Stains & Vickrey, 2017). The F2F observations made in this study included multiple levels of information, such as the COPUS analyzer (Stains & Harshman) clusters, percentages of class time that students spent on different activities, and timelines of student and instructor COPUS codes. The COPUS analyzer clusters provided information about the general 'type' of F2F environment that was implemented in each course and provided a lower-resolution picture of the F2F style. Additional data from the percentages of class time students spent on different activities provided more detail into what students were doing in the F2F environment, as this was observed to differ even in courses that were part of the same cluster. Taken together, these data provided a general picture of the F2F environment at a similar degree of resolution as observations made by Cannelas et al. (2017), who reported the percentage of class time the students spent being "active" (i.e., everything except watching, listening, and taking notes). The COPUS timelines presented higher-resolution data, specifically, what was happening in the classroom every 2 minutes. Timelines allow for more details to be presented regarding when different behaviors occur in the classroom. For instance, a study by Donnelly and Hernández (2018) presented the percentage of students engaged during the F2F portion of a flipped classroom for each 2-minute interval and found that student engagement fluctuated throughout class time and type of active learning (i.e., whole-class discussion vs. group activity). Therefore, although details of the course's structure should always be included in a study, deciding to include and/or emphasize either the low- or high-resolution data about the F2F environment should be dependent on the specific research question being asked.

Finally, although this study explored the relation between the F2F environment and students' perceptions and use of PCMs in flipped chemistry classrooms, data about every facet of the environment was not gathered. Therefore, continued research into these two tenets of flipped courses is needed to better understand how students' behaviors, engagement, and learning is affected by these environments. Qualitative studies with student focus groups and/or interviews could be used to ask students about why they do or do not engage in different behaviors with the PCMs and in the F2F environment. Additionally, qualitative or mixed-methods studies could be coupled with tracking data to gather more details about if and when students are using PCMs based on the expectations of the F2F environment or levels or assessments and if these behaviors change throughout the course of the term. Further investigation into “if”, “when”, and “why” students do or do not engage in these two tenets could provide valuable information about why outcome differences are seen between different flipped course environments.

Associated Content

Supporting Information

The Supporting Information is available in Appendix A and includes COPUS code definitions, survey development details, supplementary survey item data tables.

Author Information

Corresponding Author

*E-mail: jack.barbera@pdx.edu

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. (DUE 1611220 and 1611519). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This project would not have been possible without the participation of each instructor and the students in their flipped courses, we are therefore very thankful for their assistance in this study. Classroom observations and instructor interviews were collected by Regis Komperda, she is thanked for her time and willingness to travel to each institution. Cory Hensen is thanked for taking the lead on the initial focus group coding and for training and mentoring the secondary coders. Carrie Zografos, Andrew Isom, and Gosia Cox are thanked for their assistance in COPUS coding. Nicole James is thanked for her assistance in structuring and editing the manuscript itself.

Chapter 5: Multi-institutional Study of Self-Efficacy within Flipped Chemistry Courses

Reprinted with permission from Naibert, N., Duck, K., Phillips, M. M., & Barbera, J. (2021). Multi-institutional Study of Self-Efficacy within Flipped Chemistry Courses. *Journal of Chemical Education, 98*(5), 1489-1502. Copyright 2021 American Chemical Society.

Abstract

Active learning environments have been shown to be beneficial for student learning, however, including such activities can be limited by the class time available. One method that can provide more opportunities for active learning during face-to-face class time is the flipped learning approach. However, studies on the impacts of flipped learning environments on student motivation are limited. Therefore, in this multi-institutional study, general chemistry students enrolled in flipped courses at three institutions responded to measures of self-efficacy and self-regulatory strategies. The results from these measures were used to evaluate how students' academic self-efficacy (ASE) and chemistry self-efficacy (CSE) changed over the term at each institution, as well as to compare students' CSE between the institutions. Evidence was found for scalar measurement invariance across all measures, such that latent means could be used to compare results over time and between the institutions. Overall, students at each institution showed a decrease in ASE over the term, although their CSE increased. Comparisons between the institutions showed that students at the Southeastern institution had a higher post CSE than students at the Western and Northwestern institutions. One salient difference between the institutions was the structure of the face-to-face class time, which suggests that there may be a relation between students' post CSE scores and the

structure of the course. However, other variables, such as the demographic profiles of the institutions, may have also played a role in the observed differences.

Graphical Abstract

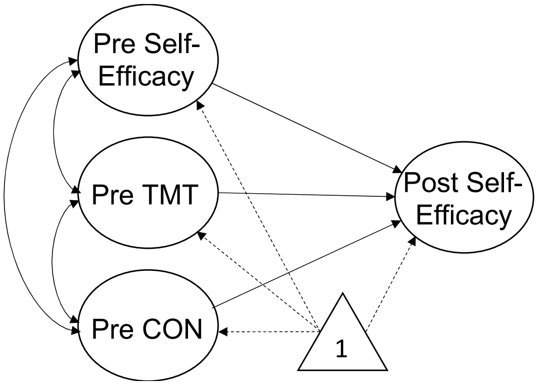


Figure 5.1. Graphical abstract for Chapter 5.

Keywords

First-Year Undergraduate / General, Chemical Education Research, Collaborative / Cooperative Learning, Student-Centered Learning

Introduction

Over the past few decades, many have advocated for the adoption of more student-centered, active-learning pedagogical approaches in college science classrooms (Freeman et al., 2014; National Research Council, 2000). The goal of moving from a more instructor-centered, lecture-based, approach is to more fully engage students in the learning and inquiry process, which may better instill higher-order learning (e.g., analysis, synthesis, and evaluation of content) and increase student-instructor and student-student interactions. Research has supported a shift from teaching approaches that focus solely on memorization to those that also incorporate greater levels of problem solving, which can lead to more developed mental models for greater meaningful learning (Freeman et al., 2014; Michael & Modell, 2003; National Research Council, 1999). With an active learning approach, the instructor becomes a facilitator during the learning process compared to the "sage on the stage" (King, 1993; Merriam et al., 2007), with potential to push students to become more self-directed and take greater ownership over their learning.

A wide variety of teaching methods have been grouped under the umbrella of "active learning" techniques, e.g., using clicker questions, peer-led team-learning (PLTL), process-oriented guided inquiry learning (POGIL), problem- and project-based learning (PBL), think-pair-share, instructor-led class discussions, and group discussions (Freeman et al., 2014; Michael & Modell, 2003). All of these contrast with a more instructor-centered approach, however they can vary based on the level of student activity and engagement generated during the learning process. The use of an active learning approach does not necessarily mean greater student engagement and motivation unless a synergy is created between the two (Barkley & Major, 2020). Thus, exploring instructional models that provide opportunities for active learning techniques is crucial to understanding the nuanced aspects between the use of active learning techniques and student motivation. One instructional model that has allowed for greater opportunities to employ active learning techniques in the classroom has been flipped learning.

Flipped Learning Model and Chemistry

The flipped learning approach moves the delivery of direct instruction from the classroom space, making room for more student-centered activities. The earliest reports of this type of inverted classroom structure date back to 2000, with a rapid and steady rise in the education research literature beginning in 2011 (see Fig. 1 in Casselman et al.

(2020)). Early reports within the higher education chemistry education literature focused on suggestions for developing and implementing the technique (Ealy, 2013), impacts compared to traditional instruction (Amaral et al., 2013), and student attitudes (Smith, 2013).

Many studies on flipped learning within chemistry education have utilized course-based measures (e.g., course evaluations, exams, grades) to report on students' perceptions of being in a flipped course and its impact on performance-based outcomes. Fewer studies have focused on measuring outcomes related to other *constructs*, with some exceptions. A 2016 study investigated student *attitudes* in a flipped organic chemistry course using the revised version of the Attitude toward the Subject of Chemistry Inventory (ASCIv2) (Mooring et al., 2016). In 2017, the Student Assessment of Learning Gains (SALG) was used to investigate student *perceptions* and *attitudes* in organic (Canelas et al., 2017) and general (Rau et al., 2017) chemistry courses. A 2018 study investigated students' *engagement* within a flipped physical chemistry course using the Behavioral Engagement Related to Instruction (BERI) protocol (Donnelly & Hernández, 2018). With specific regard to investigating *motivation*, the chemistry version of the Academic Motivation Scale (AMS-Chemistry) was utilized to explore differences in motivation between a traditional lecture course and a flipped course that included PLTL (Liu et al., 2017). As the AMS-Chemistry is based on Self-Determination Theory (SDT) (Deci & Ryan, 2012), results indicated that although students' intrinsic and extrinsic motivation were similar between the two courses at the end of the term, students' scored lower on amotivation (i.e., lack of motivation) in the flipped-PLTL course compared to the lecture-based course. In an additional study, the chemistry version of the Science Motivation Questionnaire (SMQ-II) was administered within flipped general chemistry courses to compare motivation and final course grades (Hibbard et al., 2016). Results indicated no discernable pattern between first-term grades and motivation, with a pattern arising at the end of the second-term. The SMQ-II is based upon Social-Cognitive Theory (SCT) (Bandura, 1997; Bandura, 2001), however, the subscales draw upon multiple theoretical frameworks of motivation while also seeking an overall motivational composite score, which has resulted in complications for measurement and scale adaptability (Komperda, Hosbein, et al., 2018; Komperda et al., 2020). Given this limited number of motivation-based studies of flipped learning environments, there is still a need for the use of sound motivational theories and frameworks in investigating their impacts.

Social-Cognitive Framework for Motivation

From a social-cognitive perspective, learning is viewed as being dynamic and dialectical in nature between learner's beliefs, behavior, and the environment in which the learning takes place (Bandura, 1997; Bandura, 2001; Brophy, 2010). As part of this dynamic aspect, psycho-social factors like motivation play an important role for student success in college learning environments (Robbins et al., 2004). Evidence has supported the notion that academic motivational factors have a significant impact on learning outcomes (e.g., see Anderman and Dawson (2011) for a summary). When drawing on a social-cognitive perspective, two constructs that provide insight to understanding students' goal directed actions and the reciprocal interactions within their learning environment have been self-efficacy and self-regulation (Richardson et al., 2012).

Self-Efficacy. Self-efficacy within the academic realm is the perceptual acuity one has regarding their capabilities to learn or carry out certain tasks to attain an academic outcome (Bandura, 1997). Even though academic self-efficacy is not the same as ability, it has been shown to predict academic success and performance across different age levels and content areas (Pajares & Urdan, 2006; Richardson et al., 2012; Robbins et al., 2004). One source of self-efficacy is connected to direct engagement and task completion (Bandura, 1997). The perception of success (or failure) upon completing a task can have a direct impact on increasing or decreasing one's self-efficacy (Bandura, 1997).

Many times, in academic situations, self-efficacy is measured toward the beginning of a course and used to predict academic performance at the end (DiBenedetto & Bembenutty, 2013), while mid- or end of semester assessments might provide a different perspective on the association between academic self-efficacy and performance (e.g., Galyon et al. (2012)). At these later time points, students have completed a number of assignments and assessments across their course load and thus have more feedback to inform their self-efficacy beliefs in that context. Studies in chemistry have employed self-efficacy measures to compare different groups of students or learning environments (Chase et al., 2013; Stanich et al., 2018). Other studies have measured self-efficacy for use as a predictor variable of academic outcomes (Ramnarain & Ramaila, 2018) or as one of several variables in a larger educational model (Ferrell et al., 2016; Reardon et al., 2010; Villafañe et al., 2016). For studies that explored changes, many found that self-efficacy generally increased over the term (Ferrell & Barbera, 2015; Graham et al., 2019;

Villafañe et al., 2014; Vishnumolakala et al., 2017), although, some have noted that this increase was dependent on the demographic group (Villafañe et al., 2014).

Variation in results could be based on whether self-efficacy is assessed on one's perception of performing a certain task, a specific subject area, particular topics or concepts within a subject area, performance in a specific class, or compared to all of their courses within a current semester. When the lens used to study self-efficacy is focused at a more specific level (e.g., at the subject, content, or task level), the predictive ability becomes greater for performance (Choi, 2005), future success, and re-engagement (Bandura, 2006). For example, where Galyon and colleagues (2012) found academic self-efficacy went down over a semester, Lawson and colleagues (2007) found science self-efficacy to go up over a semester. This variation could be based on the level of specificity for how self-efficacy was measured, which might contribute to the magnitude of the self-efficacy and performance association. Within chemistry, self-efficacy has commonly been measured using a variation of either the Chemistry Attitude and Experience Questionnaire (CAEQ) (Chase et al., 2013; Villafañe et al., 2014; Vishnumolakala et al., 2017) or the College Chemistry Self-Efficacy Scale (CCSS) (Ferrell & Barbera, 2015; Graham et al., 2019; Ramnarain & Ramaila, 2018). These measures primarily include items based around specific chemistry tasks, the course itself, or application of chemistry concepts to real-life situations and can be considered measures of chemistry self-efficacy (CSE). Although some studies have measured self-efficacy at a more general level (Reardon et al., 2010), none have included measures of CSE and academic self-efficacy (ASE) simultaneously.

Richardson and colleagues (2012) conducted a meta-analysis investigating the association between ASE and university success by means of grade point average (GPA). They found 9% of GPA variance could be explained by ASE. However, effect sizes varied widely between studies, indicating that there could potentially be factors that mediate or moderate this relation. For example, deep processing strategies used by students (Fenollar et al., 2007) and effort regulation (Komarraju & Nadler, 2013) have been shown to mediate the relation between self-efficacy and academic performance. Whereas, Tabak and colleagues (2009) found time on task to be a moderating factor. With the potential for mediating and moderating effects, aspects of self-regulation for how students focus their time, effort and learning strategies have the potential to highlight aspects of this relation.

Villafañe, Garcia, and Lewis (2014) noted the importance of examining gender and race/ethnicity when investigating chemistry self-efficacy over time. In chemistry, gender differences have been identified at different time points (e.g., beginning and end of semester) and for different qualitative factors. For example, Dalgety and Coll (2006) found that males had higher self-efficacy at the beginning of a semester and qualitatively worried more about specific aspects of chemistry content connected to their self-efficacy, while women were found to have lower self-efficacy overall from a qualitative analysis. Sunny and colleagues (2016) also found men to have higher chemistry self-efficacy at the end of a semester utilizing a task specific measure for chemistry adapted from the motivated strategies for learning questionnaire (MSLQ). An analysis of narrative cases in STEM (Zeldin & Pajares, 2016) found men’s self-efficacy beliefs to be tied more to mastery experiences, while women’s relational experiences in the learning environment (e.g., social persuasion and vicarious learning) were the greater influence. In connection to the classroom structure, Boz and colleagues (2016) concluded that perceptions of a chemistry learning environment mediated the relation between gender and self-efficacy at the end of a semester, after finding that when females perceived a more positive learning environment it mediated higher levels of self-efficacy beliefs. Given these prior findings, it is important to continue to examine the development of self-efficacy beliefs in addition to accounting for potential gender and race/ethnicity differences while doing so.

Self-Regulation. Self-regulated learning refers to the ability of an individual to self-generate thoughts, feelings, and behaviors and organize them to direct their abilities toward a goal before, during, and after a learning task (Pintrich, 2004; Zimmerman, 1998, 2000). As part of this process, students must use effective learning strategies to organize and manage their thoughts, behaviors, and time wisely. Individuals that tend to report using more strategic self-regulation tend to perform better than less self-regulated students (Pressley & Ghatala, 1990). Even though self-regulatory skills can be taught (Pintrich, 2004; Schunk & Ertmer, 2000), some have noted that students need the skill and will to use self-regulatory strategies (e.g., Snow (1996)) and thus is something that can be controlled when assessing learning outcomes.

One component to a number of self-regulation models includes the monitoring and management of one's learning. For monitoring, these might be potential distractions or barriers while trying to learn new material, e.g., not being able to concentrate on new material because the textbook is perceived to be boring (Huff & Nietfeld, 2009; Winne, 2004). Whereas, management connects to how a student plans and sustains their efforts toward the task (Wolters et al., 2017). Those that use self-regulatory strategies tend to be viewed as taking a more active stance toward their learning (Zimmerman, 1990). As a flipped learning environment requires students to use more self-regulated learning strategies both in and outside of the classroom, they need to take ownership over and become more involved in the learning process. For example, students must adequately manage their time and focus on the video content assigned before coming to class. Thus, it is important to assess and control for how students utilize different strategies and resources to learn, manage their effort and organize their time, and monitor and evaluate their learning outcomes (Pintrich, 2004; Zimmerman, 1990).

There is wide variability in students' perceptions of self-efficacy and their use of self-regulatory strategies in learning situations (DiBenedetto & Bembenutty, 2013). A consistent finding has been that domain-specific measures of motivation have shown a greater relation to academic achievement compared to global measures (Pintrich, 2003). Additionally, when considering a complex psychological phenomenon like motivation, taking the multi-dimensional and multi-faceted nature of the construct into account is crucial. As Anderman and Dawson (2011) note, there is no "one size fits all" when using the term motivation. It has been maintained that a one-item measure assessing students' perceptions of enjoyment do not tend to assess student motivation based on its complexity (Brophy, 2008). Thus, when examining academic motivation, it is important to identify and measure different aspects that are important for the learning context being studied.

Measurement

To gather data about students' self-regulation and self-efficacy within a learning environment, self-report survey measures are typically administered. To produce meaningful inferences, the measures must be aligned with the constructs of interest and be shown to produce valid and reliable results with the target population (American Phsychological Association & National Council on Measurement in Education., 2014; Arjoon et al., 2013). When using extant measures supported by prior psychometric studies, the primary evidence for data validity is the underlying structure. *Structural validity* provides evidence that the data derived from each indicator variable within a measure are properly associated with the *a priori* model for the latent construct being measured (Knekta et al., 2019). If structural validity of the data from the population under investigation is supported, then evidence is provided that the data maps onto the latent construct. However, if the structural validity of the data is not supported, investigations of the *Response Process* and/or *Content Validity* may need to be conducted (Komperda, Hosbein, et al., 2018; Komperda et al., 2020). Furthermore, if the measured data will be used to compare groups on the latent construct, evidence of *Consequential Validity* needs to be established. For self-reported quantitative data, this level of validity can be supported through measurement invariance to determine if group-bias is present in the data structure (Rocabado et al., 2020). Finally, when measures are only administered once per time point, an estimate of the single-administration *reliability* is warranted (Komperda, Pentecost, et al., 2018).

Purpose of this study

This study employed a social-cognitive perspective to investigate chemistry students' self-efficacy and self-regulation strategies within flipped learning environments. To broaden the generalizability, data collection spanned courses from a range of institutions and used a coordinated set of assessment instruments. In conducting this work, the following research questions were addressed: 1) What evidence supports the validity and reliability of the data generated from the coordinated assessments at our sites?, 2) How do students' self-efficacy and self-regulation change within each flipped learning environment?, and 3) How do these constructs compare across sites? To answer these questions, we examined students' self-efficacy and self-regulation at three institutions. Prior to conducting comparative analyses, data from each assessment instrument were explored for evidence of validity and reliability. Data validity was further supported by cross-validation and measurement invariance studies, following which, structural means modeling was utilized to compare outcomes within and across institutions.

Methods

Population

Three institutions from the United States were involved in this study. All three were public research universities but varied in their acceptance rate and demographic profile (Table 5.1). These data collection sites were selected based on the corresponding author's knowledge of who the flipped learning instructors were and that none were new to course flipping. As such, each instructor had a minimum of two years of experience in flipping their course and was the primary or only person involved in developing the course materials (Table 5.2). The general structure of each course followed the two basic tenets of flipping: 1) foundational information was delivered to students through pre-class materials (PCMs), and 2) the face-to-face (F2F) environment was utilized for the application or expansion of the information through active learning (Bergmann & Sams, 2012).

Table 5.1. Institution details.

Institutions by Region			
	Southeastern	Western	Northwestern
Size (Approx.)	55,000	35,000	30,000
Type	Four-year, Public, Doctoral – Very High Research Activity	Four-year, Public, Doctoral – Very High Research Activity	Four-year, Public, Doctoral – Very High Research Activity
Acceptance	50%	30%	78%
Demographics[a]	Asian – 4% Black – 13% Latino/a – 60% White – 13% Other – 7%	Asian – 27% Black – 4% Latino/a – 12% White – 39% Other – 18%	Asian – 7% Black – 1% Latino/a – 9% White – 61% Other – 22%

[a] 'Other' category includes designations of International, Pacific Islander, 2+ ethnicities, and/or other designations inconsistently reported across institutions.

At the Southeastern and Northwestern institutions, data were collected from multiple course sections across multiple years, with each taught by the same instructor or team of instructors (Table 5.2). At the Northwestern institution, a lead instructor was responsible for the development of the materials and structure employed in flipping the course, this instructor co-taught with the other instructors involved each year. A prior observational study with these courses did not reveal any substantial differences in the structure of the in-class settings across sections (Naibert et al., 2020). All data collected within this study was approved by the Institutional Review Board (IRB) at Portland State University and appropriate consent was acquired from students as required by the IRB.

Table 5.2. Course details.

	Southeastern	Western	Northwestern
Course Type	General I	General II	General I
Enrollment	793	281	974
Sections	4[a]	1	6[a]
Instructors	1	1	3[b]
Schedule	75 min, 3 times per week, morning	80 min, 2 times per week, evening	80 min, 2 times per week, morning

[a]Data collection spanned multiple years. [b]One instructor was the primary developer of the flipped learning materials used in each course and co-taught the sections with the other instructors each year.

Instruments

Chemistry Self-Efficacy (CSE). This measure was developed to be specific to students' understanding and comfort level with different chemistry concepts (Ferrell & Barbera, 2015). The measure includes 6 items that address how well students understand different areas of chemistry (e.g., properties of elements, interpreting chemical equations, explaining chemical laws and theories). The items were measured on a five-point rating scale anchored by *very poorly*, *poorly*, *average*, *well*, *very well*.

Academic Self-Efficacy (ASE). Out of the 15 subscales from the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al., 1991), we utilized the Self-Efficacy for Learning and Performance subscale, which includes 8 items related to students' expectancies related to their learning and understanding. For this study, the subscale was adapted to measure a more general aspect of academic self-efficacy by changing the phrasing from "in this class" to "in my courses" as the referent. In addition, the scale was changed from a seven-point scale (*not at all true of me* to *very true of me*) to a five-point Likert scale (*strongly disagree* to *strongly agree*) to align with the other measures used in the study.

Learning and Study Strategies Inventory (LASSI) Subscales. LASSI is an 80-item measure with 10 subscales to assess success of course or program changes regarding academic skill, will, and self-regulation (Weinstein et al., 2002). For purposes of this study, we used two of the self-regulation strategy scales to assess students' concentration (CON) and time management (TMT). Each subscale included 8 items on a five-point Likert scale (*strongly disagree* to *strongly agree*). The CON subscale centers on monitoring distractions, being able to focus one's attention, and refocusing attention after losing it during studying and in class. Whereas, the TMT subscale assesses how well students organize their schedules, procrastination, and cramming behaviors.

Data Collection

In each course, two surveys were deployed. The first took place within the first two weeks of a term (pre) and the second during the last few weeks (post), neither of which overlapped with an exam. At both time points, the survey contained the same items from the four noted instruments and was open for one week. Due to the use of two different response scales, all Likert-scale instruments were presented first (Thompson & Green, 2013). Following these items and on a stand-alone page, students were presented with a note indicating a change in the response options before being presented the last set of items on the subsequent page. Demographic information was collected at the end of a survey, following all instrument items. The instructor of each course was provided a brief script to make an initial in-class announcement regarding the survey. A note similar to the script was posted on the classroom management platform of each course. Students who were interested in participating clicked on a link to the Qualtrics survey that was part of the announcement note. Some instructors offered a nominal amount of extra-credit points for accessing the survey.

Data Analysis

For each pre and post survey, data were examined for exclusionary criteria. Cases were removed for records that started a session, but did not fill out any information. Duplicate cases were also removed that had less information, or were second attempts if both cases were complete. All analyses were completed using the *lavaan* package (version 0.6-5) in R (version 3.6.2) with a means and variance adjusted weighted least squares (WLSMV) estimator to account for the ordinal scale of the items. Descriptive statistics for the aggregated data as well as by institution are included in Tables B.7 and B.8 in Appendix B. Listwise deletion was used for incomplete responses for each scale, thus the sample size for each scale may vary slightly. A focus of the analyses was to consider differences in the measures of interest based upon gender and underrepresented minority (URM) status. For these analyses, male was used as the reference category for the by gender comparisons, and non-URM (which consisted of individuals who identified as either non-Latino/a White or Asian) was used as the reference category for the by URM comparisons. All demographics were self-reported by the students who responded to the survey.

Validity and Reliability

Structural validity of the individual scales was investigated using Confirmatory Factor Analysis (CFA). Reliability was calculated using omega. Scalar invariance was established for the four measures for longitudinal invariance, invariance between institutions, and invariance between gender and URM status. Details about the procedures and methods used for these analyses are included in Appendix B (Tables B.1 – B.6, B.9 – B.12).

Structured Means Modeling

Establishing scalar invariance provides support for the use of latent factor means when comparing groups (Rocabado et al., 2020). To do so, structured means modeling (SMM) was used. SMM includes the mean structure into the measurement model such that a relative difference between the latent means can be determined (Thompson & Green, 2013). Two types of analyses were completed using SMM: 1) the change in pre to post latent means for each factor and 2) the difference between post latent factor means while controlling for pre factors. These analyses were completed for institutional comparisons, as well as demographic comparisons (i.e., by gender and URM status).

The difference in latent means from the pre assessment to the post assessment of each factor for each institution was calculated. As SMM produces a relative mean difference, the pre factor mean for each comparison was set to zero, which allowed the value obtained for the post factor to represent the difference between pre and post factor means, or the latent mean difference, for that institution. This analysis was completed for all four measures and all institutions separately. The matched data from the Western institution included incomplete use of the entire response scale for certain items, however, for pre-post longitudinal models, the thresholds for these missing response categories could easily be removed for the appropriate factor in *lavaan*. Thus, the results for the pre to post comparisons account for those missing response categories where appropriate. The pre to post latent mean differences were also completed with the aggregated data set to compare differences based on gender and URM status. In these analyses, the male and non-URM groups were set as the reference, with female and URM groups as the comparison group, respectively.

To compare post latent means between institutions for the CSE and ASE factors, each institution's latent mean on the respective self-efficacy pre assessment and the pre assessments of TMT and CON were controlled for. This was completed by incorporating these pre factors as covariates into the model of the post factor (Hancock, 2004). Since the factors are theoretically related (Pintrich et al., 1993), the pre factors were correlated (Figure 5.2). All pairwise comparisons were made between the three institutions. Since this analysis relies on mean differences, all latent means were in comparison to a reference institution. Thus, the results from this analysis represent the difference in the latent means between the institutions and not absolute scale values. In addition, the post latent mean comparisons control for pre latent means included in the model. This analysis also used the matched data sets, in which some items for the CSE scale did not include complete use of the response scale for the Western institution. Since the thresholds between response categories cannot easily be removed from only one institution, a 'dummy' response pattern was added to the institution to account for the missing categories. A detailed description of this method is provided in Appendix B. Post latent mean comparisons were also conducted for the same demographic groups assessed in the pre-post comparisons.

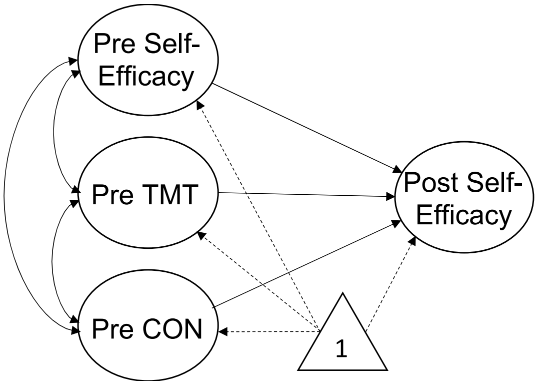


Figure 5.2. The path model with mean structure for self-efficacy (ASE or CSE) post latent mean differences with pre self-efficacy (ASE or CSE), TMT, and CON controlled for. For clarity, items are not shown.

The effect size for all latent mean differences were calculated as the absolute difference in factor means divided by the square root of the pooled variance of the factors (Thompson & Green, 2013). Although this effect size calculation is similar to Cohen's d, where effect sizes are small (~0.2), medium (~0.5), and large (~0.8) (Cohen, 1992), the magnitude guidelines for latent variables are generally accepted to differ slightly from those used for measured variables. Since latent means are free from measurement error, the magnitude of the effect size for latent mean differences should be larger than those for measured variables (Thompson & Green, 2013).

Results

Responses

The cleaned datasets by administration time and institution are detailed in Table 5.3. The response rates are based on the week-1 enrollments and therefore may not accurately reflect the percentage of participants from the actual enrollments at the time of administration. To determine if the students who ended up in the matched dataset differed significantly from those who did not, group means comparisons (i.e., t-tests) of the pre-scores for each scale at each institution were conducted. These analyses detected no significant differences between groups for any scale at any institution, indicating that the subset of students that made up each matched dataset did not represent a unique subset of the course population.

Table 5.3. Institution sample sizes and response rates by survey administration time.

Institution	Southeastern	Western	Northwestern
Pre, n (%)[a]	554 (70)	212 (75)	797 (82)
Post, n (%)[a]	293 (37)	217 (77)	710 (73)
Matched, n (%)[a]	266 (34)	170 (60)	563 (58)

[a]Percent response based on the week-1 enrollments noted in Table 5.2.

Evidence of Validity and Reliability

The initial and final data-model fits, along with details of the modifications undertaken to produce the final models, are provided in Appendix B. For each scale, the final CFA model was fit individually for each institution to cross-validate the structure with respect to each institution. Overall, there was acceptable data-model fit and evidence of good reliability (omega values above 0.80) for each final model with respect to each institution (Table B.6).

To support the use of latent means (via SMM) for comparing measurement results by group, scalar invariance was evaluated. First, as each measure was administered at two time points (i.e., pre and post), and the change from pre to post was determined, the *longitudinal scalar invariance* was evaluated (Table B.9). Next, as the results from each measure were compared across institutions, *scalar invariance by institution* was evaluated (Table B.10). Finally, as the results from each measure were compared by gender and by URM-status, *scalar invariance by gender* and *by URM-status* were also evaluated (Tables B.11 and B.12 respectively). As in evaluating the CFA data-model fits, the scalar models under all *by group* comparisons showed acceptable data-model fit based on the findings and recommendations of McNeish and colleagues (2018). Therefore, we believe that the scalar invariance for each of the *by group* comparisons is supported and structured means modeling could be used to compare latent means by each of the groupings.

Pre to Post Differences Within Each Institution

The pre to post latent mean differences for both self-efficacy factors at the three institutions are presented in Table 5.4. Pre to post latent mean differences for the TMT and CON factors for each institution are included Table B.13 in Appendix B. Each analysis was completed separately such that only one scale and one institution was modeled at a time, with the latent mean of the pre factor as the reference. This allowed for the difference in the pre to post latent factor means for each scale to be determined at the institution level. For reference purposes, the observed average pre score for each institution is also included in Table 5.4, however, as the latent mean differences represent a relative difference, these values cannot be used to determine the observed average post scale scores. For example, as shown in Table 5.4, the Southeastern institution had an observed pre score of 2.88 on the CSE scale and a latent mean difference of 1.49, which was a large effect (1.17). However, this data does ***not*** imply that the observed post score for this institution was 4.37 (i.e., 2.88 + 1.49).

Table 5.4. Pre to post latent mean differences for each institution. Bolded values indicate the difference was statistically significant ($p < 0.05$).

Scale	Institution	Responses, n	Observed Pre Score[a]	Pre to Post Latent Mean Difference (Effect Size)
Chemistry Self-efficacy (CSE)	Southeastern	265	2.88	**1.49** (1.17)
	Western	168	3.53	0.14 (0.14)
	Northwestern	551	3.34	**0.31** (0.28)
Academic Self-efficacy (ASE)	Southeastern	263	4.18	**-0.23** (0.19)
	Western	169	3.75	**-0.32** (0.26)
	Northwestern	554	3.77	**-0.71** (0.62)

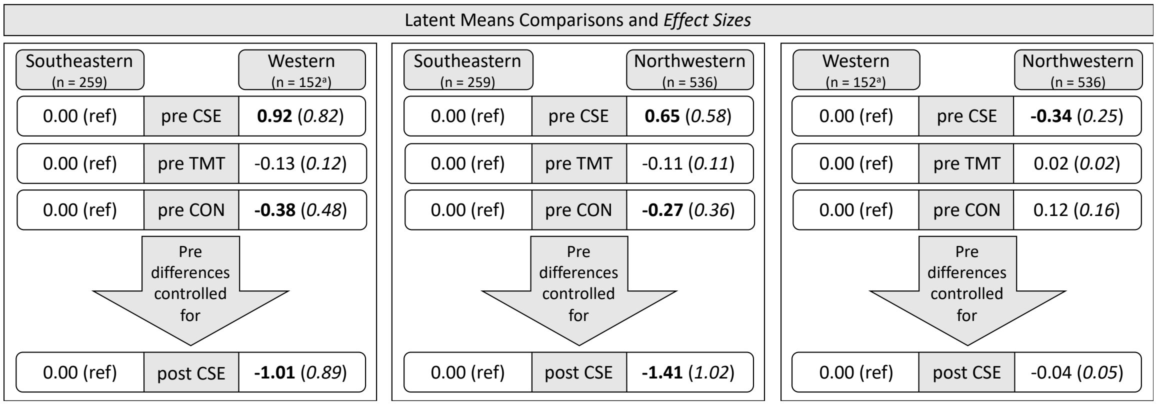
[a]Observed pre scale scores were calculated as an unweighted average of the items included in the final version of each scale.

Overall, the difference in pre to post latent means for the CSE factor showed a positive change for all institutions (Table 5.4). These differences were significant for the Southeastern and Northwestern institutions, with a large and small effect size, respectively. Although the Western institution also saw a small positive change, it was not significant. These results were in contrast to students' ASE scores overtime, which decreased significantly at all three institutions, although this change was only a small effect size at the Southeastern and Western institutions. Pre to post differences for TMT and CON were also examined and nonsignificant changes for most institutions were found (see Table B.13 in Appendix B). The exceptions to this were the Western institution, which showed a significant decrease in TMT from pre to post, and the Northwestern institution with an increase in CON. However, these differences only represented small effects.

Post Differences Between Institutions

The model used for pairwise comparisons of the post latent means of ASE and CSE between institutions included the respective pre factor (i.e., ASE or CSE), TMT, and

CON as covariates, such that they were controlled for when comparing the post factors (see Figure 5.2). As SMM only allows for relative differences to be determined, one of the institutions was used as the reference for each pairwise comparison and the latent mean differences represent the difference between the two institutions. For example, as shown in Figure 5.3, when compared to the Southeastern institution, the pre CSE latent mean for the Western institution was 0.92 *higher* and this difference was found to be a medium to large effect size (0.82). When pre CSE, TMT, and CON factors were taken into account as covariates, the post CSE latent mean difference for the Western institution was 1.01 *lower* when compared to the Southeastern institution, with a medium to large effect size (0.89). Latent mean differences for all pairwise comparisons of post CSE and ASE between institutions are presented in Figures 5.3 and 5.4, respectively.



Latent Means Comparisons and *Effect Sizes*

Southeastern (n = 259)		Western (n = 152[a])
0.00 (ref)	pre CSE	**0.92** (*0.82*)
0.00 (ref)	pre TMT	-0.13 (*0.12*)
0.00 (ref)	pre CON	**-0.38** (*0.48*)
	Pre differences controlled for	
0.00 (ref)	post CSE	**-1.01** (*0.89*)

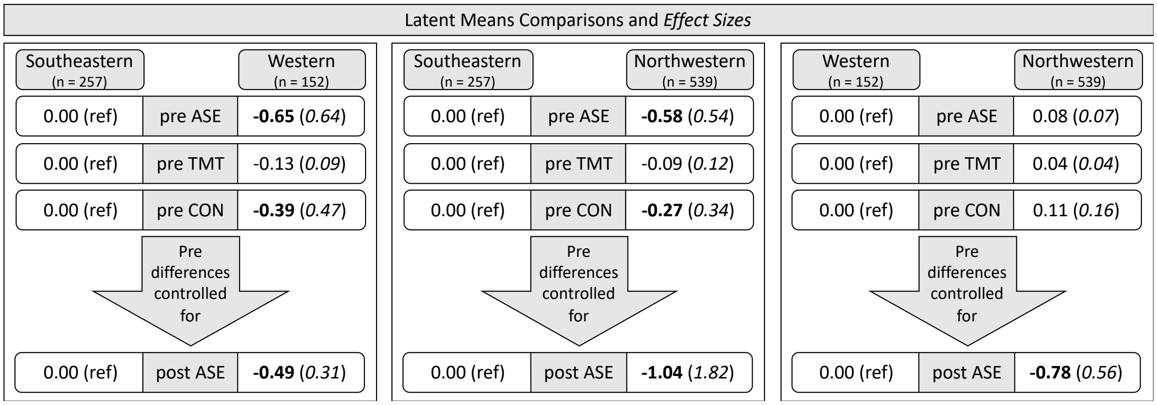
Southeastern (n = 259)		Northwestern (n = 536)
0.00 (ref)	pre CSE	**0.65** (*0.58*)
0.00 (ref)	pre TMT	-0.11 (*0.11*)
0.00 (ref)	pre CON	**-0.27** (*0.36*)
	Pre differences controlled for	
0.00 (ref)	post CSE	**-1.41** (*1.02*)

Western (n = 152[a])		Northwestern (n = 536)
0.00 (ref)	pre CSE	**-0.34** (*0.25*)
0.00 (ref)	pre TMT	0.02 (*0.02*)
0.00 (ref)	pre CON	0.12 (*0.16*)
	Pre differences controlled for	
0.00 (ref)	post CSE	-0.04 (*0.05*)

Figure 5.3. Pairwise post chemistry self-efficacy (CSE) latent mean differences between institutions with pre CSE, TMT, and CON factors as covariates. Each comparison is between two institutions while accounting for the pre latent means. The listed reference institution was used as the reference group for the designated pairwise analysis. Bolded values indicate the difference was statistically significant ($p < 0.05$). [a]This data set included one dummy response pattern to account for missing response categories. See Appendix B for details.

When comparing post CSE latent means (Figure 5.3) between the Southeastern and the Western and Northwestern institutions, the Southeastern institution was found to have a higher post CSE latent mean than both of the other institutions, each with a large effect size. These differences accounted for the higher pre CSE latent means of the

Western and Northwestern institutions when compared to the Southeastern institution. Although a pre CSE latent mean difference was also found between the Western and Northwestern institutions, the post CSE latent mean difference was small and not significant.



Latent Means Comparisons and *Effect Sizes*

Southeastern (n = 257)		Western (n = 152)
0.00 (ref)	pre ASE	**-0.65** (*0.64*)
0.00 (ref)	pre TMT	-0.13 (*0.09*)
0.00 (ref)	pre CON	**-0.39** (*0.47*)
	Pre differences controlled for	
0.00 (ref)	post ASE	**-0.49** (*0.31*)

Southeastern (n = 257)		Northwestern (n = 539)
0.00 (ref)	pre ASE	**-0.58** (*0.54*)
0.00 (ref)	pre TMT	-0.09 (*0.12*)
0.00 (ref)	pre CON	**-0.27** (*0.34*)
	Pre differences controlled for	
0.00 (ref)	post ASE	**-1.04** (*1.82*)

Western (n = 152)		Northwestern (n = 539)
0.00 (ref)	pre ASE	0.08 (*0.07*)
0.00 (ref)	pre TMT	0.04 (*0.04*)
0.00 (ref)	pre CON	0.11 (*0.16*)
	Pre differences controlled for	
0.00 (ref)	post ASE	**-0.78** (*0.56*)

Figure 5.4. Pairwise post academic self-efficacy (ASE) latent mean differences between institutions with pre ASE, TMT, and CON factors as covariates. Each comparison is between two institutions while accounting for the pre latent means. The listed reference institution was used as the reference group for the designated pairwise analysis. Bolded values indicate the difference was statistically significant ($p < 0.05$).

Post ASE latent mean differences (Figure 5.4) indicated that students at the Southeastern institution had the highest post ASE, with students at the Northwestern institution having the lowest post ASE. Pre ASE at the Southeastern institution was also higher than the other two institutions, with no difference between the pre ASE latent means of the Western and Northwestern institutions.

Discussion

Pre to Post Differences Within Each Institution

The increases in CSE latent means from pre to post for all institutions suggest that students perceived their chemistry ability to be higher at the end of the term than at the beginning. As the items used for the CSE scale are based on specific tasks students are expected to accomplish in general chemistry (i.e., “How well can you interpret chemical equations?”), it makes sense that students would generally have a higher CSE at the end of the term. Increases in students’ chemistry self-efficacy throughout the term has also been found in previous studies of non-flipped general chemistry courses (Ferrell & Barbera, 2015; Graham et al., 2019; Villafañe et al., 2014; Vishnumolakala et al., 2017). Although positive changes in CSE were seen, ASE latent mean differences were found to be significantly lower from pre to post for all institutions. The effect size of this difference for the Southeastern and Western institutions represented a small effect, while the difference for the Northwestern institution represented a medium effect. A decrease in ASE over the term has been reported in other studies (e.g., Young et al. (2018)). In contrast to the CSE items, the items included on the ASE were targeted toward general statements about the courses a student was taking, not just their chemistry course (i.e., “I expect to do well in my courses.”). Self-efficacy scales that are more specific (i.e., task-based statements) have been shown to be a better predictor of academic performance than more general self-efficacy scales (Choi, 2005). Thus, the difference in specificity between the CSE and ASE items could have contributed to how chemistry and academic self-efficacy trended in different directions throughout the term.

Post Differences Between Institutions

Differences in post CSE latent means were seen between the Southeastern institution and both the Western and Northwestern institutions. The pre latent mean comparisons between the Southeastern institution and the other two institutions also showed significant differences, with the Southeastern institution having a lower latent mean for pre CSE and a higher latent mean for pre CON. Thus, although students at the Southeastern institution initially had lower CSE than students at both the Western and

Northwestern institutions, Southeastern institution students had the highest reported CSE at the end of the term. The pairwise comparison between the Western and Northwestern institutions also showed a significant difference between pre CSE latent means, with the Western institution having a higher pre CSE; however, the post CSE latent means showed no significant difference between the two institutions. Although there were some initial differences between the pre CSE and CON latent means between all the institutions, the pre factors were included as covariates in the model and the latent mean differences for post CSE account for any differences the students may have had in their incoming time management, concentration, or chemistry self-efficacy. However, even though these pre factors were accounted for in the model, other possible confounds could have influenced the results, such as differences in the class structure and differences in demographics.

As this study was completed across multiple institutions, there may have been course differences that contributed to the measured post CSE differences between institutions. In flipped courses there are two main aspects that are usually incorporated: information is provided to students through pre-class materials (PCMs), which is then reinforced through active learning during the face-to-face (F2F) time. The differences and similarities of these two aspects for the courses at these institutions were detailed in a prior study (Naibert et al., 2020), here we address the most salient features. The PCMs for all institutions were in the form of online videos, however, there were slight details that differed, such as instructor-curated versus instructor-created, video length, etc. Results indicated that a higher percentage of students at the Western and Northwestern institutions reported watching all of the videos compared to students at the Southeastern institution, where most students reported watching only some of the videos. The

Northwestern institution also had a significantly higher percentage of students report that they utilized the PCMs before the F2F time compared to the other institutions. In addition, the structure of the F2F environments differed between institutions, with the Southeastern institution including more student and instructor questioning (~80% of time-blocks) than the Western and Northwestern institutions (~20% of time-blocks each), which incorporated more groupwork into their F2F time (Figure B.1 in Appendix B). While these course-level differences in PCMs and F2F time cannot be said to be the cause, it is possible that these variations in how the classes were structured influenced students' CSE. Although one study has found that including POGIL discussion sections in general and organic chemistry did not have a significant effect on students' CSE over traditional discussion sections (Chase et al., 2013), different course structures have been found to influence students' performance, as well as the time they spent preparing for class, with moderate-structured courses having a larger impact on these variables (Eddy & Hogan, 2014). Others have found, in a more controlled study with case-based learning, that a gradual shift to more autonomous active learning environments over a semester benefit students' motivation and learning compared to an abrupt shift (Baeten et al., 2013). Also, group work as a constructivist practice needs scaffolding (Hanrahan, 1998) as these practices can support or undermine student motivation (Pintrich et al., 1993). Therefore, the potential impact of the flipped course structure cannot be ruled out when considering the differences in post-term CSE values.

Demographic differences could have also contributed to the differences seen in post CSE. However, due to the small group-level sample sizes at some of the institutions (Table B.14), differences based on minority status and gender could only be evaluated using the aggregated data set. In doing so, it was found that both male and female students increased in CSE from pre to post (Table B.15) and that there was no significant difference in post CSE factors based on gender (Table B.16). However, results indicated that although both non-URM and URM groups increase in CSE from pre to post (Table B.17), URM students reported significantly higher post CSE than non-URM students (Table B.16). While other studies have also found differences in CSE by demographic group (Villafañe et al., 2014), it should be noted that the differences seen in this study could be influenced by the demographic profiles of the institutions themselves. The Southeastern institution had a larger percentage of URM students than the Western and Northwestern institutions, which had equal percentages of URM and non-URM and a larger percentage of non-URM students, respectively. Therefore, as the Southeastern institution was found to have a higher post CSE latent mean than the Western and Northwestern institutions and the URM group was also found to have a higher post CSE latent mean than the non-URM group, these results could be conflated. Since the sample sizes for the different groups were not large enough to complete SMM analyses on institutional subsets (Table B.14), it is unknown whether the differences were due to course-level differences or the different demographic profiles of the institutions.

Latent mean differences of post ASE between institutions were also significant, with small to medium effect sizes. The Southeastern institution was found to have the highest post ASE latent mean, with the Northwestern institution having the lowest. However, as mentioned earlier, the items used to assess ASE were more general than the CSE items and related to all the classes the students were taking. Thus, it is unknown whether the differences between the chemistry courses, which may have only been one of many courses a student was taking, influenced these results. When the aggregated data set was analyzed for demographic differences, both non-URM and URM students were found to decrease on ASE from pre to post (Table B.17), with URM students having a higher post ASE latent mean than non-URM students (Table B.18). Thus, it is unknown whether the differences in post ASE could be a result of course differences, institutional differences, or demographic differences across the institutions.

Conclusions, Limitations and Implications

This project investigated the self-efficacy of students enrolled in general chemistry courses structured within flipped learning environments. The conclusions from this multi-institutional investigation are framed by our research questions.

What evidence supports the validity and reliability of the data generated from the coordinated assessments at our sites?

Measures of self-efficacy and self-regulation were administered and their data evaluated for structural validity and single-administration reliability via Confirmatory Factor Analysis (CFA). The final CFA models included a reduced set of items for each measure, which were found to have consistently strong factor loadings across administration times. All models had acceptable data-model fit and reliability. As Structured Means Modeling (SMM) was used to compare the latent means of each measure pre-post and by institution, gender, and URM status, the scalar invariance of each was evaluated. Measurement invariance analyses help to support that measures are equally made across groups prior to comparing their results (Rocabado et al., 2020). As the data from each measure were treated as ordinal, evaluating scalar invariance required supporting the data-model fit with both the factor loadings and response thresholds fixed across groups. Each scalar invariance model (i.e., longitudinal, by institution, by gender, and by URM status) showed acceptable data-model fit.

How do students' self-efficacy and self-regulation change within each flipped learning environment? and How do changes in each construct compare across sites?

With regard to self-regulation (i.e., the TMT and CON measures), when differences from pre to post were detected, they were small effects. However, when examining differences across institutions for the pre measures, several differences were found between institutions (Figures 5.3 and 5.4). Thus, the pre-TMT and pre-CON factors were added as controls for the analyses of CSE and ASE. The evaluation of pre to post SMMs for CSE and ASE produced disparate results. Students at each institution reported significant decreases in ASE and increases in CSE at the end of the term. This may be due to the task-based focus of the CSE items compared to the more general academic focus of the ASE items. As the CSE scores showed an increase over the term for all institutions, and since more specific self-efficacy measures have been found to be a better predictor of performance than general self-efficacy measures (Choi, 2005), the decrease in ASE scores may not be representative of a change in students' self-efficacy as a result of the structure of their chemistry course. Therefore, as the focus of this study was to explore differences between different flipped environments, the CSE measure was examined in more depth.

While two out of the three institutions showed significant CSE increases, students at the Western institution showed a small, nonsignificant, increase in CSE. This nonsignificant change could be due to the term of the course that was surveyed. At the Southeastern and Northwestern institutions, the surveyed courses were the first-term of general chemistry, whereas, at the Western institution it was the second-term (Table 5.2). It is possible that the students in the second-term course have already had experiences informing their CSE by the beginning of this course and thus, their CSE did not change significantly by the end. Between the two institutions that consisted of the first-term general chemistry courses, students at the Southeastern institution showed the largest increase in CSE and had the highest post CSE. To better explore the difference between these courses, the in-class structures were examined. In a prior study (Naibert et al., 2020), the instructional practices at these institutions were found to vary based on the structure of the F2F active learning techniques employed and students' reported use of the PCMs. The F2F structure of the Southeastern institution course primarily included instructor-student interactions in the form of whole-class questioning, whereas the Northwestern institution course relied heavily on peer-to-peer interactions during groupwork (see Figure B.1 in Appendix B). These differences in F2F structure could have been a contributing factor to the differences seen in post CSE. Differences in the amount of structure included in a task has been found to contribute to differences in students' self-efficacy on those tasks in secondary classrooms (Lodewyk & Winne, 2005). In this prior study, both a "well-structured" task and an "ill-structured" task were provided to the students with the difference between the tasks described as varying "in the structural cues they provided for students." When students' self-efficacy during those tasks was measured, they found that students reported significantly higher self-efficacy when they were working on the well-structured task compared to the ill-structured task. In our study, it could be argued that the course at the Southeastern institution, consisting of primarily instructor-guided questioning during F2F sessions, may have provided more “structured” tasks to students than the predominant use of peer-to-peer small group interactions found in course at the Northwestern institution. Given this result, future studies would be needed to further test the impacts of these types of structural differences in a learning environment.

When exploring the benefits of flipped learning environments on individual differences, both males and females reported increases in CSE over the semester and there were no by gender differences detected at the end of the term. In regard to minority status, the results were a bit more complicated. URM differences were detected, although a potential confound by institution could be at play, since the majority of students with URM status were at the Southeastern institution, thus conflating a potential difference. However, previous studies have found differences in students’ CSE based on demographic profiles. For example, Villafane et al. (2014) found that even though most demographic groups showed an overall increase of CSE over the term, Black and Hispanic males reported a decrease in CSE. Thus, further research into differences and changes of CSE based on demographic profiles may be beneficial, if there is a large enough sample size to explore these differences at the institution level.

Overall, while students in each of the courses reported higher CSE at the end of the term, the study was not designed to evaluate which structural features may have led to the differential increases detected. Bandura postulated that one’s self-efficacy is derived from four experiential sources of information: mastery experiences, vicarious learning, social persuasion, and psychological state (Bandura, 1997). We reflect on these sources to postulate on why the experiences of the students in the predominantly whole-class questioning F2F environment might have led to higher self-efficacy than those in the peer-to-peer small groupwork environment.

Mastery experiences require that individuals experience success, or failure, in a task (Bandura, 1997). Therefore, whole-class questioning may provide more *individual* opportunities to experience success (or failure), compared to small groupwork. Students may find more value in the frequent instructor feedback that occurs with whole-class questioning compared to less feedback during longer groupwork activities (Wiggins et al., 2017). Vicarious experiences occur through seeing a peer perform a task (i.e., modeling success) or in comparing one’s own performance to that of others (i.e., comparative success) (Bandura, 1997). While small groupwork, in theory, should provide consistent opportunities for both, this may be highly dependent on the makeup of a group, its discourse, and how it is facilitated by learning assistants or the instructor (Chapman & van Auken, 2016). It cannot be assumed that groupwork is equally supportive for all members (Chang & Brickman, 2018). While whole-class questioning may not provide many opportunities to observe peer success (i.e., modeling success), each individual should at least have the chance to compile their own answers and compare them to those discussed (i.e., comparative success). Group dynamics may also encourage or discourage the social persuasion experiences (i.e., messages about ability) (Bandura, 1997) of students. Individuals in groups with established and well facilitated group-norms may receive more supportive feedback than those in groups dominated by one or more individuals (Chapman & van Auken, 2016; Oakley et al., 2004). In contrast, students may experience supportive social persuasion (Bandura, 1997) when the answers to instructor-initiated (i.e., whole-class or clicker) questions are discussed, as these types of questions are typically followed up with clarifying information to support learners understanding (Naibert et al., 2020). Lastly, negative feelings (e.g., stress or anxiety) in a learning situation may be interpreted as an indicator that one is not capable (Sawtelle et al., 2012). Therefore, the feelings associated with groupwork (Cantwell & Andrews, 2002; Livingstone & Lynch, 2002; Shortlidge et al., 2019) or group relationships (Lavy, 2016) may not be supportive of the self-efficacy development of all students. In concluding, Murphy and colleagues (2011) posited in their review that the success of discussions with regard to learning and motivation is less about small groups or instructor-led, but more about the level of structure provided during the discussion sessions.

This study, and many more in the extant literature within discipline-based education research, document the quantitative impacts of a learning environment on students' self-efficacy. However, few have actually studied what types of self-efficacy opportunities (SEOs) actually exist within a given learning environment. One study in physics did investigate the SEOs provided through the interactions among three learners performing a task from the Modeling Instruction (Brewe, 2008) curriculum (Sawtelle et al., 2012). This observation-based study was able to identify a variety of SEOs while performing the task. Given the broad ways that active learning can be defined (Freeman et al., 2014), or that flipped courses can be structured (Naibert et al., 2020), these types of in-depth observational designs might be needed if researchers or practitioners wish to understand the nature of detected differences in self-efficacy across learning environments. Therefore, it is recommended that further research on flipped learning environments continue to account for the structural components connected to F2F active learning to examine each environment's specific benefits to students' motivation and learning, while controlling for different elements.

Limitations

This nonexperimental research has several limitations that should be considered when interpreting the outcomes presented. The outcomes are based on voluntary and self-reported student data and therefore only reflect the results of those students who agreed to participate in the study. As such, the data may not reflect the outcomes of other students, especially for cases where lower response rates were obtained. While the pre-score comparisons did not detect differences between students who appeared in the matched dataset and those who did not, other unmeasured factors could not be ruled out given the design of the study. Within any self-report study, students' responses could be influenced by social desirability; that is, students might respond on the basis of what would make them "look best". However, as no data from this study was collected within the authors' institutions and none of the course instructors were involved in the data collection process, this influence was potentially diminished as the research team had no connections to the students. A potential confounding aspect with regard to the consistency (or priming) of students' responses in this study may come from item- and/or scale-order effects, as neither were randomized. To reduce any potential order-effects, future researchers are encouraged to randomize their administrations at both levels. Finally, as changes in self-efficacy were measured from pre to post, students' self-reported pre-term self-efficacy could be over-estimated based on their perceived incoming ability, which may be more targeted by the end of a term. However, these potential discrepancies would not impact the post self-efficacy comparisons conducted within this project, as these were not ‘gain scores’ but comparisons of students reported self-efficacy at the end of each course at each institution.

While this study employed a coordinated set of measures, and validated the data produced within each environment, these measures may not be supported for use in other course types or institutions. Therefore, those interested in conducting similar analyses are encouraged to support the validity of their data as appropriate (Arjoon et al., 2013; Knekta et al., 2019). This study utilized latent means to make comparisons among different groups. While the scalar invariance models of each measure were supported, SMM comparisons by gender and by URM status could only be conducted using data from all institutions combined. This was due to the low number of students within one or more groups at certain institutions. For example, at the Southeastern institution, only 26 students (10% of the pre-post matched data) were categorized as non-URM (i.e., non-Latino/a White or Asian), with 211 students (80%) reporting as Latino/a. Therefore, not only was there an insufficient number of non-URM students to conduct an intra-institution comparison, there was also an insufficient number of Latino/a students in the other datasets to support inter-institution comparisons at this specific level. Future studies interested in exploring these measured outcomes by demographic groups within a single student population are encouraged to not only seek to collect data in large-enrollment course environments, but also those with more balanced demographics, such that large enough group-level populations are available. Another strategy would be to oversample students of minority status in order to conduct analyses based on race/ethnicity stratifications.

Finally, while evidence of the structural validity, single-administration reliability, and consequential validity (via scalar measurement invariance) of the data from this study were provided, some items from each measure were flagged, evaluated, and subsequently dropped to produce the final models. These decisions were based on analysis of the *a priori* initial CFA model data; response process validity interviews were not conducted. This type of qualitative data could have provided insights to the functioning of the flagged items. However, this was beyond the scope of this multi-year and multi-institutional study. In future uses of these measures, qualitative data should be gathered to evaluate if the dropped items can be improved upon and therefore retained.

Implications

In contrast to an increase in CSE over the term at all institutions, students' ASE was found to decrease. The main difference between these two measures was the specificity of the items. Whereas the CSE measure included specific task-based items, the ASE items were more general and referred to all of a students' courses that they were taking. This brings into light the importance of the specificity of the items when assessing self-efficacy. Other studies which measured students' CSE with task-specific items also found increases in CSE over the term (Ferrell & Barbera, 2015; Graham et al., 2019; Villafañe et al., 2014; Vishnumolakala et al., 2017), whereas a study that used a measure with more general items found that students' self-efficacy decreased by the end of the term (DiBenedetto & Bembenutty, 2013). Therefore, when self-efficacy is assessed, or prior studies are interpreted, it is important to keep in mind the specificity of the items to ensure that they align with the goals of the study. As more task-specific measures have been found to be better predictors of performance (Choi, 2005) and future success

(Bandura, 2006), it may be more beneficial to use a task-specific measure when assessing self-efficacy at the course-level.

It is important that the validity and reliability of the data collected with a measure in a new environment are assessed, even if the measure has been previously shown to produce good data. In this study, evidence of both structural validity and single administration reliability were gathered for the data collected with each of the measures and at each of the institutions. Even if evidence of structural validity is found, group comparisons are not recommended without additional evidence of consequential validity (Rocabado et al., 2020). Evidence of consequential validity is gathered through the evaluation of different levels of measurement invariance. If latent means are to be compared, scalar invariance of the different groups should be established. However, if observed scores are to be compared across groups, then strict invariance is recommended. Without evidence of scalar or strict invariance, comparisons between the groups would not be supported (Rocabado et al., 2020). The requirement of measurement invariance necessitates a reasonably large sample size with relatively equal populations in the different groups. For example, although it may have been beneficial in this study to compare latent means based on URM status *within* the different institutions, this analysis was limited by the sample sizes of these different populations within each institution (e.g., the Southeastern institution only had 26 non-URM students) and so were only assessed at the aggregate level where scalar measurement invariance could be established. Therefore, future studies that wish to focus on group differences within a factor analysis framework are encouraged to consider the sample sizes of the individual groups.

In this study, students' CSE was detected to increase over the term for all three institutions, suggesting that the students were more confident in their abilities by the end of the term. It should be noted that studies of other chemistry classrooms (Ferrell & Barbera, 2015; Villafañe et al., 2014) and active learning environments (Vishnumolakala et al., 2017) have also found increases in CSE over the term. Thus, the inclusion of a flipped classroom structure cannot be said to be the cause of these increases and did not seem to negatively affect students' CSE. Within a flipped learning environment, students are provided with the opportunity to initially engage in the course material before coming to class, leaving the F2F time for exploration of the material in a variety of manners. In this study, each of the three institutions structured their F2F time differently. The Southeastern institution primarily focused on student-instructor interactions through whole-class questioning, while the Northwestern institution included more peer-to-peer groupwork. Since significant differences were found in students' post CSE between these two institutions, instructors who flip their course are encouraged to consider the active learning techniques that will be incorporated during the F2F time. Considering that the structure of the F2F time may lead to different student outcomes. Some demographic groups have been shown to increase more in their performance outcomes than other groups when additional structure is added to the course (e.g., Eddy and Hogan (2014)).

Associated Content

Supporting Information

The Supporting Information is available in Appendix B and includes CFA data for scale modifications, protocol for unused response categories, descriptive statistics, supplemental measurement invariance and structured means modeling tables, course observation.

Author Information

Corresponding Author

*E-mail: jack.barbera@pdx.edu

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. (DUE 1611220 and 1611519). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This project would not have been possible without the participation of each instructor and the students in their flipped courses, we are therefore very thankful for their assistance in this study.

Chapter 6: Modifying the ASPECT Survey to Support the Validity of Student Perception Data from Different Active Learning Environments

Reprinted with permission from Naibert, N., Shortlidge, E.E., & Barbera, J. (2021). Modifying the ASPECT Survey to Support the Validity of Student Perception Data from Different Active Learning Environments. *Journal of Microbiology and Biology Education, 22*(3), 1-14. Copyright 2021 Naibert et al.

Abstract

Measuring students' perceptions of active learning activities may provide valuable insight to their engagement and subsequent performance outcomes. A recently published measure, the Assessing Student Engagement in Class Tool (ASPECT), was developed to assess student perceptions of various active learning environments. As such, we sought to use this measure in our courses to assess student perceptions of different active learning environments. Initial results analyzed with confirmatory factor analysis (CFA) indicated that the ASPECT did not function as expected in our active learning environments. Therefore, prior to administration in an introductory biology course that incorporated two types of active learning strategies, additional items were created and the wording of some original items were modified to better align with the structure of each strategy, thereby producing two modified ASPECT (mASPECT) versions. Evidence of response process validity of the data collected were analyzed using cognitive interviews with students, while internal structure validity evidence was assessed through exploratory factor analysis (EFA). When data were collected after a "Deliberative Democracy" (DD) activity, 17 items were found to contribute to three factors related to 'personal effort', 'value of the environment', and 'instructor contribution'. However, data collected after a "clicker" day resulted in 21 items that contributed to four factors; three similar to the DD activity and a fourth related to ‘social influence’. Overall, these results suggest that the same measure may not function identically when used within different types of active learning environments, even with the same population, and highlights the need to collect data validity evidence when adopting and/or adapting measures.

Keywords

Engagement, structural validity, factor analysis, active learning, undergraduate

Introduction

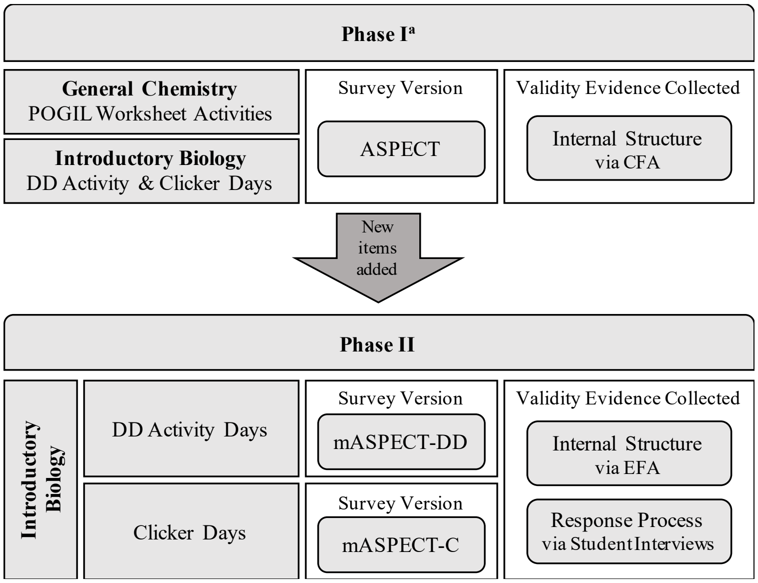
The continued shift in undergraduate science courses from instructor-centered classrooms to student-centered learning has been influenced in part by national reports aimed at improving higher education within the science, technology, engineering, and mathematics (STEM) fields (National Research Council, 2012; President’s Council of Advisors on Sciences and Technology (PCAST), 2012). Many studies have found that including active learning strategies in the classroom positively impacts student outcomes (e.g., higher exam grades, lower withdrawal rate, etc.) (Freeman et al., 2014; Rahman & Lewis, 2019). However, while including these strategies may increase student performance outcomes, the extent of these benefits may vary in different student populations (Eddy & Hogan, 2014) and it cannot be assumed that every student in the classroom engages in or benefits from an active learning environment to the same extent (Wiltbank et al., 2019). Since active learning strategies are inherently student-centered, it is up to the student to decide to interact with and “buy-in” to the activity and learning environment (Cavanagh et al., 2016). Student buy-in, along with other perceptions such as trust in the instructor and growth mindset, have been shown to influence students’ engagement and course outcomes (Cavanagh et al., 2016; Cavanagh et al., 2018). Thus, measuring students' perceptions of active learning environments could provide valuable information about how students engage with and benefit from different active learning environments.

Measuring student perceptions of active learning environments

As multiple types of active learning strategies are implemented in our classrooms at Portland State University (PSU), we were interested in measuring students' perceptions of these various environments. Although individual student perceptions can be gathered through qualitative methods (e.g., Shortlidge et al. (2019)), quantitative methods, such as a self-report survey, can be used to easily and efficiently collect perceptions from every student in the class. Recently, the Assessing Student Perspective of Engagement in Class Tool (ASPECT), was developed by Wiggins et al. (2017) to measure students' perceptions of their cognitive and affective engagement in different active learning environments incorporated in a large-format introductory biology classroom. Their results suggested that student perceptions of the value of the activity and the instructor contribution differed based on the activity type (i.e., students perceived there to be less value and less instructor contribution during worksheet activity days compared to clicker-question activity days), as well as demographic group. No significant differences were detected for student perceptions of their personal effort across different activity types.

Active learning environments can vary, even between classes that implement the same active learning strategy; therefore, evidence of validity and reliability of data generated by an instrument should be gathered before interpreting any results in a different environment and/or with a different population (Knekta et al., 2019). The types and amount of validity evidence collected for a measure depend on the goals of the project as well as what types of validity evidence had previously been assessed. Collecting evidence of internal structure validity of a previously published measure provides evidence that the constructs are being measured in a similar way in the different learning environment (Knekta et al., 2019). Additionally, gathering evidence of response process validity can provide confidence that students are interpreting the items correctly in the new environment (Arjoon et al., 2013; Knekta et al., 2019), especially if modifications are made to the original measure.

We evaluated the ASPECT in our learning environments through two experimental phases. Phase I focused on gathering evidence of internal structure validity for data collected with the original ASPECT measure in our learning environments. The initial results from Phase I led to modifications of the ASPECT (results and details from Phase I are included in Appendix C). This chapter will focus on Phase II, in which the modified ASPECT (mASPECT) was used to measure student perceptions of two different types of active learning strategies. Since the mASPECT included additional items, evidence of both internal structure and response process validity were gathered. An overview of Phases I and II, ASPECT versions, and types of validity evidence collected is shown in Figure 6.1.



[a]Details and results from Phase I are included in Appendix C.

Figure 6.1. Overview of the active learning strategies, survey versions, and validity evidence collected during Phase I and Phase II.

We hypothesized that data collected with the mASPECT during Phase II would show evidence of similar factors related to student perceptions that were discovered with the original ASPECT (i.e., personal effort, value of the activity, instructor contribution), as well as an additional group-related factor. To this end, evidence of response process and internal structure validity and reliability of the data collected with the mASPECT in two active learning environments was gathered and the resulting survey structures and scale scores were evaluated in both environments. This work sought to answer two research questions: 1) What modifications could be made to the ASPECT to obtain sufficient evidence of internal structure validity of the collected data? and 2) What factor structure best represented the modified ASPECT (mASPECT) data from our active learning environments? Answering these questions would provide support for the student perception data from our course and could serve as a model for others seeking to use the ASPECT or mASPECT when evaluating their active learning environments.

Methods

Course information and active learning environments

This study was completed in a third-term introductory biology course at Portland State University (PSU) with a week-1 enrollment of 266 students. Demographic information of students who consented to participate in this study is provided in Table C.1 in Appendix C. Two types of active learning strategies were assessed within the same class; 1) Deliberative Democracy (DD) modules; and 2) classroom response systems (clickers). DD is a small-group active learning strategy that includes a multi-day deliberation exercise where students are introduced to a real-world problem that correlates with their course content, and through reading, deliberation, and research they are asked to come to a consensus on a policy recommendation (Komperda, Barbera, et al., 2018; Rain-Griffith et al., 2020; Weasel & Finkel, 2016). In this study, the DD activity consisted of a two-day module where students gathered information on their own between DD activity days and brought the information back to class to inform group discussion and consensus making. Students were assigned readings, quizzes, and group worksheets to build a consensus statement. Students worked in the same randomly assigned groups of 3 – 5 on DD activity days, and the professor, graduate Teaching Assistant (TA), and multiple undergraduate Learning Assistants (LAs) (~15) facilitated the group work. The TA functioned in an instructor role during DD activity days and trained the LAs in each DD activity. The second active learning strategy investigated were clicker days. These were 'normal' lecture days where students were regularly encouraged to 'think-pair-share' with other students nearby in response to clicker prompts given by the professor. Although no undergraduate LAs were in class during the clicker days, the graduate TA was present.

All data collected within this study was approved by the Institutional Review Board (IRB no. 196410-18) at Portland State University and appropriate consent was gathered from instructors and students as required by the IRB.

Survey items

The surveys administered in both environments consisted of a modified ASPECT (mASPECT) survey based on the original ASPECT (Wiggins et al., 2017). Two versions of the mASPECT were created; one for a DD activity day (mASPECT-DD) and one for a clicker day (mASPECT-C). The modifications to the surveys included minor wording changes to the 19 original ASPECT items (Wiggins et al., 2017), as well as the creation of new items based on the structure of the active learning environments and the results from Phase I (details provided in Appendix C). The mASPECT-DD version contained 35 items and the mASPECT-C version contained 31 items. Both versions included the 19 original ASPECT items in a slightly modified form (Items 1 – 19, Table 6.1), 8 new items related to personal effort (Items 20 – 27, Table 6.2), and 4 new items related to group function (Items 28 – 31, Table 6.2). The four-item difference between mASPECT-DD and mASPECT-C versions was due to the addition of ‘LA-worded’ items (Items 13B, 14B, 15B, and 16B, Table 6.1) that paralleled the ‘Professor/TA’ items. As the LAs were not present during the clicker day, the items were not applicable to that environment. All survey items were administered on a 6-point Likert-type scale from *strongly disagree* (1) to *strongly agree* (6).

Table 6.1. Original ASPECT items (Wiggins et al. (2017)) (Items 1 – 19) and modifications for mASPECT-DD and mASPECT-C versions. Shading is included to indicate the original ASPECT factors: personal effort (PE), value of group activity (VGA), and instructor contribution (IC). Wording differences between mASPECT-DD and mASPECT-C are underlined.

Factor	ASPECT wording	Item	mASPECT-DD wording	mASPECT-C wording
PE	I was focused during today's group activity.	1	I was focused during today's class.	I was focused during today's class.
	I worked hard during today's group activity.	2	I worked hard during today's class.	I worked hard during today's class.
	I made a valuable contribution to my group today.	3	I made valuable contributions when working with other students during today's class.	I made valuable contributions when having discussions with other students during today's class.
VGA	Explaining the material to my group improved my understanding of it.	4	Explaining the material to my group members improved my understanding of it.	Explaining the material to other students improved my understanding of it.
	Having the material explained to me by my group members improved my understanding of the material.	5	Having the material explained to me by my group members improved my understanding of it.	Having the material explained to me by other students improved my understanding of it.
	Group discussion during the activity contributed to my understanding of the course material.	6	Working with other students during today's class contributed to my understanding of the material.	Discussion with other students during today's class contributed to my understanding of the material.
	Overall, the other members of my group made valuable contributions during the group activity.	7	The students I worked with made valuable contributions during today's class.	The students I had discussions with made valuable contributions during today's class.
	I had fun during today's group activity.	8	I had fun during today's class.	I had fun during today's class.
	I would prefer to take a class that includes this [topic] group activity over one that does not include this [topic] activity.	9	I would prefer to take a class that included today's activity over one that does not include it.	I would prefer to take a class that included today's clicker questions over one that does not include them.
	I am confident in my understanding of the material presented during today's group activity.	10	I am confident in my understanding of the material presented during today's class.	I am confident in my understanding of the material presented during today's class.
	The group activity increased my understanding of the course material.	11	Today's class increased my understanding of the material.	Today's class increased my understanding of the material.
	The group activity stimulated my interest in the course material.	12	Today's class stimulated my interest in the course material.	Today's class stimulated my interest in the course material.

Table 6.1 cont.

Factor	ASPECT wording	Item	mASPECT-DD wording	mASPECT-C wording
IC	The instructor's enthusiasm made me more interested in the group activity.	13A	The Professor/Teaching Assistant's enthusiasm made me more interested in today's class.	The Professor/Teaching Assistant's enthusiasm made me more interested in today's class.
		13B	The Learning Assistant's enthusiasm made me more interested in today's class.	*n/a*
	The instructor put a good deal of effort into my learning for today's class.	14A	The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.	The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.
		14B	The Learning Assistant put a good deal of effort into my learning for today's class.	*n/a*
	The instructor seemed prepared for the group activity.	15A	The Professor/Teaching Assistant seemed prepared for today's class.	The Professor/Teaching Assistant seemed prepared for today's class.
		15B	The Learning Assistant seemed prepared for today's class.	*n/a*
	The instructor and TAs were available to answer questions during the group activity.	16A	The Professor/Teaching Assistant was available to answer questions during today's class.	The Professor/Teaching Assistant was available to answer questions during today's class.
		16B	The Learning Assistant was available to answer questions during today's class.	*n/a*
n/a	I felt comfortable with my group.	17	I felt comfortable working with other students during today's class.	I felt comfortable having discussions with other students during today's class.
	I knew what I was expected to accomplish during the group activity.	18	I knew what I was expected to accomplish during today's class.	I knew what I was expected to accomplish during today's class.
	One group member dominated discussion during today's group activity.	19	One of the students I worked with dominated discussion during today's class.	One of the students I had discussions with dominated discussion during today's class.

Table 6.2. New survey items (Items 20 – 31) created for mASPECT-DD and mASPECT-C related to personal effort and group function. Wording differences between mASPECT-DD and mASPECT-C are underlined.

Item	mASPECT-DD wording	mASPECT-C wording
20[a]	I completed the pre-work for today's class.	I completed the pre-work for today's class.
21[a]	I did not make much of an effort during today's class.	I did not make much of an effort during today's class.
22[a]	I guessed or made stuff up so that I could finish today's activity.	I guessed or made stuff up so that I could finish today's activity.
23[a]	I skipped or guessed on the hard parts of today's activity.	I skipped or guessed on the hard parts of today's activity.
24[a]	I found it difficult to maintain my concentration during today's class.	I found it difficult to maintain my concentration during today's class.
25[a]	I tried to relate today's class to prior material from the course.	I tried to relate today's class to prior material from the course.
26[a]	I was not very engaged in today's class.	I was not very engaged in today's class.
27[a]	I was fully engaged in today's class.	I was fully engaged in today's class.
28[b]	The students I worked with were focused during today's class.	The students I had discussions with were focused during today's class.
29[b]	The students I worked with worked hard during today's class.	The students I had discussions with worked hard during today's class.
30[b]	The students I worked with had fun during today's class.	The students I had discussions with had fun during today's class.
31[b]	Each student I worked with made an equal contribution during today's class.	Each student I had discussions with made an equal contribution during today's class.

[a]Personal effort related items. [b]Group related items.

Quantitative methods

Quantitative survey data were collected after both a final day of a DD activity and after a clicker day. Students were notified of the surveys during in-class announcements, as well as an announcement posted on the course's learning management site with a link to the Qualtrics survey. Students were given 48 hours following completion of the in-class activity to access and complete the survey. Students who accessed the survey were given a nominal amount of extra-credit regardless of consent or completion. Before analysis, the responses were cleaned by removing 1) students that did not consent, 2) any duplicate submissions by the same student, 3) incomplete responses, and/or 4) responses that did not correctly respond to the 'check items'. One check item asked students to select a specific response (i.e., *somewhat agree*). Students who did not respond to this check item correctly were assumed to have responded to the survey randomly without reading the items, therefore their responses were removed. Additionally, a topic-based check item was included that asked students to select the topic covered during the day of the activity. Students who responded with the incorrect class topic were assumed to have not attended class and were also removed from the data set. As some items contained statements about interactions with others, surveys also included an item asking students if they worked with a group or discussed with other students during class that day. Only students who selected that they worked or discussed with other students were included in the final data set. Overall, 183 responses were collected for the DD activity day and 215 for the clicker day, which were a 69% and 81% response rate, respectively, based on the week-1 enrollment of the course of 266 students. After data cleaning, there were 149 remaining student responses for the DD activity day and 136 student responses for the clicker day. Item descriptive statistics are provided in Appendix C.

To gather evidence of internal structure validity, the survey data were analyzed using exploratory factor analysis (EFA), as EFA does not require an *a priori* structure to be specified. This allowed for the factor structure of both mASPECT versions to be explored. The number of factors used for the EFAs was selected based on results from both the Kaiser criterion and the scree test (Brown, 2015). These analyses were completed using the 'stats' package (version 3.6.2) and EFAs were completed with the 'psych' package (version 1.9.12.31) in R. All EFAs used principle axis factoring with a promax (oblique) rotation, as an oblique rotation method allows for correlation between the factors. Negatively worded items were reverse coded before EFAs were completed.

The data were analyzed using an iterative process consisting of an EFA, removal of items that did not meet certain criteria, and then a subsequent EFA with the remaining items (Hancock et al., 2010). Items were removed at each step if they had factor loadings of less than 0.4, cross-loaded on two or more factors, or loaded on factors that contained less than three items. For exploratory purposes, items with cross-loadings between 0.3 and 0.4 were flagged but not immediately removed. This process was repeated until all remaining items met the criteria and produced well-formed factors. All items included in final EFAs had loadings of less than 0.35 on the non-primary factors.

Reliability evidence of the data collected was evaluated using the final factor structure found through EFA. As EFA allows all items to load onto each factor, individual factor models were not evaluated, therefore it is unknown if the final factor models contained equal item loadings (i.e., a tau-equivalent model). Thus, the decision was made to estimate the single-administration reliability using omega instead of alpha for each individual factor, as the criteria for omega does not require equal item loadings or errors (Komperda, Pentecost, et al., 2018). Although there are no formal cutoffs for good single-administration reliability, values above 0.7 are generally considered acceptable.

Student Interviews

Because modifications were made to the original items and new items were also included in the mASPECT surveys, evidence of response process validity was gathered from students through the use of cognitive interviews (Willis, 2005). At the end of the associated quantitative surveys, students were given the option to include their email address to indicate they were interested in participating in a short in-person interview about the survey. After the survey closed, emails were sent to randomly selected students, and interviews were scheduled. Response process interview data were gathered for both types of active learning environments (i.e., DD activity day and clicker day) separately. Four students participated in on-campus interviews about the survey items related to the DD activity day (mASPECT-DD), and eight students about the items related to the clicker day (mASPECT-C). Each student was interviewed individually and all interviews were audio recorded. During each interview, students were directed to read each item aloud, state which response they selected, and then explain their reasoning for choosing that response. When needed, students were asked follow-up questions to gain more details about their understanding of the items themselves and/or their response reasoning. All students who participated in an interview were compensated with a $20 gift card.

The audio recording of each interview was initially analyzed by two researchers individually. Student responses to each item were recorded as either being in alignment with the intention of the item or were flagged for possible confusion or irrelevance to the active learning environment. The two researchers then came together and discussed the responses for each item and came to a consensus on which items seemed unclear to the students or were not relevant to the type of active learning environment. The student responses to these items were then provided to a third researcher, who similarly analyzed the items for clarity and relevance. The items that all three researchers agreed were unclear or irrelevant to the type of environment based on the student interviews were removed prior to quantitative analysis and provided insights for items that were not found to contribute to the final factor structure.

Results

Evaluation of mASPECT-DD data

Through the interview results (n = 4), Item 10, “I am confident in my understanding of the material presented during today’s class” was found to be irrelevant to this type of activity. When students were asked to explain their response to this item, they would refer to the out-of-class assignment of finding articles to bring in instead of their confidence in what was learned during the activity itself. Additionally, Item 22, “I guessed or made stuff up so that I could finish today’s activity” and Item 23, “I skipped or guessed on the hard parts of today’s activity” were also found to be irrelevant to the students based on the structure of the DD activity, which required students to work together towards finding a solution to a ‘real-world’ problem which was intentionally nuanced with no ‘right answers’. Thus, students said that there was no reason to guess and that there were no ‘hard parts’ to the activity. Two more items were also removed based on student interviews. Item 25, “I tried to relate today’s class to prior material from the course” was removed as students were unable to consistently justify their response, and Item 30, “The students I worked with had fun during today’s class” was removed as students indicated they were unsure how to gauge how much fun other students had. These five items (Items 10, 22, 23, 25, and 30) were removed before quantitative analysis through EFA.

An iterative EFA process was used to determine which items created well-formed factors. A summary of the entire process, including the items that were removed at each step, is displayed in Figure 6.2 (details provided in Appendix C). The final EFA for the mASPECT-DD version consisted of 17 items, which were found to load onto three factors related to ‘personal effort’ (PE, 6 items), ‘value of environment’ (VE, 5 items), and ‘classroom support (instructors and LAs)’ (CS, 6 items) (Table 6.3). The descriptions given to these factors were based on their relation to the original ASPECT factors (Wiggins et al., 2017) and observed similarities of the included items. These three factors were found to explain 18% (PE), 22% (VE), and 16% (CS) of the variance in responses, for a total of 56%. The single-administration reliability coefficient, omega, was calculated for each of the three factors and found to be 0.85 (PE), 0.84 (VE), and 0.90 (CS), which suggested good reliability for each.

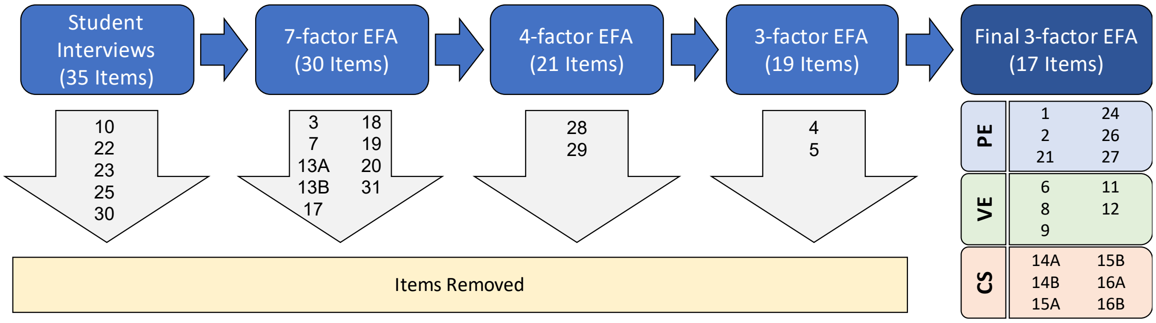


Figure 6.2. Summary of the analysis process for the mASPECT-DD survey. Final factors were ‘personal effort’ (PE), ‘value of environment’ (VE), and ‘classroom support’ (CS).

Table 6.3. Factor loadings for the final 3-factor EFA structure for the mASPECT-DD survey given during a DD activity (n = 149). Item loadings above 0.4 are bolded. Items that were reverse-coded are marked with (*rev*).

Survey Item		Personal Effort	Value of Environment	Classroom Support
1	I was focused during today's class.	**0.69**	0.14	0.07
2	I worked hard during today's class.	**0.56**	0.05	0.06
21	I did not make much of an effort during today's class. (*rev*)	**0.97**	0.32	0.03
24	I found it difficult to maintain my concentration during today's class. (*rev*)	**0.52**	0.20	-0.09
25	I was not very engaged in today's class. (*rev*)	**0.71**	0.02	-0.02
26	I was fully engaged in today's class.	**0.67**	0.15	-0.02
6	Working with other students during today's class contributed to my understanding of the material.	0.21	**0.42**	0.13
8	I had fun during today's class.	0.04	**0.70**	0.06
9	I would prefer to take a class that included today's activity over one that does not include it.	-0.14	**0.97**	-0.17
11	Today's class increased my understanding of the material.	-0.01	**0.69**	-0.03
12	Today's class stimulated my interest in the course material.	0.05	**0.69**	0.08
14A	The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.	-0.17	0.20	**0.73**
14B	The Learning Assistant put a good deal of effort into my learning for today's class.	0.09	0.07	**0.68**
15A	The Professor/Teaching Assistant seemed prepared for today's class.	0.11	-0.07	**0.73**
15B	The Learning Assistant seemed prepared for today's class.	0.06	-0.14	**0.88**
16A	The Professor/Teaching Assistant was available to answer questions during today's class.	-0.08	0.02	**0.75**
16B	The Learning Assistant was available to answer questions during today's class.	-0.04	-0.13	**0.91**

Average scale scores were calculated using the final mASPECT-DD factor structure (Table 6.4). Since EFAs allow items to load on all factors, weighted means could not be calculated and as such, the values presented assume each item contributed equally to the factor.

Table 6.4. Average scale scores for mASPECT-DD factors (n = 149). All item responses were collected on a six-point Likert-type scale from 1 (*strongly disagree*) to 6 (*strongly agree*).

Factor	**Average Scale Score (standard deviation)**
Personal Effort (Items 1 – 2, 21, 24 – 26)	4.64 (0.79)
Value of Environment (Items 6, 8 – 9, 11 – 12)	4.26 (0.92)
Classroom Support (Items 14A – 16B)	5.15 (0.71)

Evaluation of mASPECT-C data

Data collected with the items administered during the clicker day were also analyzed using student interviews and EFAs. Response process interviews (n = 8) about the mASPECT-C items led to the removal of three items. Item 18, "I knew what I was expected to accomplish during today's class" and Item 20, "I completed the pre-work for today's class" were removed as students mentioned that these items did not relate to clicker days since their only expectation during class was to understand the material and there was no required "pre-work" to complete before attending class that day. Additionally, Item 30, "The students I had discussions with had fun during today's class" was removed as students indicated they were unsure of how to respond to this statement.

The remaining items were quantitatively analyzed with an iterative EFA process. A summary of the entire process, including the items that were removed at each step, is displayed in Figure 6.3 (details provided in Appendix C). The final EFA for the clicker day mASPECT-C items was found to contain 21 items with four factors related to 'personal effort' (PE, 5 items), 'social influence' (SI, 8 items), 'value of environment' (VE, 4 items), and 'classroom support (instructors only)' (CS, 4 items) (Table 6.5). The three factors of 'personal effort', 'value of environment', and 'classroom support' were similar to the factors found with mASPECT-DD and thus were named accordingly. The fourth factor was named ‘social influence’, as the included items appeared to be related to working with other students. The four factors were found to explain 17% (PE), 19% (SI), 13% (VE), and 7% (CS) of the variance, for a total of 55%. Omega was calculated for each of the four factors and found to be 0.85 (PE), 0.89 (SI), 0.81 (VE), and 0.81 (CS), which indicated good single-administration reliability.

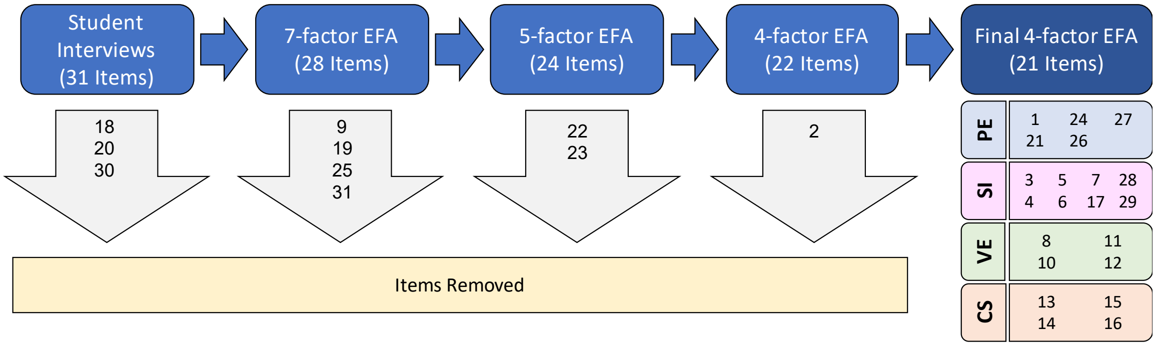


Figure 6.3. Summary of the analysis process for the mASPECT-C survey. Final factors were ‘personal effort’ (PE), ‘social influence’ (SI), ‘value of environment’ (VE), and ‘classroom support’ (CS).

Table 6.5. Factor loadings for the final EFA structure for the mASPECT-C (n = 136). Item loadings above 0.4 are bolded. Items that were reverse-coded are marked with (*rev*).

Survey Item		Personal Effort	Social Influence	Value of Environment	Classroom Support
1	I was focused during today's class.	**0.61**	0.04	0.19	0.14
21	I did not make much of an effort during today's class. (*rev*)	**0.67**	0.14	-0.24	0.10
24	I found it difficult to maintain my concentration during today's class. (*rev*)	**0.74**	-0.13	0.32	-0.20
26	I was not very engaged in today's class. (*rev*)	**0.78**	-0.03	0.06	-0.06
27	I was fully engaged in today's class.	**0.52**	0.13	0.25	0.04
3	I made valuable contributions when having discussions with other students during today's class.	-0.09	**0.63**	0.23	-0.08
4	Explaining the material to other students improved my understanding of it.	-0.05	**0.52**	0.15	0.12
5	Having the material explained to me by other students improved my understanding of it.	0.04	**0.76**	-0.27	0.10
6	Discussion with other students during today's class contributed to my understanding of the material.	0.09	**0.80**	-0.15	0.04
7	The students I had discussions with made valuable contributions during today's class.	-0.01	**0.94**	-0.08	-0.15
17	I felt comfortable having discussions with other students during today's class.	-0.08	**0.62**	0.27	-0.16
28	The students I had discussions with were focused during today's class.	0.14	**0.54**	0.14	-0.01
29	The students I had discussions with worked hard during today's class.	0.05	**0.75**	-0.15	0.08
8	I had fun during today's class.	0.03	0.18	**0.58**	0.04
10	I am confident in my understanding of the material presented during today's class.	0.04	-0.11	**0.73**	-0.09
11	Today's class increased my understanding of the material.	0.05	-0.02	**0.65**	0.15
12	Today's class stimulated my interest in the course material.	0.10	-0.13	**0.66**	0.22
13	The Professor/Teaching Assistant's enthusiasm made me more interested in today's class.	-0.12	0.09	0.28	**0.60**
14	The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.	-0.15	0.01	0.21	**0.61**
15	The Professor/Teaching Assistant seemed prepared for today's class.	0.12	-0.17	-0.11	**0.86**
16	The Professor/Teaching Assistant was available to answer questions during today's class.	0	0.08	-0.03	**0.66**

Average scale scores were calculated for the mASPECT-C version using the final factor structure (Table 6.6). The values presented assume each item contributed equally to the factor, as EFAs allow all items to load on each factor.

Table 6.6. Average scale scores for mASPECT-C factors (n = 136). All item responses were collected on a six-point Likert-type scale from 1 (*strongly disagree*) to 6 (*strongly agree*).

Factor	**Average Scale Score (standard deviation)**
Personal Effort (Items 1, 21, 24, 26 – 27)	4.64 (0.87)
Social Influence (Items 3 – 7, 17, 28 – 29)	4.94 (0.70)
Value of Environment (Items 8, 10 – 12)	4.75 (0.74)
Classroom Support (Items 13 – 16)	5.34 (0.61)

Discussion

Interview and EFA results provided evidence of response process and structural validity for the data collected with both mASPECT versions and resulted in well-formed factor structures.

Comparisons among the factor structures of mASPECT-DD and mASPECT-C

Although similarly worded items were used in both mASPECT versions, different factor structures were discovered for the two environments; a 3-factor solution was found to describe the DD activity day (mASPECT-DD) data, while a 4-factor solution described the clicker day (mASPECT-C) data (Table 6.7). The data from both active learning environments included factors related to ‘personal effort’, ‘value of environment’, and ‘classroom support’, however, these factors included different items for the different environments. Thus, although they could be considered *similar* constructs, they were not found to be identical. Additionally, a fourth factor related to ‘social influence’ was discovered for data collected in the clicker day environment with mASPECT-C. This factor was not found for data collected for the DD activity (mASPECT-DD), nor was it an original ASPECT factor (see Appendix C). This result suggests that students' perceptions of the clicker day environment included a social component, which may not have been an important factor in the DD activity environment. However, as open-ended student interviews asking about their general perceptions of the active learning environments were not conducted during this study, we cannot say that students did not find social influence to contribute to their perceptions of the DD activity, just that none of the included items were found to measure this perception.

Table 6.7. Comparison of the final factor structures found for mASPECT-DD and mASPECT-C.

Factors	mASPECT-DD	Item Number	mASPECT-C	Factors
Personal Effort	I was focused during today's class.	1	I was focused during today's class.	Personal Effort
	I worked hard during today's class.	2	*Removed*	
	I did not make much of an effort during today's class	21	I did not make much of an effort during today's class	
	I found it difficult to maintain my concentration during today's class.	24	I found it difficult to maintain my concentration during today's class.	
	I was not very engaged in today's class.	26	I was not very engaged in today's class.	
	I was fully engaged in today's class.	27	I was fully engaged in today's class.	
	Removed	3	I made valuable contributions when having discussions with other students during today's class.	Social Influence
Value of Environment	*Removed*	4	Explaining the material to other students improved my understanding of it.	
	Removed	5	Having the material explained to me by other students improved my understanding of it.	
	Working with other students during today's class contributed to my understanding of the material.	6	Discussion with other students during today's class contributed to my understanding of the material.	
	Removed	7	The students I had discussions with made valuable contributions during today's class.	
	Removed	17	I felt comfortable having discussions with other students during today's class.	
	Removed	28	The students I had discussions with were focused during today's class.	
	Removed	29	The students I had discussions with worked hard during today's class.	

Table 6.7 cont.

Factors	mASPECT-DD	Item Number	mASPECT-C	Factors
Value of Environment (cont.)	I had fun during today's class.	8	I had fun during today's class.	Value of Environment
	I would prefer to take a class that included today's activity over one that does not include it.	9	*Removed*	
	Removed	10	I am confident in my understanding of the material presented during today's class.	
	Today's class increased my understanding of the material.	11	Today's class increased my understanding of the material.	
	Today's class stimulated my interest in the course material.	12	Today's class stimulated my interest in the course material.	
Classroom Support (Instructors and LA)	*Removed*	13A	The Professor/Teaching Assistant's enthusiasm made me more interested in today's class.	Classroom Support (Instructors only)
	Removed	13B	*not applicable*	
	The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.	14A	The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.	
	The Learning Assistant put a good deal of effort into my learning for today's class.	14B	*not applicable*	
	The Professor/Teaching Assistant seemed prepared for today's class.	15A	The Professor/Teaching Assistant seemed prepared for today's class.	
	The Learning Assistant seemed prepared for today's class.	15B	*not applicable*	
	The Professor/Teaching Assistant was available to answer questions during today's class.	16A	The Professor/Teaching Assistant was available to answer questions during today's class.	
	The Learning Assistant was available to answer questions during today's class.	16B	*not applicable*	

Student perceptions of the environments

Although the factor names for 'personal effort', 'value of environment', and 'classroom support' for mASPECT-DD and mASPECT-C are identical, since the factors contain different items, the final scale scores cannot be compared to each other. However, independently considering the scale scores from each environment can still provide insight into how students viewed the environments. For example, based on the average scale scores it appeared that students positively recognized the classroom support that was present during both the DD activity (Table 6.4) and the clicker day (Table 6.6). They also perceived their personal effort and the value of the environment to be fairly high for both types of environments, as all of the averaged scale scores were above 4 (i.e., *somewhat agree*). Within the clicker day environment, it appeared that students also perceived the social influence positively. These results suggest that the students thought fairly highly of both the DD activity and the clicker day learning environments, as measured by these factors.

Limitations

The relatively low survey response rates (~50%) were a limitation of this study. However, these percentages only represent the students who consented to be part of the study and do not include the students who accessed the surveys for extra-credit only. Overall, 69 – 81% of enrolled students accessed the surveys; however, as these surveys were given in the course as part of a research study, students could not be required to complete it. Additionally, student interviews only captured the perceptions of a small subset of the classroom population who self-selected to participate.

While the scale scores indicate that students generally perceived both environments positively, these results should be interpreted cautiously. Even with the well-formed factor structure found for both surveys, the amount of variance explained by each factor only ranged from 7 – 22%. This indicates that there could have been additional factors that contributed to students' perceptions of the environment that were not measured with this survey. Additionally, although the general descriptions given to the factors aligned with the original ASPECT factor descriptions and appear to describe the items that were contributed to each factor, neither the original study (Wiggins et al., 2017) nor this study evaluated test content validity (Arjoon et al., 2013; Knekta et al., 2019) in relation to theoretical definitions of the different constructs. As such, these factors cannot be said to measure theoretically defined constructs of personal effort, value of environment, classroom support, or social influence.

Implications for Research

Collecting data with the mASPECT may provide insight into students' perceptions of in-class active learning environments, which could be an important contributor to the variation in student performance outcomes found in these environments. There are a number of opportunities for comparisons of students' perceptions of personal effort, value of environment, classroom support, and social influence and how those might change based on the type of environment. However, as evidenced by the differences in factor structures between mASPECT-DD and mASPECT-C, these measures should not be used to directly compare results from different active learning environments unless evidence of validity has been gathered in each environment for data collected *with the same version* of the survey. Therefore, we encourage users of the mASPECT or ASPECT to continue to collect evidence of response process validity to ensure that the items on either measure make sense to students and are relevant for a given type of active learning environment. Although this could take the form of student interviews, a larger number of student response process data could alternatively be collected through the use of open-ended written survey responses. As active learning strategies can take many forms, the use of response process data could be used to determine what students find important in different types of active learning environments and ensure that these or related items are worded properly to appropriately capture those perceptions. Additionally, as Wiggins et al. (2017) noted, an important potential use of the data collected by these scales is to better understand if there are equitable outcomes and experiences across student and/or demographic groups in the same classroom. However, evidence of measurement invariance between different groups would first have to be evaluated (Rocabado et al., 2020).

Finally, although the mASPECT versions provide information on students' perceptions of these active learning activities, the measures were not developed to directly align with theoretical definitions of student engagement. The ASPECT was developed as a measure of students' *perceived* cognitive and emotional engagement during in-class active learning activities, however, the original authors note that the psychometric properties of the ASPECT were not evaluated with respect to the theoretical definition of engagement (Wiggins et al., 2017). To assess the extent to which the ASPECT or mASPECT measures are a representation of engagement, evidence of test content validity that is aligned with a theoretical definition of engagement would have to be gathered and evaluated (Arjoon et al., 2013; Knekta et al., 2019).

Alternatively, future studies could administer both a measure of engagement and mASPECT or ASPECT to evaluate the overlap between constructs.

Implications for Teaching

Instructors who want to learn more about how students' perceptions differ across active learning environments could use the mASPECT measure to gather feedback about different active learning strategies. For example, the mASPECT could be used to gather pre- and post-data that could be used to inform the instructor if group-level dynamics improved after a certain strategy was implemented or adapted. As evidenced by the differences in factor structures between mASPECT-DD and mASPECT-C, the scale scores (i.e., item averages within a scale) from these measures should not be used to directly compare results from different active learning environments unless evidence of validity has been gathered in each environment for data collected *with the same version* of the survey. However, even if scale scores cannot be compared, instructors may still wish to use one or more of the individual mASPECT survey items as formative feedback for environments that are similar to the ones described in this study. For example, if an instructor implements a group-work focused activity similar to DD or includes clicker questions in their course, they could collect feedback about students' perceptions using common items from the mASPECT-DD and mASPECT-C, which could be used to inform changes or modifications to the environment or facilitation of the activity.

Although the mASPECT versions provide information on students' perceptions of these active learning activities, the measures were based on the original ASPECT items, thus do not directly align with theoretical definitions of student engagement (Wiggins et al., 2017). Therefore, if an instructor's goal is to measure student *engagement* in the classroom, other measures may be better suited. For example, some observational protocols have been developed to evaluate student engagement during class such as the Behavioral Engagement Related to Instruction (BERI) (Lane & Harris, 2015) and the ICAP framework (Chi & Wylie, 2014). Additionally, some survey measures have been developed to assess different dimensions of student engagement in higher education STEM classrooms (Skinner et al., 2017) and laboratories (Smith & Alonso, 2020).

Supplemental Materials

Appendix C: Initial analysis of the original ASPECT in our environments, survey modification, item descriptive statistics for mASPECT-DD and mASPECT-C, iterative EFA process details, final survey structure comparison (mASPECT vs. ASPECT)

Acknowledgements

We thank the students who participated in the surveys and instructors for allowing data collection in their courses. We thank Eric Earle for assistance with the interview coding. Author E.E.S acknowledges partial-support by an award to PSU under the Howard Hughes Medical Institute Science Education Program (award 52008105). The authors declare no conflicts of interest.

Chapter 7: Exploring Student Perceptions of Behavioral, Cognitive, and Emotional Engagement at the Activity Level in General Chemistry

Reprinted with permission from Naibert, N., Vaughan, E. B., Brevick, K., & Barbera, J. (2022). Exploring Student Perceptions of Behavioral, Cognitive, and Emotional Engagement at the Activity Level in General Chemistry. *Journal of Chemical Education, 99*(3), 1358-1367. Copyright 2022 American Chemical Society.

Abstract

Although active learning strategies are being incorporated into many higher-education STEM courses, not all students benefit from these activities to the same extent. As these types of activities are designed to engage students in their learning, differences in student engagement may explain some of the differences in learning outcomes. However, before student engagement in active learning activities can be meaningfully measured using a self-report survey, it is important to evaluate if students perceive engagement similarly to the literature definitions these measures are based on. Therefore, this study sought to explore students' perceptions of the behavioral, cognitive, and emotional dimensions of engagement with respect to specific worksheet activities incorporated into a general chemistry course. This was completed through the use of open-ended written responses and interviews. Results indicated that although students generally perceived behavioral, cognitive, and emotional engagement similarly to the literature definitions of these dimensions, students tended to conflate many ideas of behavioral and cognitive engagement. Additionally, social themes were also discovered to be threaded throughout student responses to the three dimensions of engagement, suggesting students also perceived the presence of a social engagement dimension when considering engagement at the activity level.

Graphical Abstract

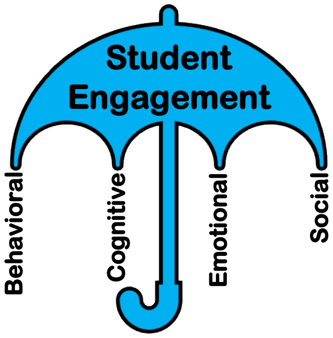


Figure 7.1. Graphical abstract for Chapter 7.

Keywords

First-Year Undergraduate / General, Chemical Education Research, Collaborative / Cooperative Learning, Student-Centered Learning

Introduction

Incorporating active learning strategies in the classroom has been found to generally improve student performance outcomes with respect to exam scores, course grades, withdrawal rate, and other measures when compared to traditional lecture classes (Freeman et al., 2014; Rahman & Lewis, 2019). However, the benefits of active learning may not be realized to the same extent for every student. Case-studies of individual students found that university students' experiences in the same active learning environment varied and were not necessarily reflected in their course grades (Wiltbank et al., 2019). Additionally, another study found that students' grades were not dependent on their attitude toward the active learning environment (Shortlidge et al., 2019). In general, higher student engagement in an environment has been shown to be positively related to student learning outcomes (Chi & Wylie, 2014; Skinner et al., 2017). As active learning environments focus on engaging students in their learning through the use of discussion and activities (Freeman et al., 2014), variations in how students engage in these tasks may influence if and how they benefit from them.

Student engagement in the classroom has been conceptualized through two different but related perspectives. The “behavioral perspective” of student engagement focuses solely on the behavioral dimension of engagement, such as time and effort, and the relation of certain behaviors to students’ achievement (Kahu, 2013). However, this perspective may underrepresent equally important aspects of engagement related to students’ psychological state, such as the students’ investment in their learning and emotions (Kahu, 2013). Therefore, the “psychological perspective” of student engagement encompasses several dimensions of engagement, including ones related to behavioral, cognitive, and emotional aspects (Kahu, 2013). Frameworks based in the psychological perspective can be single or multidimensional in nature. For example, the ICAP (interactive-constructive-active-passive) framework (Chi & Wylie, 2014) focuses solely on categorizing different modes (i.e., levels) of students’ cognitive engagement. Other frameworks consist of multiple overlapping dimensions. One such framework, defined by Fredricks et al. (2004), characterizes engagement as including interrelated behavioral, cognitive, and emotional components and emphasizes that these dimensions should be evaluated simultaneously to better assess the complex construct of engagement and account for any effects due to the overlapping nature of the dimensions. Therefore, when evaluating student engagement, a multidimensional engagement framework can provide a more complete perspective of student engagement than simply focusing on a single component.

Within this multidimensional framework of engagement, behavioral engagement focuses on students' positive conduct and involvement in the classroom, which can include behaviors related to asking questions, putting in effort to do the work, and paying attention (Fredricks et al., 2004; Sinatra et al., 2015). Cognitive engagement is often considered in relation to students' psychological investment in their learning, which includes putting in effort to understand and master the material, as well as going above and beyond the requirements (Fredricks et al., 2004; Sinatra et al., 2015). Emotional engagement centers around students' affective reactions to interactions they have in the classroom. Many different types of emotions are often included when considering emotional engagement, such as, interest, boredom, value, etc., and can be related to students' interactions with peers, instructors, the course material, or in-class activities (Fredricks et al., 2004; Sinatra et al., 2015).

Observational measures can be used to evaluate student engagement in the classroom (e.g., Chi and Wylie (2014), Harris and Cox (2003), Lane and Harris (2015)); however, they are often difficult to implement in large enrollment courses and generally are only used to assess the engagement of a subset of students. Additionally, observational measures are generally cautioned against when evaluating cognitive and emotional engagement, as indicators of these dimensions tend to be internal to the students (Appleton et al., 2006). Instead, self-report surveys can be used to collect information from all students in a class, as well as allow the simultaneous evaluation of behavioral, cognitive, and emotional engagement. Multiple self-report survey measures have been used to collect data on student engagement in higher-education STEM courses (e.g., Aceti (2017), Gasiewski et al. (2012), Seery (2015a), Skinner et al. (2017), Smith and Alonso (2020)). Measures focused on specific dimensions of engagement often rely on the literature definitions, expert feedback, and factor analysis to create initial items and to group the items into the different dimensions (e.g., Skinner et al. (2017), Smith and Alonso (2020)). While these methods provide evidence of item groupings and alignment of items to the literature definitions through the perceptions and interpretations of experts, it is unknown whether students' perceptions of the dimensions of engagement align with their definitions in these environments. Exploring students' perceptions of engagement would provide additional evidence for the validity of data collected with engagement measures. Therefore, the main goal of this study was to assess students' perceptions of behavioral, cognitive, and emotional engagement with relation to active learning worksheet activities incorporated into a general chemistry lecture course and to evaluate the alignment of student perceptions with the literature definitions of these dimensions. This goal is met by answering the research question: How do students perceive behavioral, cognitive, and emotional engagement in these general chemistry worksheet activities?

Course Information

All students included in this study were part of the general chemistry (GC) course sequence at Portland State University (PSU). Student data from two different academic years and terms (GC I and GC II) were collected (Table 7.1). During both years the course included both lecture and activity days. The activity days focused around completing worksheet activities that were similar in structure to Process-Oriented Guided Inquiry Learning (POGIL) materials. Because of the Covid-19 pandemic, this course was transitioned from meeting in-person during the Winter of 2020 to meeting remotely (i.e., online) by Fall 2020. Due to the change in classroom environment, the implementation of the activity days differed between the two years.

Table 7.1. Course information for each general chemistry (GC) classroom environment.

Classroom environment	Term	Sections included in study	Week 1 enrollment
In-person	Winter 2020 (GC II)	1	N = 249
Remote	Fall 2020 (GC I)	2	N = 629

In the in-person environment, students were expected but not required to work on the worksheets in groups of 2 – 4 students. In this environment, the instructor, graduate teaching assistant (TA), and multiple undergraduate learning assistants (LAs) all moved throughout the room to facilitate group discussion and answer questions. Clicker questions were used periodically to gauge students' understanding of the content. Data were collected from a single section of the course, which was taught by an instructor who also conducted one of the remote instruction sections the following year.

In the remote environment, both sections were facilitated using the video platform software Zoom. During activity days, students were directed to work through the worksheets in randomly assigned groups in breakout rooms, although some students opted to stay in the main room to work on the worksheet by themselves. Students who did work with others in a breakout room generally worked with 2 – 6 other students. Multiple undergraduate and graduate LAs each rotated through an assigned set of breakout rooms to facilitate discussion and answer questions. The main role of the instructor and graduate TA during the activity days was to manage the remote breakout rooms through check-ins with the LAs. Although two instructors taught this course, both implemented the worksheets similarly and students from both sections are represented in the data collected.

Methods

Data Collection

Institutional Review Board (IRB) approval from Portland State University was received for all data collected within this study and appropriate consent was obtained from students as required by the IRB.

Student responses from the in-person environment were collected through the use of open-ended survey items. Students were notified about the survey two weeks before the end of the course through an in-class announcement, as well as one posted on the course's learning management site. The posted announcement included a link to the Qualtrics survey. Students were given one week to access and complete the survey and were offered a nominal amount of extra credit for accessing the survey regardless of completion. Student responses from the remote environment were collected through the use of interviews. Students who were interested in participating in an interview were asked to provide their email at the end of a related survey. All interested students were contacted and directed to fill out a consent form and provide their availability. These responses were used to schedule group interview sessions with at least two students per time slot. Each interview only contained students from the same class section and all students who participated in the interviews had worked on the worksheet in breakout rooms with other students for at least some of the activities. A total of 14 students from both sections participated in 8 interviews. Although the goal was to have at least two students to create a group interview (i.e., focus group) environment, to respect participants time, an interview proceeded even if only one student showed up to the agreed upon time. Therefore, 3 interviews were conducted with individual students, 4 interviews included two students, and one interview included three students. All interviews were completed over Zoom with audio and visual recording and recordings were transcribed prior to analysis.

Open-ended Survey Items

As part of a larger survey, each student was randomly given a definition of engagement that aligned with either behavioral, cognitive, or emotional engagement. These definitions were based on the descriptions of the three engagement dimensions given by Fredricks et al. (2004) They were initially created by the primary researcher (author N.N.) and then were slightly modified for meaning and clarity with the input of a secondary researcher. These definitions are presented in Table 7.2. Based on the presented definition, students were asked to respond to the following two items:

1) How would you describe a student who is **NOT engaged** in the worksheet activities based on the definition above?
2) How would you describe a student who is **VERY engaged** in the worksheet activities based on the definition above?

Table 7.2. Definitions of the three dimensions of engagement presented to students.

Engagement Dimension	**Definition presented to students**
Behavioral	Engagement is the physical participation or involvement in the worksheet activities.
Cognitive	Engagement is [exerting mental effort][a] to comprehend ideas or skills presented in the worksheet activities.
Emotional	Engagement is the positive feelings towards the worksheet activities.

[a]The phrase “exerting mental effort” was replaced with “trying” for the version presented to students during the interviews.

Interview Protocol

Interviews were completed remotely over Zoom using a semi-structured interview approach. The students were first asked to generally describe their engagement in the worksheet activities. They were then presented with definitions of engagement related to behavioral, cognitive, and emotional engagement (see Table 7.2) one at a time using the chat function in Zoom. These definitions were similar to the definitions given during the open-ended survey; however, the cognitive definition was slightly modified to remove the word "effort", as it has been associated with both behavioral and cognitive engagement components (Fredricks et al., 2004). One student was asked to read each definition out loud and then all students were asked to describe what it meant to be *very engaged* and *not engaged* in the worksheet activities based on the definition provided. Follow-up questions were asked as needed for further clarification.

Data Analysis

Data collected from the open-ended surveys were coded using an inductive approach where codes were created from the data. Before coding, the responses were first cleaned to remove any illogical responses from the dataset (i.e., responses that only included random characters such as "n", ".", etc.). To create the initial codebook, the two coders individually separated each response to the two items (i.e., *not* engaged and *very* engaged) into statements that each coder felt represented different topics related to any dimension of engagement defined by Fredricks et al. (2004). These statements were then organized by each coder into groupings that related to a possible code. They then came together to discuss these possible codes and through consensus compiled the initial codebook (version 1). The coders used the initial codebook to individually code the full set of responses to the two items related to behavioral engagement (n = 55 each) and then met to discuss any discrepancies or possible new codes that appeared. Based on the discussion, codes were modified or added as needed and the codebook was updated to create a second version. The set of responses for the cognitive (n = 57 for each item) and emotional (n = 58 for each item) definitions were then coded in a similar manner and subsequent discussion between the two coders resulted in some modifications and updates to the codebook to create a third and final version. This final version of the codebook was then used by both coders to recode approximately 20% of the responses to the items for each of the definitions. Inter-coder reliability (ICR) (O'Connor & Joffe, 2020) was calculated to evaluate agreement between the coders through the use of Cohens' kappa (Cohen, 1960), which accounts for agreement due to chance. Cohen's kappa was calculated using the *irr* package (version 0.84.1) in the statistical software R (version 3.6.2). The ICR for the responses coded by both coders was 0.87, which is considered substantial agreement (Landis & Koch, 1977). The remainder of the responses were then recoded by the primary researcher using the final version of the codebook. The percentage of student responses that mentioned each code was calculated out of the total number of responses for each definition of engagement.

Responses from the interviews were coded using both an inductive and deductive approach, such that codes were created from a subset of the interview data and then applied to the remaining interviews. The two coders started by individually reading through two transcripts and highlighting phrases and statements from the students that corresponded to their engagement in the activities. The coders then came together to discuss any discrepancies in the highlighted sections to reach a consensus on which statements did or did not relate to engagement. Agreed upon phrases were then grouped together into categories to create the initial codebook (version 1) and similar codes to the in-person environment were used when possible. Version 1 of the codebook was then used by the coders to separately code two additional transcripts. The coders came together to discuss any discrepancies and modified the codebook as needed, producing the second and final version. This final codebook was then used to code the remainder of the interview transcripts and to recode the first two transcripts. During this process, both coders would separately code two transcripts at a time and then come together to discuss any discrepancies. No new codes were discovered during this process and each transcript was coded to consensus. ICR was evaluated throughout the process through the use of Cohen's kappa and found to be between 0.43 – 0.63 among the transcripts, which is considered moderate agreement (Landis & Koch, 1977). Although ICR was lower when coding the interviews compared to the short-answer survey responses, ICR is generally expected to be lower when coding something with a larger amount of text (O'Connor & Joffe, 2020). The number of students who mentioned each code at least once throughout the three engagement definitions was calculated to determine overall how many students perceived engagement related to that code. Additionally, the number of students who mentioned each code during the section of the interview where they were provided one of the engagement definitions (i.e., behavioral, cognitive, emotional) was also calculated. If a student mentioned the same code in relation to multiple definitions, they were counted as having mentioned that code during each definition section. Therefore, the number of students who mentioned the code *overall* throughout the interview may be lower than the sum of the number of students who mentioned a code for each of the definitions.

For both sets of data (i.e., open-ended and interview responses), each code that was discovered was organized into either behavioral, cognitive, or emotional engagement. Codes were first categorized by the primary researcher based on if they were related to positive engagement or negative engagement (i.e., disengagement). The primary researcher then further categorized the codes into the different dimensions of engagement based on the theoretical definitions given by Fredricks et al. (2004). Behavioral codes included ones that were related to participation and involvement, as well as behaviors involving staying focused, paying attention, and asking questions. Cognitive codes were selected based on if they described students' investment in their learning. This included codes related to putting effort into understanding, applying and/or connecting the activity to prior material or classes, and learning from mistakes. As there is some known overlap between behavioral and cognitive engagement (Fredricks et al., 2004; Sinatra et al., 2015), codes that focused specifically on behaviors related to student participation (i.e., talking with others, asking questions, etc.) were coded as behavioral and codes that focused on going "above and beyond" to understand the material (i.e., discussion, writing extra notes, helping others, etc.), were categorized as cognitive. Codes were grouped into emotional engagement if they were related to students' feelings about the activity, working with others, or the content material in general. A secondary researcher independently reviewed the categorizations and some uncertain codes related to cognitive and behavioral engagement were discussed among the researchers until a consensus was reached.

Results and Discussion

From the in-person environment survey responses, 63 codes were discovered related to how students perceive behavioral, cognitive, and emotional engagement in this environment (details included in Tables D.1 – D.3 in Appendix D). These codes represented 31 positive aspects of engagement and 32 aspects related to negative engagement (i.e., disengagement). From the remote environment interview responses, a total of 58 codes related to how students perceive engagement in this environment were discovered (details included in Tables D.4 – D.6 in Appendix D), which included 33 codes related to positive aspects of engagement and 25 related to disengagement.

Behavioral Engagement

When provided with the behavioral definition of engagement, students' perceptions of what constitutes behavioral engagement were described by 31 codes across environments (Table 7.3). This included 15 codes that were similar between the two environments, and 11 and 5 codes only discovered in the in-person and remote environments, respectively.

Table 7.3. Number of students that mentioned each behavioral code when provided with the behavioral definition of engagement.

	Number of Students (%)	
Behavioral Codes[a]	In-person Environment n = 55	Remote Environment n = 14
Engagement		
Asked questions	21 (38.2)	2 (14.3)
Worked on worksheet	13 (23.6)	7 (50.0)
Wrote things down	--	11 (78.6)
Focused/paid attention	8 (14.5)	5 (35.7)
Was prepared	8 (14.5)	4 (28.6)
Tried to do worksheet	8 (14.5)	2 (14.3)
Completed worksheet	8 (14.5)	--
Talked to/worked with others (positive)	8 (14.5)	4 (28.6)
Read question to self	--	3 (21.4)
Shared screen	--	3 (21.4)
Participated	5 (9.1)	3 (21.4)
Asked for group feedback	5 (9.1)	2 (14.3)
Engaged with others	3 (5.5)	--
Put in general effort	2 (3.6)	--
Listened to others	--	1 (7.1)
Disengagement		
Didn't work on worksheet	18 (32.7)	3 (21.4)
Just “there”	--	10 (71.4)
Was on a non-class related device	19 (34.5)	--
Worked on other things	8 (14.5)	6 (42.9)
Distracted	6 (10.9)	9 (64.3)
Didn't try to do worksheet	6 (10.9)	5 (35.7)
Participated in off-topic conversations	6 (10.9)	--
Didn’t talk to/work with others	6 (10.9)	4 (28.6)
Didn't ask questions	4 (7.3)	1 (7.1)
Wasn't prepared	3 (5.5)	1 (7.1)
Left class early	3 (5.5)	--
Talked to others (negative)	3 (5.5)	--
Didn't put in general effort	2 (3.6)	--
Didn't participate	2 (3.6)	--
Didn't complete worksheet	2 (3.6)	--
Copied answers from others	2 (3.6)	--

[a]Although some codes are the same between the two environments, there may be slight differences in the type of responses included in each due to the different data collection formats. Details are included in Tables D.1 and D.4 in Appendix D.

One of the more common perceptions related to positive behavioral engagement was working on the worksheet. Students described many different actions related to working on the worksheet, such as writing things down and reading the questions. Additionally, students talked about asking questions, staying focused, paying attention, being prepared for the activity, and participating as indications of behavioral engagement.

Behavioral engagement was also perceived as working with other students. This included coded actions such as asking for feedback and/or assistance on problems and, in the remote environment, taking a leadership role in the group, sharing their screen over Zoom, and listening to others.

When students talked about disengagement, they would generally mention not working on the worksheet at all, doing or working on other things (e.g., other coursework, ALEKS, etc.), being distracted, and not asking questions. In the in-person environment, students also perceived specific actions as being related to behavioral disengagement such as using a phone or laptop (i.e., "device") in class when not used for the activity and participating in off-topic conversations with fellow students. In the remote environment, the most prevalent perception of behavioral disengagement was the idea of simply "being there" in the Zoom meeting without doing anything. In both environments, students also perceived not working or interacting with the other group members as an aspect of behavioral disengagement. However, not all the students perceived working with others to be necessary for their own engagement, although they did mention that it was generally beneficial to work with others. For example, in an interview, one student said, "*I mean, you can be on your own and doing it engaged. Just being in a group helps because you can share answers and point out mistakes.*"

The behavioral engagement definition provided to students during this study emphasized physical participation and involvement. Students' perceptions of what constitutes behavioral engagement based on this definition closely matches behavioral engagement as described by Fredricks et al. (2004), which includes students' positive conduct in the classroom, as well as "effort, persistence, concentration, attention, asking questions, and contributing to class discussion." Many of the students' perceptions of behavioral engagement included these aspects, which can be seen through the codes t*ried to do the worksheet, focused/paid attention, asked questions,* and *talked to/worked with others*. Additionally, students' perceptions of behavioral disengagement included concepts related to negative conduct, such as using a phone or working on other things.

Cognitive Engagement

A total of 19 codes were found to relate to students' perceptions of cognitive engagement and disengagement across both environments (Table 7.4). The in-person and remote environments shared 8 similar codes, with 7 unique codes discovered in the in-person responses and 4 codes in the remote responses.

Overall, students perceived more aspects of cognitive engagement compared to disengagement in both environments. Specifically, when students were asked about cognitive engagement in the worksheet activities, they referred to the idea of trying to understand the material. This idea was often expanded on by students in the interviews by describing techniques they perceived as being related to cognitive engagement, including checking their work and understanding their mistakes. Students also described aspects related to thinking through how to solve the problems, going through problems step-by-step, and trying to connect material with prior course information, in addition to writing down extra notes (i.e., interacting with the worksheet) and using resources to help solve the problems. Students also perceived cognitive engagement in relation to working with other students to better understand the material, including discussing the worksheet with their peers and helping others.

Students' perceptions of cognitive disengagement centered around not contributing to the group (i.e., not discussing with or helping others), as well as not trying to understand the worksheet or material being covered and just writing the answers down on the worksheet without trying to understand how to actually solve the problems. Additionally, some students perceived disengagement as only doing the minimum required for the worksheets or giving up on trying to solve the problems. Overall, more students mentioned ideas related to positive cognitive engagement compared to cognitive disengagement. This suggests that students might have found it more difficult to conceptualize cognitive disengagement compared to cognitive engagement.

Students in this study were presented with a cognitive engagement definition that centered around trying to comprehend the skills and ideas present in the worksheets. Results from the coding indicated that students' perceptions of cognitive engagement align with the definition from Fredricks et al. (2004), which includes concepts related to students putting in mental effort to understand the material, as well as going above and beyond the minimum requirements. For example, many students perceived cognitive engagement as trying to understand the material. Additionally, codes such as *interacted with the worksheet* (e.g., wrote down extra notes, read the worksheet thoroughly, etc.), *used resources* (to help them work through the problems), and *discussed with others*, suggest that students perceived cognitive engagement as doing more than what was required of them.

Table 7.4. Number of students that mentioned each cognitive code when provided with the cognitive definition of engagement.

	Number of Students (%)	
Cognitive Codes[a]	In-person Environment n = 57	Remote Environment n = 14
Engagement		
Tried to understand	15 (26.3)	12 (85.7)
Helped others	16 (28.1)	5 (35.7)
Checked work/answers	--	5 (35.7)
Discussed with others	9 (15.8)	4 (28.6)
Thought about how to solve problems	--	4 (28.6)
Interacted with worksheet	4 (7.0)	2 (14.3)
Put effort into learning	2 (3.5)	--
Tried to solve problems a different way	2 (3.5)	--
Used resources	1 (1.8)	4 (28.6)
Connected or applied material	1 (1.8)	3 (21.4)
Went through problems step-by-step	--	2 (14.3)
Tried their best/didn't give up	1 (1.8)	--
Learnt from and/or corrected mistakes	1 (1.8)	--
Disengagement		
Didn't try to understand	7 (12.3)	--
Just wrote down answers	5 (8.8)	6 (42.9)
Didn't discuss with others	4 (7.0)	1 (7.1)
Only did the minimum required	4 (7.0)	--
Didn’t help others	--	1 (7.1)
Didn't try their best/gave up	3 (5.3)	--

[a] Although some codes are the same between the two environments, there may be slight differences in the type of responses included in each due to the different data collection formats. Details are included in Tables D.2 and D.5 in Appendix D.

Emotional Engagement

Students’ perceptions of emotional engagement were described by a total of 27 emotional engagement codes in the two environments (Table 7.5). Of these codes, 9 were similar between the two environments, 7 were unique to the in-person environment and 11 to the remote environment.

Table 7.5. Number of students that mentioned each emotional code when provided with the emotional definition of engagement.

	Number of Student Responses (%)	
Emotional Codes[a]	In-person Environment n = 58	Remote Environment n = 14
Engagement		
Positive feelings	6 (10.3)	9 (64.3)
Felt confident	--	10 (71.4)
Felt activity was beneficial	8 (13.8)	7 (50.0)
Liked/enjoyed the activity	4 (6.9)	6 (42.9)
Excited about activity	--	4 (28.6)
Wanted to learn	--	3 (21.4)
Liked chemistry/science	--	3 (21.4)
Didn't feel frustrated	1 (1.7)	2 (14.3)
Wanted to/liked working with others	--	3 (21.4)
Interested in content	2 (3.4)	--
Looked forward to activity	2 (3.4)	--
Disengagement		
Negative feelings	6 (10.3)	7 (50.0)
Felt activity wasn't beneficial	8 (13.8)	3 (21.4)
Felt self-doubt	--	6 (42.9)
Felt frustrated	3 (5.2)	5 (35.7)
Felt disconnected	--	4 (28.6)
Didn’t want to learn	--	3 (21.4)
Didn't like/enjoy activity	5 (8.6)	2 (14.3)
Didn't like chemistry/science	1 (1.7)	2 (14.3)
Felt confused or discouraged	7 (12.1)	--
Didn't care about activity	3 (5.2)	--
Not interested in content	3 (5.2)	--
Didn't look forward to activity	1 (1.7)	--
Didn't want to do activity	1 (1.7)	--
Didn't want to/like working with others	--	1 (7.1)
Felt left behind/rushed	--	1 (7.1)
Felt bored	--	1 (7.1)

[a]Although some codes are the same between the two environments, there may be slight differences in the type of responses included in each due to the different data collection formats. Details are included in Tables D.3 and D.6 in Appendix D.

Overall, many students perceived emotional engagement as feeling like the activities were beneficial for their learning. Students also described feelings of confidence, especially when getting problems correct, and generic positive feelings, such as feeling “good”. Additionally, students perceived emotional engagement as wanting to learn, as well as wanting to work with and/or help others in the group. Other positive feelings that students mentioned in relation to their perception of engagement were liking, enjoying, and being interested in the material/content, or chemistry and science in general. In the interviews, one student summed up many of these feelings by saying,

> "*I actually really enjoy learning about chemistry. It can be super interesting at times, it can be really hard at times, but keeping that positive outlook on it – like, I really enjoy this topic – is important to be engaged*."

Students perceived emotional disengagement as being related to negative (i.e., "not good") feelings and self-doubt, as well as not feeling like the activity was beneficial or useful for learning the material, feeling confused and/or discouraged, and not liking or caring about the worksheet or activity. Students also described the idea of feeling disconnected with the material in relation to big picture ideas and how the content fit together.

The idea of feeling frustrated also appeared in both environments, although student responses in the interviews indicated that some students may perceive frustration as disengagement, while others perceive it as positive engagement. For example, when talking about working through some of the problems, one student said that, "*When I actually tried to do it, I was extremely frustrated and gave up multiple times.*" This instance of frustration seems to indicate disengagement, as the student gave up when they felt frustrated. Although not engagement-specific, measures of emotional satisfaction with learning chemistry, which have included frustration as a negative component (Bauer, 2008; Xu & Lewis, 2011), have found that lower emotional satisfaction is related to lower student performance outcomes (Xu & Lewis, 2011). However, the idea that frustration always indicates disengagement was not universal throughout the interviews. In a different session, a student perceived frustration as an indicator of engagement,

> "*I also think that if they're really engaged, sometimes they might not get the right answer, so they might feel frustrated. But that, that isn't a bad thing in my opinion. Because you can feel frustrated, but that could be a good type of frustration because you're, you just encountered a roadblock, but it's not ultimately preventing you from understanding the ideas. So you can feel frustrated, but it doesn't mean you aren't really engaged because you are engaged already by feeling frustrated because you only feel frustrated when you actually do the activity, because if you don't do that activity in the first place, then you can't really experience any feeling of frustration in the first place.*"

Some definitions of emotional engagement include the idea of positive and negative emotions being either activating or deactivating, where an activating emotion would increase engagement and a deactivating emotion would decrease engagement (Pekrun & Linnenbrink-Garcia, 2012; Sinatra et al., 2015). From these interviews, it appears that although frustration may be a negative affective reaction, students may perceive it as deactivating or activating for engagement depending on how they approach frustration in the context of these worksheets. One study that looked at frustration intolerance (i.e., the inability to continue working on an activity based on negative feelings) found that frustration intolerance influenced college students' academic outcomes, such that students who were more willing to feel discomfort or frustration had higher performance outcomes (Wilde, 2012). Therefore, although how students handle frustration may be an important factor to consider for student achievement, frustration itself may not be a good indicator of students' emotional engagement in classroom worksheet activities due to potential differences in how students may perceive it in relation to engagement.

The Fredricks et al. (2004) definition of emotional engagement centers around students' affective reactions and includes emotions such as value, "interest, boredom, happiness, sadness, and anxiety." Overall, students' perceptions of emotional engagement in these worksheet activities included many of these aspects, indicated through codes related to *felt activity was beneficial, interested in content,* and *felt bored*, for example. Additionally, although not specifically listed as examples in Fredricks et al. (2004), students perceived other affective reactions in relation to the worksheet, including *felt confident, excited about activity, felt self-doubt,* and *felt disconnected*.

Conflation of Behavioral and Cognitive Engagement

Codes related to all three dimensions of engagement were identified throughout all short-answer responses and interview transcripts regardless of the definition students were provided. The prevalence of codes across definitions are provided in Tables D.7 – D.12 in Appendix D. Most students only mentioned emotional engagement codes when asked to describe engagement based on the emotional definition. However, more overlap was seen with the behavioral and cognitive codes across definitions in both the in-person and remote environments. When students were provided with the cognitive definition of engagement, they often mentioned ideas related to the behavioral dimension. For example, many students would indicate behavioral aspects related to working with others, reading the question out loud, and asking for help/feedback. Similarly, when the behavioral definition was given, students would indicate ideas related to cognitive engagement codes (i.e., discussing the worksheet with their group, thinking through how to solve the problems, using resources, etc.). This overlap may indicate that students perceived cognitive and behavioral engagement to be very similar constructs when considering engagement at the activity-level.

As the different dimensions all assess engagement, they are inherently interconnected. However, it has been noted that the overlap between behavioral and cognitive engagement may make it difficult to clearly distinguish between these two dimensions. Specifically, an overlap between cognitive and behavioral engagement may be prevalent when cognitive engagement is perceived primarily through the lens of students' investment in their learning (e.g., putting effort into understanding the material, going above and beyond, etc.) (Sinatra et al., 2015). As 'putting in effort' can be perceived as both behavioral engagement (e.g., putting in effort by doing multiple examples) or cognitive engagement (e.g., putting in effort by trying hard to understand mistakes), there may be a lack of distinction between these two constructs that make them difficult to separate (Fredricks et al., 2004; Sinatra et al., 2015). Overlap between the two dimensions is also seen in one of the cognitive engagement frameworks. The ICAP framework (Chi & Wylie, 2014) of engagement includes four levels, or modes, of cognitive engagement with respect to learning activities. Passive engagement is the lowest level followed by active and then constructive, with interactive as the highest mode. In this framework, each mode of cognitive engagement is defined through observing students' overt behaviors (Chi & Wylie, 2014). For example, if a student is listening to a lecture, they could just *passively* listen, *actively* take notes, *constructively* draw a concept map to connect ideas, or *interactively* discuss the material with a small group (see Table 1 in Chi & Wylie (2014)). However, although these are characterized as indicators of cognitive engagement within the ICAP framework, the actions that indicate lower levels of cognitive engagement, such as listening and taking notes, also readily describe behavioral engagement as defined by Fredricks et al. (2004). Therefore, when focused on evaluating engagement related to specific actions students take in relation to an activity, it may be more difficult to clearly delineate between which actions indicate behavioral or cognitive engagement.

Social Engagement

One theme that appeared throughout all three dimensions of engagement was a social aspect. When talking about cognitive engagement, students also talked about discussion and collaboration with their group, including sharing ideas back and forth and helping others. Behaviorally, many students mentioned working with others and asking their group for feedback. Students also would refer to ideas related to liking or wanting to work with others when discussing emotional engagement. The idea of a social engagement dimension has previously been presented in a qualitative study conducted by Fredricks et al. (2016) who found the presence of a social component threaded throughout behavioral, cognitive, and emotional engagement dimensions when exploring middle- and high-school students' perceptions of engagement. They found that when students talked about engagement in their science classes, they included many social aspects related to sharing ideas and working together to solve problems (Fredricks, Wang, et al., 2016). Additionally, they found evidence for modeling social engagement as a separate, but related, dimension of engagement in a subsequent quantitative study (Wang et al., 2016).

The idea of social engagement also appears in the ICAP framework, where the highest level of cognitive engagement, *interactive*, is defined as dialoguing between students where both make generative comments during the discussion (Chi & Wylie, 2014). This suggests that students are expected to work with each other in order to be cognitively engaged at the highest mode. Collaboration and working with others has been seen as an important component to students' engagement at the college level, especially when considering STEM students. One study noted that STEM students found that positive collaborative environments encouraged them to be more engaged (Gasiewski et al., 2012), while another study noted that STEM students were generally represented within a 'culture of engagement' that included collaboration and problem-solving with peers (Brint et al., 2008).

Students in the interviews also noted that social interactions influenced their engagement. For example, when talking about their general engagement, one student stated that,

> "*I would say my engagement almost is based off the group. So if you have a group that's willing to actually work together and put in the effort, then my engagement's great. I have no problem keeping up. I have no problem, you know, being invested in the activity. But if you have a group that's just going to sit there silently over Zoom, then I'll try and do the activity by myself and that's fine, but you definitely lose engagement very quickly that way.*"

This sentiment, along with the presence of social themes throughout all three dimensions of engagement, indicate that students perceived a social engagement dimension when participating in these activities.

Perceptions of Engagement by Activity Environment

Results from one study completed immediately after the beginning of the Covid-19 pandemic found that students' perceptions about their engagement in class were influenced by the changes in environment and human interactions resulting from the shift from in-person to online learning (Jeffery & Bauer, 2020). Although in our study direct comparisons between in-person and remote environments cannot be made due to the differences in method and depth of data collection, it is worth noting some of the general similarities and differences in how students perceived engagement between the two environments. Many of the codes that were discovered were the same or very similar, which suggests that students perceived engagement similarly in both environments. For example, students from both environments discussed helping others, using resources, being prepared, and feeling like the activity was beneficial to their learning. Additionally, the overlap of perceptions related to both behavioral and cognitive engagement was found in both environments. This suggests that students conflated behavioral and cognitive engagement regardless of whether the activity was completed in-person or remotely.

There were also some codes that only appeared in one environment. Although some of these differences in codes may be due to the different methods of data collection, there are some inherent aspects to the environments that may have facilitated some of the differences. For example, leaving class early may be more obvious to others when physically in a classroom and, in the remote environment, students may have felt more comfortable reading questions out loud to themselves since other students couldn't hear them, whereas it might have been considered distracting in the in-person environment.

Additionally, some of the different codes were related to their interactions with other students, which was an inherently different experience in the two environments. For example, in the remote environment it is often difficult for people to have side conversations with a subset of group members since only one person can effectively talk at a single time over Zoom, which may have made off-topic conversations between just a couple people less likely to occur. For similar reasons, having a person step up as "group leader", sharing screens with each other, and listening to others may have been more important to engagement in the remote setting. These differences indicate that although social engagement may be an important aspect for students in both environments, how students perceive social engagement in these environments may vary slightly. For example, one student summarized how their interactions with peers in a general classroom had changed with the switch to remote instruction and how that influenced the social situation,

> "*In a [in-person] class setting, you find that one buddy. And if that, if you and that one buddy can go make other friends, that's great! But you don't have a time, you don't have a chance online to find that one buddy. And then maybe find another pair of buddies that you're also able to communicate with in that small group. Because when I go into a breakout room and it's six strangers, I can't, I can't be the first one and be like, 'Hey guys, what's up?'*"

Conclusions

This study explored students' perceptions of behavioral, cognitive, and emotional engagement in relation to worksheet activities completed in a general chemistry course. The results indicated that students perceived a variety of positive and negative aspects of behavioral, cognitive, and emotional engagement to be present during the worksheet activities in both in-person and remote learning environments; however, there were some overlaps between how students perceived behavioral and cognitive engagement. These overlaps may indicate a lack of distinction between the two dimensions of engagement when focusing on the worksheet-based activities. Studies that have quantitatively measured multiple dimensions of engagement in middle-school aged students have found evidence for both separating behavioral and cognitive dimensions (Wang et al., 2016), as well as combining the two to create a single "behavioral & cognitive" dimension (Ben-Eliyahu et al., 2018). One of the differences between these two studies was the focus of the measures. In the study completed by Wang et al. (2016), where the behavioral and cognitive dimensions remained separate, students were asked about their engagement in relation to their science class. However, in the study by Ben-Eliyahu et al. (2018), students were asked about their engagement related to specific science activities and the results indicated the presence of a combined dimension. Additionally, a recent quantitative study by Naibert and Barbera (2022) in higher-education, which assessed student engagement in active learning activities of general chemistry students, found further evidence for combining behavioral and cognitive engagement into a combined behavioral/cognitive dimension. The results of our qualitative and quantitative studies provide support for students' perceiving a large overlap between behavioral and cognitive engagement when asked specifically to think about the specific activities. Therefore, it may be that students conflate the two dimensions when the focus is on engagement in a specific activity instead of the class as a whole.

The results from this study also provided support for the existence of a social engagement dimension. Throughout all three dimensions of engagement (behavioral, cognitive, and emotional), students mentioned social aspects such as working with other students, including discussion, asking for feedback, and wanting to work with others. In other studies, collaboration with peers has been included as an indicator of higher student engagement in learning activities (Chi & Wylie, 2014) and has been found to be one of the defining aspects of the 'culture of engagement' seen in STEM courses (Brint et al., 2008). Ideas related to social engagement have also been discovered in qualitative studies of middle- and high-schoolers' perceptions of engagement in their science classes (Fredricks, Wang, et al., 2016), where the presence of a social engagement dimension centered around students' interactions with others in the classroom and a willingness to invest in those relationships was further supported during a subsequent quantitative study (Wang et al., 2016). Results from our study supported the presence of a similar social engagement dimension, as students would talk about their interactions with their group members (i.e., having discussions, helping their group members, etc.), as well as ideas related to *wanting* to work with other students and help them understand the material.

Responses from the in-person and remote environments indicated that students perceived behavioral, cognitive, and emotional engagement to be similar between the two environments. For example, students from both environments perceived engagement as trying to understand the material, helping others, staying focused, and liking/enjoying the activity. However, there were some differences in how students perceived engagement in the worksheet activities between the two environments, specifically related to social interactions (i.e., social engagement). For example, students' perceived disengagement in the in-person environment included participating in off-topic conversations, whereas in the remote environment disengagement was viewed as simply “being there” in the Zoom meeting. Many of the aspects that students found important in only one environment could have been due to the inherently different experience the two environments had to offer in terms of interacting with others (i.e., face-to-face vs. over a computer screen).

Limitations

One of the limitations of this study was that data were collected from a single institution and in relation to worksheet activities developed for a single course. Although data saturation was reached when coding responses from both types of environments, collecting data from other institutions, courses, and/or types of active learning activities would provide more generalizable insights into how students perceive engagement in active learning classrooms. Additionally, the data only included responses from students who self-selected to participate in the short-response survey or interviews; therefore, it is unknown if the results encompass the perceptions of all students in the course or those of varying engagement levels. Although students’ perceptions of activity-level engagement within the in-person and remote environments were reported, direct comparisons cannot be made due to the differences in the data collection methods employed.

Implications for Research

This study aimed to explore how students perceived engagement in worksheet activities implemented in the PSU general chemistry course. Although the results provided evidence that students’ perceive engagement similarly to the definitions provided by Fredricks et al. (2004), this may not be the case for every student population, learning environment, or active learning activity. Therefore, future work may benefit from collecting qualitative data about students' perceptions of engagement in other environments through the use of open-ended surveys, focus groups, or interviews.

One of the major findings of this study was the conflation in how students perceive behavioral and cognitive engagement. This finding, combined with support from our quantitative study (Naibert & Barbera, 2022), implies that these two dimensions of engagement cannot be measured and evaluated separately. Therefore, it is suggested that future studies combine these dimensions when seeking to evaluate students' engagement in learning activities. Although an overlap between these two dimensions has been noted in literature definitions (Fredricks et al., 2004; Sinatra et al., 2015), further studies are needed to understand why students might perceive behavioral and cognitive engagement similarly in these types of active learning activities. As studies at the middle-school level have found evidence for modeling behavioral and cognitive engagement as separate constructs when asking students about class-level engagement (Wang et al., 2016), as well as for modeling behavioral and cognitive engagement as a single construct when asking about activity-level engagement (Ben-Eliyahu et al., 2018), investigating student perceptions of engagement at various levels of focus (e.g., activity-specific, class-specific, etc.) could provide information about whether students perceive there to be more distinction between the two dimensions when focused on engagement in the course as a whole.

Another major finding of this study was support for the presence of a social engagement dimension, suggesting that students perceived social interactions and relationships to be an important factor in their engagement in the worksheet activities. Collaboration and working with other students have been noted previously as being an important aspect to students' engagement (Brint et al., 2008; Chi & Wylie, 2014; Wang et al., 2016). Therefore, future studies into students' engagement may benefit from considering the presence of a possible social engagement dimension. Additionally, further research into how students perceive social engagement in different active learning activities and/or the class in general may provide more insight into the importance of a social dimension in different environments and focuses of engagement.

Implications for Practice

Although results from this study indicated that students perceived behavioral, cognitive, and emotional engagement to be similar to the literature definitions of these dimensions, specific indicators of engagement may vary between different student populations and environments. For example, the data from the remote environment suggested that some students felt that sharing their screen was an aspect of engagement. Therefore, providing students with the ability to share their screen over Zoom may have encouraged students to engage more. Similarly, instructors may gain useful information about how to better engage their students in the material by asking their students for feedback on how they perceive engagement in a class or activity through the use of open-ended surveys, focus groups, or interviews.

Associated Content

Supporting Information

The Supporting Information is available in Appendix D and includes code descriptions and examples, coding results.

Author Information

Corresponding Author

*E-mail: jack.barbera@pdx.edu

Acknowledgments

We thank the students who participated in the surveys and interviews, as well as the instructors for allowing data collection in their courses. We thank Safaa El-Mansy for assistance with categorizing codes.

Chapter 8: Development and Evaluation of a Survey to Measure Student Engagement at the Activity Level in General Chemistry

Reprinted with permission from Naibert, N. & Barbera, J. (2022). Development and Evaluation of a Survey to Measure Student Engagement at the Activity Level in General Chemistry. *Journal of Chemical Education, 99*(3), 1410-1419. Copyright 2022 American Chemical Society.

Abstract

Student engagement is an important consideration when incorporating active learning activities into a classroom. To facilitate the large-scale assessment of students' engagement in activities, a survey measure must first be developed and evaluated. Therefore, the goal of this study was to create a self-report measure of student engagement for use with active learning activities in general chemistry classes. The Activity Engagement Survey (AcES) was modified from an existing survey of engagement of middle and high school science students that contained behavioral, cognitive, emotional, and social engagement items. Multiple rounds of response process interviews and factor analyses were used to modify the measure and provide evidence of validity for the data collected. The results provided support for the presence of a bifactor model, where an overall general engagement factor and the individual dimensions of engagement were simultaneously modeled. Additionally, support was found for combining behavioral and cognitive components into a single factor due to a large overlap in students' perceptions between the constructs.

Graphical Abstract

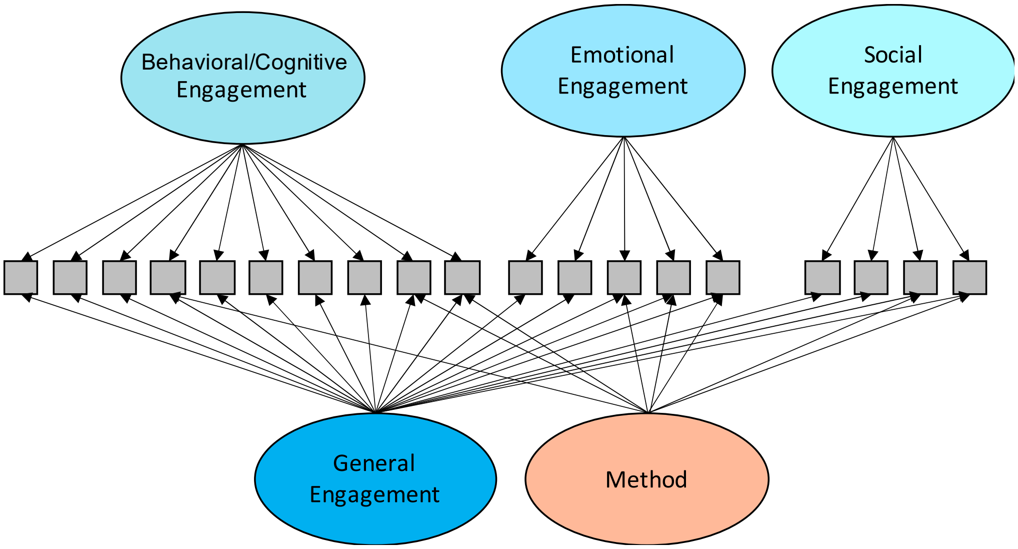


Figure 8.1. Graphical abstract for Chapter 8.

Keywords

First-Year Undergraduate / General, Chemical Education Research, Collaborative / Cooperative Learning, Student-Centered Learning

Introduction

One of the underlying goals of active learning is to shift the focus of learning from the instructor to the student by providing the opportunity for students to directly engage with the course material. As student engagement is considered to be a malleable state that can be easily influenced by the surrounding environment (Fredricks et al., 2004; Furlong & Christenson, 2008), understanding how students engage in individual activities within a class could be used to assess which activities are more engaging and for which students. As higher student engagement in a class has been linked to higher student performance outcomes (Chi & Wylie, 2014; Skinner et al., 2017), higher engagement in specific activities may be a contributing factor to increasing students' understanding of specific content material covered through a learning activity. Furthermore, students' engagement has also been included in models related to motivation (Zumbrunn et al., 2014) and student buy-in (Wang et al., 2021). However, before studies can model and explore the role of student engagement in active learning activities, an appropriate measure of student engagement is required.

Student engagement is generally considered to be a meta-construct composed of multiple dimensions (Fredricks et al., 2004). These often include behavioral, cognitive, and emotional engagement. Behavioral engagement is generally defined as students' involvement in the classroom (Fredricks et al., 2004; Sinatra et al., 2015). For a specific activity, this may include behaviors such as asking questions, participating, etc. Cognitive engagement is often conceptualized as students' investment in their learning (Fredricks et al., 2004; Sinatra et al., 2015), which can include trying to understand their mistakes and trying to connect new ideas to previously learned concepts. Emotional engagement is the students' affective reactions to their surrounding environment, including their peers, the instructor, and the activity itself. Emotional engagement can be composed of many different emotions, such as value, boredom, and interest (Fredricks et al., 2004; Sinatra et al., 2015). As engagement is theorized to be a meta-construct, simultaneously assessing all three dimensions of engagement can allow greater insight into how these dimensions relate and interact with each other and performance outcomes. In addition to these three dimensions of engagement, a social aspect to engagement has also been conceptualized as an important component in definitions of engagement related to STEM students (Brint et al., 2008). Social engagement is described as students' interactions with others and the willingness to build relationships (Fredricks, Wang, et al., 2016; Wang et al., 2016).

Collaborative social interactions have also been included as a defining component when considering cognitive engagement (Chi & Wylie, 2014) and may be especially important when students are directed to work with others.

Multiple methods have been used to gauge student engagement in the classroom. A recent study by McNeal et al. (2020) found that biosensors could be used to assess engagement through measuring changes in the galvanic skin response, although the researchers noted that this technique may be difficult to generalize due to individuals' physical differences. Observational methods have also been developed for measuring student engagement with respect to behavioral engagement (Harris & Cox, 2003; Lane & Harris, 2015) and an observational protocol has been developed to assess students' cognitive engagement through mapping overt student behaviors onto different levels of engagement (Chi & Wylie, 2014). However, these methods are often limited by only allowing the assessment of a subset of students. Additionally, assessing students' cognitive and emotional engagement through observations is not recommended due to the internal nature of these dimensions (Appleton et al., 2006).

Self-report survey measures can be used to simultaneously collect data about all dimensions of engagement, while also allowing data to be collected for every student. In higher-education STEM courses, there have been a few survey measures designed or adapted to measure single dimensions of engagement in the classroom. For example, a measure created by Gasieswski et al. (2012) focused on assessing behavioral engagement in introductory STEM courses and a study completed by Seery (2015a) used a cognitive engagement scale to gather data about student engagement in a flipped classroom. Additionally, a survey to measure both behavioral and emotional engagement in STEM courses was developed by Skinner et al. (2017). Although these surveys measure different dimensions of engagement, none of them were designed to simultaneously assess behavioral, cognitive, emotional, and social engagement.

In the K-12 literature, multiple surveys of student engagement have been created (Fredricks & McColskey, 2012). In 2016, Wang et al. (2016) developed a survey to measure behavioral, cognitive, emotional, and social engagement of middle and high school science students. The measure was developed through a series of interviews with students, where they were asked to describe what they were doing, thinking, and feeling when engaged in class (Fredricks, Wang, et al., 2016). Although this survey was not developed for higher-education students or for active learning activities, this measure contained items for all four dimensions of engagement and was specific to science students, which made it a good candidate as the basis for a modified measure of student engagement in active learning activities in general chemistry. However, when surveys are modified and used in a different student population and/or environment, it is essential to collect evidence of validity and reliability of the data gathered with the measure before results are interpreted. This evidence provides support that the results of the survey reflect the construct of interest within the new population or setting.

There are multiple types of validity evidence: test content, response process, internal structure, association with other variables, and consequential validity (Arjoon et al., 2013; Knekta et al., 2019; Wren & Barbera, 2013). Collecting sufficient evidence of *test content* validity provides confidence that the survey measures the construct it is intended to measure. This can include evidence of content validity, which assesses if the measure is a representation of the construct (Arjoon et al., 2013), and face validity, which assesses if students perceive the measure as a representation of the construct (Crocker & Algina, 1986). Evidence of *response process* validity provides confidence that students interpret the survey items in the manner intended (Arjoon et al., 2013). Collecting evidence of response process validity is important when survey items are modified, as students from a different population and/or environment may not interpret the items the same way as was originally intended. *Internal structure* validity focuses on evaluating whether the relation between the survey items and the construct matches the hypothetical structure. Sufficient evidence of internal structure validity provides confidence that the survey items are related to the construct in the expected manner and is important to collect when surveys are modified and/or used in a new population or environment (Arjoon et al., 2013; Knekta et al., 2019). Validity evidence related to *association with other variables* can be used to evaluate the construct by assessing its relation to other related constructs or outcomes (Knekta et al., 2019) and providing evidence of *consequential* validity allows for the comparison of the construct between different groups (e.g., demographic groups) (Rocabado et al., 2020).

The main goal of this study was to develop a measure of student engagement (the Activity Engagement Survey (AcES)) for use with activities employed in the general chemistry classroom. To do this, we modified the measure of engagement originally developed by Wang et al. (2016) for middle and high school students and collected evidence of response process and internal structure validity, which included evaluating different factor structures that would best represent the multidimensional nature of engagement. Evidence of test content validity was not directly assessed in this study, as there was support for test content validity from other sources: first, the original survey measure provided evidence of test content validity in that items were developed through student interviews and the theoretical definitions of engagement (Fredricks, Wang, et al., 2016); second, in an initial study (Naibert et al., 2022), we found that students in general chemistry courses at our institution perceived the dimensions of engagement similarly to the theoretical definitions. Additionally, although association with other variables and consequential validity are important, they are beyond the scope of this development manuscript and will be fully evaluated in a separate study.

Research Questions

1) How well does a modified measure of student engagement function with our student population and active learning environment?
2) Which factor structure best represents the data collected with the measure?

Course Information

Data were collected over all three sequential terms of the 2020-2021 academic year in the general chemistry (GC) course at Portland State University (PSU). This course consisted of two sections taught by a total of three instructors (Table 8.1). Self-reported demographics data for students who selected to participate in this study during Fall term are provided in Table E.1 in Appendix E. As this study took place during the COVID-19 pandemic, the courses were all taught in a synchronous remote format, where the students were expected, but not required, to attend each class session remotely during assigned class times using the video platform software Zoom. Each course section consisted of both lecture days and activity days and students were aware of the schedule for each class period.

Table 8.1. Course information for each general chemistry (GC) section and term.

Term	**Instructor**	**Week-1 enrollment**
Fall (GC I)	A	N = 297
	B	N = 332
Winter (GC II)	A	N = 190
	B	N = 266
Spring (GC III)	A	N = 186
	C	N = 156

Activity days were centered around worksheet activities, which included key questions, exercises, and problems related to provided models about a single chemistry topic, similar in design to process-oriented guided-inquiry learning (POGIL) materials (Hanson et al., 2018). During the fall term, students were randomly assigned into Zoom breakout rooms with around 3 – 7 students for the first two activities. For all subsequent activities, students were asked to select the type of breakout room they wanted to join for the remainder of the term. For these pre-assigned breakout rooms, students had the option of choosing to 1) work with others with their video camera on, 2) work with others with their video camera off, 3) work independently in a breakout room with the ability to ask a learning assistant (LA) questions, or 4) work completely independently in the main room. Students who chose to work with others (i.e., options 1 or 2) were assigned to the same breakout room for the term, which typically contained about 4 students per room. All students were expected to work through the worksheet regardless of if they chose to work in a breakout room or the main room. During the winter and spring terms, the same options were utilized but students made their selection at the start of each term (i.e., prior to conducting the first activity). For all activities each term, undergraduate and graduate LAs were each assigned to 1 – 2 rooms, where they answered questions and helped to guide students through the worksheet. The instructor and lead teaching assistant (TA) facilitated the activity through check-ins with the LAs. Although classes were taught by different instructors, each instructor implemented the activities similarly.

Methods

Survey Items

The survey contained items related to students' behavioral, cognitive, emotional, and social engagement in the worksheet activities. The items were adapted from Wang et al. (2016) and underwent several rounds of modifications during the prior year (2019 – 2020) when the class was conducted in-person, which resulted in the pilot survey. Details about these modifications, including results from response process and initial factor analyses, are included in Appendix E (see Table E.2 for summary). The pilot survey contained a total of 26 items, which included 7 behavioral, 6 cognitive, 7 emotional, and 6 social engagement items. The social engagement items were only presented to students who selected an option on the survey indicating that they worked on the worksheets with other students in a Zoom breakout room during class. Students who didn't select this option were only presented with the behavioral, cognitive, and emotional engagement items, for a total of 20 items. All survey items were administered on a 6-point Likert-type scale from *strongly disagree* (1) to *strongly agree* (6).

Data Collection

All data collected within this study was approved by the Institutional Review Board (IRB) at Portland State University and appropriate consent was obtained from students as required by the IRB.

Quantitative survey data were collected after three activity days each term, for a total of nine different activities over the course of the year. Students were notified of each survey through an in-class announcement on the day of the activity, as well as through an announcement posted to each section's learning management site, which included a link to the Qualtrics survey. Students were given 48 hours after each activity to access and complete the survey. This brief window was selected to give students adequate time to access and complete the survey while still ensuring it was completed before the next class period in order to encourage the students to reflect on the activity specifically and not the class as a whole. Additionally, it was intended that this brief window would help to reduce the inherent recall bias of retrospective reports. All students who accessed the survey were awarded a nominal amount of extra credit regardless of consent or completion. Between 71 – 80% of students each term consented to participate in at least one survey. Survey responses collected during each activity were aggregated into an overall data set covering all nine activities. Only responses from students who consented to a survey and indicated that they completed an activity during the class time were included in the data. Two types of check items were used to clean the response data; a topic-based check item and a response-based check item. The topic-based check item asked students to select the topic of the material that was covered during the activity. This check item was used to remove responses from students who likely did not participate in the activity. The response-based check item was included with the survey items and asked students to select a specific response (e.g., *somewhat agree*) to screen for students who did not read the survey items before responding. A total of 8% of responses were removed due to incorrect check items. After cleaning, a total of 1287 responses, across all nine activities, remained.

At the end of select surveys during the first two terms, students were asked if they were interested in participating in a short interview about the survey. During the first term, all interested students were sent an email asking them to fill out an online consent form and select their availability. During the second term, students who indicated they were interested in participating in an interview were randomly selected based on their response to the type of breakout room they participated in during the activity. A total of 21 students participated in an interview; 13 students during the first term and 8 students during the second term. Of the second term interviews, 5 students had worked in a breakout room with their cameras on, 1 had worked in a breakout room with cameras off, and 2 had worked by themselves in the main room. All interviews were completed remotely over Zoom with audio and visual recording. During the interview, students were first provided a link to a blank copy of the survey through the chat function in Zoom. Students were asked to think back to the activity and then fill out the survey again. After they had finished, they were directed to read each of the items out loud, state which response they selected, and then explain why they selected the response they did. If needed, follow-up questions were asked by the researcher to gain additional details or clarification. Students who indicated they did not work with other students on the activity were not asked to respond to the items related to social engagement.

Response Process Interview Analysis

Interviews were analyzed to provide evidence of response process validity. Two researchers first individually listened to and watched two interviews and noted if the students had difficulty reading any items and whether their explanation matched their given response to each item. The two researchers then came together to discuss explanations that seemed unclear or not in alignment with the given response. The researchers then individually analyzed additional interviews (2 – 4 each cycle) before coming together to discuss the responses from those interviews. This cycle continued until all interviews were analyzed and discussed. The researchers then consolidated the findings for each item and assessed whether each item appeared to function well or was unclear or confusing to students. Data from the interviews were used to both remove any confusing items prior to quantitative analysis, as well as to support the removal of items during quantitative analysis.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) was used to gather evidence of internal structure validity of the survey response data. CFA is generally used when there is an *a priori* hypothesis of the factor structure to be assessed. All CFAs were completed using the statistical program R (version 3.6.2) with the package lavaan (version 0.6-5). Maximum likelihood with Satorra-Bentler adjustment and robust standard errors were used to account for any non-normality of the data (Satorra & Bentler, 1988). Each model was identified by setting the factor variance of each factor to 1. All negatively worded items were reverse coded before analysis. As engagement is considered to be a malleable state that can vary based on the environment (Fredricks et al., 2004; Furlong & Christenson, 2008), all responses were treated as independent since each survey was directed towards students' engagement in a specific activity. Listwise deletion was used to remove incomplete responses prior to analysis. Data-model fit was assessed using standard recommended values for good fit: CFI ≥ 0.95, TLI ≥ 0.95, RMSEA ≤ 0.06, and SRMR ≤ 0.08 (Hu & Bentler, 1999). Factor loadings (λ) and modification indices (MIs) were used to evaluate possible modifications to the scales during initial analysis. Reliability evidence was collected for individual factors through the use of omega (Komperda, Pentecost, et al., 2018), where values greater than 0.7 are generally considered to indicate acceptable reliability.

Results and Discussion

Before using CFA to evaluate the combined engagement scale, response process interviews and individual scale CFAs were used to analyze the data-model fit of the separate behavioral, cognitive, emotional, and social pilot scales. After some modifications, evidence of good internal structure validity and good reliability were found for each scale. Details about the item modifications made to each individual scale, as well as fit and reliability statistics, are included in Appendix E (Tables E.3 – E.7). The final 19-item engagement survey (titled the Activity Engagement Survey (AcES)) included 5 behavioral, 5 cognitive, 5 emotional, and 4 social items. Descriptive statistics for each item are included in Appendix E (Table E.8).

Model Selection

As not all students worked with others, the data collected with only the 15 AcES behavioral, cognitive, and emotional items were evaluated first. This included students who worked on the activities independently, as well as students who worked socially (i.e., in a breakout room with other students). Different models were assessed using CFA to determine which was the most appropriate through evaluating data-model fit statistics. The first was a unidimensional model (Figure 8.2) where engagement was modeled as an overall factor without separable dimensions. As expected, due to the theoretical multi-dimensional nature of engagement, the data-model fit statistics indicated poor fit with this unidimensional model (Table 8.2). Thus, the next model tested was a three-factor (B-C-E) correlated model, where the factors of behavioral (B), cognitive (C), and emotional (E) engagement are separate but correlated (Figure 8.3). The fit statistics with the B-C-E correlated model showed evidence of reasonable data-model fit (Table 8.2). However, the factor correlation between the behavioral and cognitive factors was high (0.953). High factor correlations can indicate that the two factors are measuring an identical construct. Thus, a two-factor (BC-E) correlated model was tested where the behavioral and cognitive factors were combined into a behavioral/cognitive (BC) factor (Figure 8.4). The data-model fit statistics for this model also indicated reasonable fit (Table 8.2). Although both models showed similar fit, the high factor correlation between behavioral and cognitive engagement in the B-C-E model provides support for the BC-E model, where behavioral and cognitive engagement is conceptualized as a single combined factor. Additionally, in a concurrent study (Naibert et al., 2022), students mentioned aspects related to both behavioral and cognitive engagement when asked to describe their engagement in the worksheet activities. This conflation of engagement types was irrespective of the definition they were provided at the time (i.e., behavioral or cognitive engagement). These qualitative results suggested that students perceived the behavioral and cognitive dimensions as similar ideas when considering engagement in these worksheet activities and provide additional support for combining the behavioral and cognitive dimensions. Although all dimensions of engagement are interconnected, the literature definitions for behavioral and cognitive engagement have noted that these two dimensions may be difficult to separate, especially when cognitive engagement is conceptualized as students' psychological investment in their learning (Sinatra et al., 2015). Other quantitative studies have also found support for combining the behavioral and cognitive engagement factors into a single factor when measuring behavioral, cognitive, and emotional engagement at the activity level (Bathgate & Schunn, 2017; Ben-Eliyahu et al., 2018).

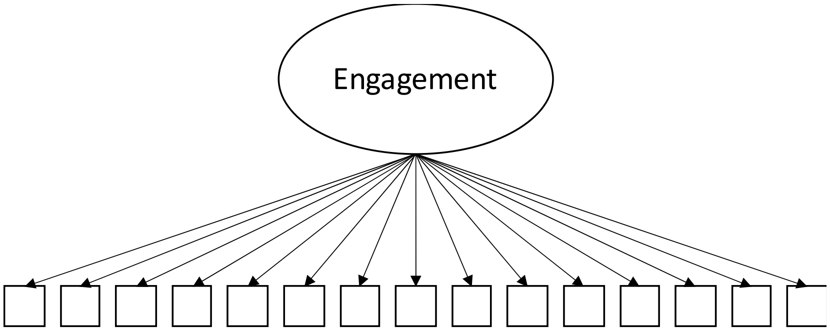


Figure 8.2. Unidimensional model of engagement, where all 15 AcES items are related to a general engagement factor.

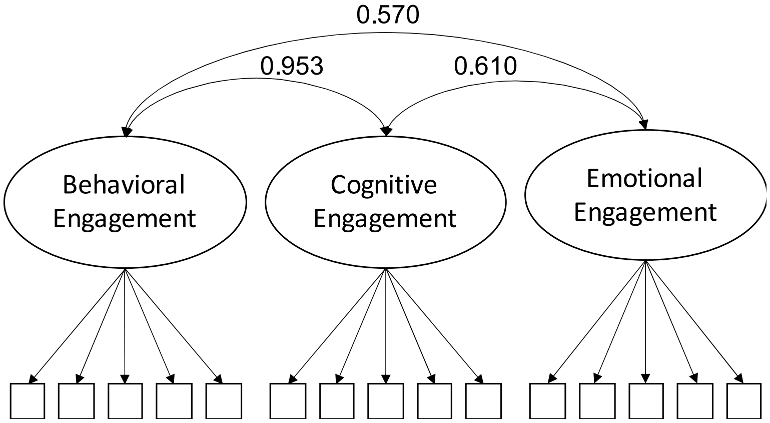


Figure 8.3. B-C-E correlated model where the individual behavioral, cognitive, and emotional engagement dimensions are correlated. Standardized factor correlations found through CFA are included.

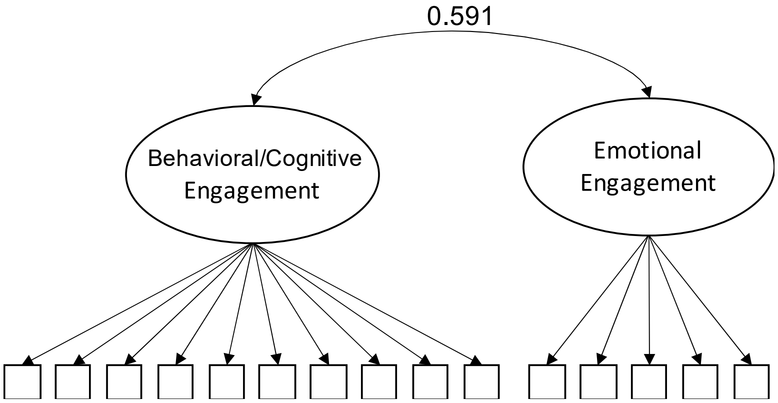


Figure 8.4. BC-E correlated model with a combined behavioral/cognitive engagement factor and individual emotional engagement factor. The standardized factor correlation found through CFA is included.

In addition to a correlated model, there are other types of models which may better represent the data. One of these models is the bifactor model. The bifactor model includes a general engagement factor that takes into account the commonality among the dimensions, as well as the separate factors for the individual dimensions that represent the contribution of those dimensions above and beyond the general factor. This model allows for both the individual dimensions and general overarching engagement factor to be evaluated simultaneously and has been suggested as the most appropriate model when analyzing multidimensional constructs similar to engagement (Chen et al., 2012; Chen et al., 2006). Additionally, as both positively and negatively worded items were included in this survey, a bifactor model that included the addition of a method factor was also tested. Method factors can be used to account for possible differences in how students respond to positively and negatively worded items (Wang et al., 2016; Wang et al., 2015; Ye & Wallace, 2013). Although including both positively and negatively worded items can encourage students to think about their responses instead of simply responding *agree* to all the items, students may respond differently to these two types of items (Wang et al., 2015). For example, students may be more likely to *strongly agree* with a positively worded item but less likely to *strongly disagree* with the same negatively worded item. Therefore, first an engagement-only bifactor model was analyzed that included a general engagement factor and separate behavioral/cognitive and emotional engagement factors (Figure 8.5). Then, a bifactor model with a negative method factor included was analyzed to account for the relation between the negatively worded items (Figure 8.6). The fit statistics from the engagement-only bifactor model indicated reasonable to good data-model fit, which was found to improve to good data-model fit with the addition of the negative method factor (Table 8.2).

During the development of the original measure, Wang et al. (2016) also found evidence to support a bifactor model with an added method factor as the best model in their study. Additionally, other studies on engagement of middle-school students have also found support for the bifactor model (Ben-Eliyahu et al., 2018). Beyond theoretical considerations for modeling student engagement with a bifactor model due to the multidimensional nature of engagement, there are also additional practical applications. Some studies that have focused on the relation between student engagement and motivation of middle-school students have found that behavioral/cognitive and emotional engagement relate differently to various aspects of motivation (Bathgate & Schunn, 2017; Ben-Eliyahu et al., 2018). For example, one study found that emotional engagement was related to changes in fascination, value, and competency beliefs, while behavioral/cognitive engagement was not found to be related to any change in motivation (Bathgate & Schunn, 2017). Although a correlated model (e.g., Figure 8.4) can be used to investigate the relations between the engagement dimensions and other constructs and outcomes, it does not allow relations with the general engagement factor to be explored.

A bifactor model was used in one of the middle-school studies, which allowed them to find that the students' general engagement in an activity was related to their self-efficacy, mastery, and performance (Ben-Eliyahu et al., 2018). In addition to representing the multidimensional nature of engagement with a bifactor model, including a method factor to account for response bias due to the wording of the items not only was also supported in the original survey development by Wang et al. (2016), but can also reduce bias when estimating the relation between the construct of interest (i.e., engagement) and other constructs (Gu et al., 2015). Therefore, due to results from similar surveys, practical applications, and evidence of good data-model fit, the bifactor model with a negative method factor was selected as the most appropriate model for evaluating the data set containing both students who worked independently and students who worked socially.

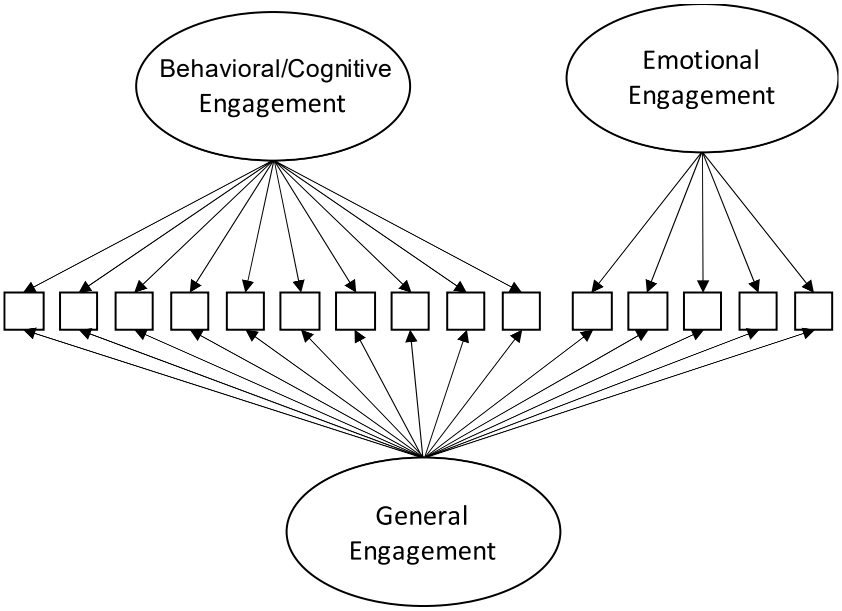


Figure 8.5. The 15-item AcES BC-E bifactor model with an overarching general engagement factor that accounts for commonality between all items, as well as separate behavioral/cognitive and emotional engagement factors that represent engagement for those dimensions above and beyond what is included in the general engagement factor.

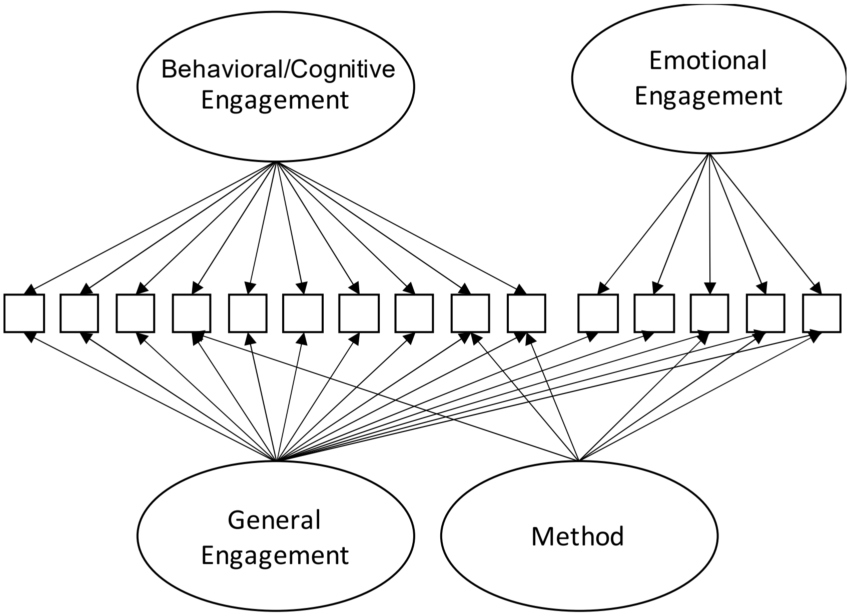


Figure 8.6. The 15-item AcES BC-E bifactor model with a negative method factor that accounts for the relation between negatively worded items.

Table 8.2. Data-model fit statistics for different models with the entire data set (individual and social students) (n = 1248). Bold values indicate the results met the suggested criteria based on recommendations from Hu and Bentler (1999).

Model	**χ^2 (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**
Unidimensional	1508.221 (90)	<0.001	0.770	0.731	0.136 [0.130 – 0.142]	0.095
B-C-E Correlated	495.374 (87)	<0.001	0.937	0.924	0.072 [0.066 – 0.078]	**0.051**
BC-E Correlated	518.093 (89)	<0.001	0.935	0.923	0.073 [0.067 – 0.079]	**0.052**
BC-E Bifactor	377.243 (75)	<0.001	**0.955**	0.937	0.066 [0.060 – 0.073]	**0.037**
BC-E Bifactor with negative method factor	244.350 (69)	<0.001	**0.974**	**0.960**	**0.053** **[0.046 – 0.060]**	**0.030**

Item Analysis

For the BC-E bifactor model with the negative method factor (Figure 8.6), all item loadings were statistically significant ($p < 0.05$) on their respective factors including both the general and dimension-specific engagement factors (Table 8.3). Although most behavioral/cognitive items loaded strongly ($\lambda > 0.4$) on both the behavioral/cognitive and overall engagement factors, most had higher loadings on the general engagement factor than the specific behavioral/cognitive factor. This indicates that the general engagement factor explained more variance in these items than the specific factor. Additionally, three cognitive-based items had loadings of less than 0.4 on the specific behavioral/cognitive factor: *I made sure I understood my work on today's activity, I tried to connect what I was learning during today's activity to concepts I have learned before,* and *I wrote down the answers to today's activity without trying to understand them.* This suggests that indicators related to students' understanding of the work and making connections to previous material may be better indicators of students' general engagement instead of specifically behavioral/cognitive engagement. Additionally, the item, *I didn't think very hard when I came across a challenging problem on today's activity,* was found to have a low loading related to the general engagement factor, suggesting it was a better indicator of students' specific behavioral/cognitive engagement above and beyond general engagement. Approximately half of the emotional engagement items loaded higher on the general engagement factor compared to the emotional engagement factor, although most had reasonably high loadings on both factors ($\lambda > 0.4$). One item, *I didn't care about doing today's activity*, was found to have a low loading on the specific emotional engagement factor, which suggests that this item better assessed students' general engagement.

Table 8.3. Standardized factor loadings for each AcES item on the behavioral/cognitive (BC) and engagement (E) scales. Negatively worded items were reverse coded before analysis (rev).

AcES BC-E Items	Standardized Factor Loadings		
	General Engagement	Dimension-specific Engagement	Negative Method
Behavioral/Cognitive Engagement			
I stayed focused during today’s activity.	0.576	0.418	--
I put effort into doing today’s activity.	0.566	0.516	--
I kept working on today’s activity even if something was hard.	0.557	0.534	--
I didn’t do much work on today’s activity. (rev)	0.570	0.420	0.167
I attempted to answer most of the items on today’s activity.	0.525	0.544	--
I made sure I understood my work on today’s activity.	0.718	0.244	--
I tried to connect what I was learning during today’s activity to concepts I have learned before.	0.569	0.268	--
I tried to understand my mistakes when I got something wrong during today’s activity.	0.530	0.444	--
I wrote down the answers to today’s activity without trying to understand them. (rev)	0.442	0.308	0.124
I didn’t think very hard when I came across a challenging problem on today’s activity. (rev)	0.337	0.477	0.288
Emotional Engagement			
I looked forward to today’s activity.	0.511	0.740	--
I enjoyed learning the class material during today’s activity.	0.688	0.397	--
I thought that today’s activity was boring. (rev)	0.527	0.400	0.375
I didn’t want to do today’s activity. (rev)	0.517	0.543	0.363
I didn’t care about doing today’s activity. (rev)	0.567	0.275	0.426

Social-Focused Engagement Scale

The students who indicated they worked with others on the activity were also asked to respond to items related to their social engagement. The entire survey, which included all 19 AcES items related to behavioral, cognitive, emotional, and social (S) engagement, was analyzed. Similarly to the prior analysis, unidimensional, correlated (B-C-E-S and BC-E-S), and bifactor models were tested (Table 8.4). The most appropriate model for the 19-item AcES was also found to include a combined behavioral/cognitive factor within a bifactor model with a negative method factor (Figure 8.7, Table 8.4).

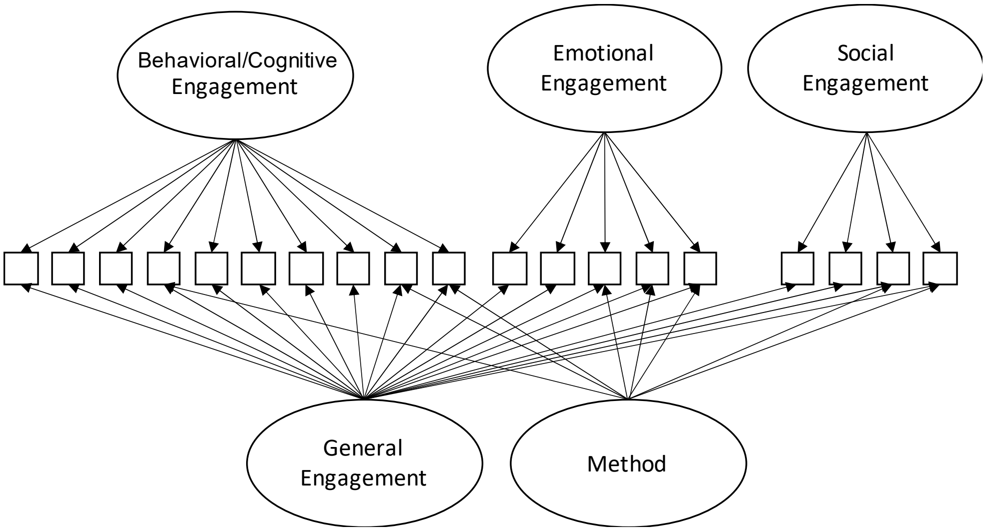


Figure 8.7. The 19-item AcES BC-E-S bifactor model with a negative method factor.

Table 8.4. Data-model fit statistics for different models using the social data set (n = 853). Bolded values indicate results met the suggested criteria based on recommendations from Hu and Bentler (1999).

Model	**χ^2 (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**
Unidimensional	1236.803 (152)	<0.001	0.775	0.747	0.113 [0.108 – 0.119]	0.083
B-C-E-S correlated	497.499 (146)	<0.001	0.930	0.918	0.064 **[0.058 – 0.071]**	**0.052**
BC-E-S correlated	507.586 (149)	<0.001	0.929	0.918	0.064 **[0.058 – 0.707]**	**0.053**
BC-E-S Bifactor	408.351 (133)	<0.001	0.946	0.930	**0.059** **[0.053 – 0.066]**	**0.039**
BC-E-S Bifactor with negative method factor	312.407 (124)	<0.001	**0.963**	0.949	**0.051** **[0.044 – 0.058]**	**0.037**

Most of the item factor loadings in the BC-E-S bifactor model with negative method factor (Figure 8.7) were found to be statistically significant ($p < 0.05$) on both the general and their respective dimension-specific engagement factors (Table 8.5). The exception to this included the behavioral/cognitive item, *I tried to connect what I was learning during today's activity to concepts I have learned before*, which did not significantly load on the specific behavioral/cognitive factor and suggests that this item was not an indicator of behavioral/cognitive engagement, although it was a significant indicator of students' general engagement. Although other items loaded at a statistically significant level, many of the items from the behavioral/cognitive factor did not load strongly on the specific factor (λ < 0.4). This suggests that the behavioral/cognitive engagement items may be better indicators of students' overall engagement when social engagement is included in the model rather than specifically their behavioral/cognitive engagement.

Table 8.5. Standardized factor loadings for each AcES item on the behavioral/cognitive (BC), engagement (E), and social (S) scales. Negatively worded items were reverse coded before analysis (rev). Factor loadings that were not statistically significant with a $p < 0.05$ cutoff are in italics.

AcES BC-E-S Items	**Standardized Factor Loadings**		
	General Engagement	**Dimension-specific Engagement**	**Negative Method**
Behavioral/Cognitive Engagement			
I stayed focused during today's activity.	0.702	0.134	--
I put effort into doing today's activity.	0.723	0.167	--
I kept working on today's activity even if something was hard.	0.687	0.295	--
I didn't do much work on today's activity. (rev)	0.663	0.304	0.206
I attempted to answer most of the items on today's activity.	0.646	0.573	--
I made sure I understood my work on today's activity.	0.705	0.124	--
I tried to connect what I was learning during today's activity to concepts I have learned before.	0.642	*0.057*	--
I tried to understand my mistakes when I got something wrong during today's activity.	0.646	0.140	--
I wrote down the answers to today's activity without trying to understand them. (rev)	0.525	0.144	0.132
I didn't think very hard when I came across a challenging problem on today's activity. (rev)	0.540	0.185	0.207
Emotional Engagement			
I looked forward to today's activity.	0.418	0.805	--
I enjoyed learning the class material during today's activity.	0.552	0.498	--
I thought that today's activity was boring. (rev)	0.473	0.467	0.361
I didn't want to do today's activity. (rev)	0.452	0.574	0.376
I didn't care about doing today's activity. (rev)	0.528	0.315	0.391
Social Engagement			
I built on other students' ideas during today's activity.	0.505	0.661	--
I tried to understand other students' ideas during today's activity.	0.568	0.285	--
I didn't share ideas when working with other students during today's activity. (rev)	0.567	0.335	*0.072*
I didn't like working with other students during today's activity. (rev)	0.449	0.331	0.247

Conclusions

Both research questions, 1) *How well does a modified measure of student engagement function with our student population and active learning environment?* and 2) *Which factor structure best represents the data collected with the engagement measure?* were addressed through the development process to create the Activity Engagement Survey (AcES) measure. Response process interviews and initial CFAs were used to modify an existing survey of student engagement (Wang et al., 2016) for our student population (i.e., university students) and environment (i.e., activity-focused). Evidence of both internal structure validity and reliability for each of the independent dimensions was found for the data collected with this modified measure. Results from CFA of the data collected with the AcES indicated that the bifactor model with a negative method factor and combined behavioral/cognitive factor showed evidence of good data-model fit. This was the most supported model for the AcES BC-E (which only included the behavioral, cognitive, and emotional items) and the AcES BC-E-S (which included the behavioral, cognitive, emotional, and social items). The presence of the bifactor model suggests that engagement in these worksheet activities can be modeled as students' overall engagement in the activities with separate components related to the specific factors, allowing for both the overall and separate components to be measured simultaneously. In future studies, the bifactor model could allow for further simultaneous exploration into how certain student learning outcomes relate to overall engagement, as well as separate factors of behavioral/cognitive, emotional, and social engagement. The evidence for combining the behavioral and cognitive factors provides support that students tend to perceive these two engagement dimensions similarly when focused on engagement in these specific activities. This result supports what was found in a separate qualitative study when students in this course were asked to describe engagement in these activities based on the definitions of behavioral, cognitive, and emotional engagement (Naibert et al., 2022). Although a combined behavioral/cognitive factor was not discovered in the study by Wang et al. (2016), other studies with middle-school students found support for combining the two factors when measuring student engagement in specific science activities (Bathgate & Schunn, 2017; Ben-Eliyahu et al., 2018). As the original items from Wang et al. (2016) were focused on engagement at the class-level instead of the activity-level, it may be that students are more likely to conflate behavioral and cognitive engagement when asked to reflect on their engagement in specific activities or tasks. The idea that individual behaviors and cognitive engagement are closely related in specific tasks has also been used as the basis of an observational measure of student engagement through the ICAP (Chi & Wylie, 2014) (interactive-constructure-active-passive) framework, which relates students' overt behaviors to different levels or modes of cognitive engagement.

Limitations

This study was completed at a single institution and with worksheet activities developed for general chemistry. Thus, the results of this study may not be generalizable to other student populations, classroom environments, or activities. Additionally, this study took place while classes were held remotely over Zoom and breakout rooms were used to facilitate groupwork with the activities. Therefore, without further study, it is unknown if the Activity Engagement Survey (AcES) measure will function similarly in an in-person active learning environment.

Implications for Research

The goal of this project was to create a measure of student engagement that could be used with the active learning activities included in our general chemistry courses. Therefore, evidence of validity and reliability of data collected with the Activity Engagement Survey (AcES) measure should be gathered before results are interpreted in a different student population or environment or with different active learning activities (Knekta et al., 2019). However, if sufficient evidence of validity and reliability is found, data collected with AcES could be used for many different types of analyses. The presence of the bifactor model allows for the simultaneous measurement of students' overall engagement, as well as the behavioral/cognitive, emotional, and social engagement components above and beyond students' general engagement. Future studies could focus on exploring relations between students' overall, behavioral/cognitive, emotional, and social engagement in learning activities to student performance outcomes (e.g., grades, etc.). Additionally, if evidence of consequential validity for AcES data is supported through measurement invariance, differences in each of the factors could be compared between different activities or different student groups (e.g., demographic groups). Another possibility is the use of latent profile analysis to investigate whether different engagement profiles exist among students. In a study focused around engagement of high school students, five engagement profiles were discovered (Wang & Peck, 2013). They found that some students were moderately engaged across behavioral, cognitive, and emotional dimensions, while others were highly engaged in all three, minimally engaged in all three, minimally engaged in only the emotional dimension, or minimally engaged in only the cognitive dimension. These profiles were found to be related to certain outcomes such as GPA, college enrollment, etc. Therefore, if different engagement profiles with respect to specific activities are discovered using the AcES measure, they could also be used to explore if and how outcomes are related to different profiles.

Implications for Practice

As higher student engagement in general has been associated with higher student performance outcomes (Chi & Wylie, 2014; Skinner et al., 2017), discovering potential differences in student engagement between different activities could be useful for instructors and may allow them to adjust activities to increase student engagement overall. Additionally, feedback about student engagement may also provide valuable information about how facilitation techniques may influence students' experiences in the course. For example, one study that explored student argumentation found that differences in instructor facilitation and expectations influenced how students' interacted with POGIL activities (Stanford et al., 2016). Therefore, using the AcES measure to gather feedback about student engagement in learning activities may not only provide information about differences between activities, but also feedback about the structure and facilitation of the activities.

Associated Content

Supporting Information

The Supporting Information is available in Appendix E and includes demographics, preliminary survey item modifications, and survey scale and item modifications.

Author Information

Corresponding Author

*E-mail: jack.barbera@pdx.edu

Chapter 9: Investigating Student Engagement in General Chemistry Active Learning Activities using the Activity Engagement Survey (AcES)

Reprinted with permission from Naibert, N. & Barbera, J. (2022). Investigating Student Engagement in General Chemistry Active Learning Activities using the Activity Engagement Survey (AcES). *Journal of Chemical Education* (In Review 03/30/2022). Copyright 2022 American Chemical Society.

Abstract

Investigating student engagement in active learning activities could provide valuable insight into variations of student learning outcomes when active learning is included in a course. This study sought to explore students' engagement in relation to active learning activities incorporated in a general chemistry lecture course using the Activity Engagement Survey (AcES). The AcES can be used to simultaneously assess students' overall engagement, as well as dimensions above and beyond overall engagement including their combined behavioral/cognitive engagement, emotional engagement, and social engagement. As students' engagement may be influenced by aspects related to the learning environment and context, differences in engagement were explored between students who chose to work on active learning activities with others and those that chose to work independently, as well as between activities completed in a remote environment versus an in-person environment. Results indicated that students who worked with others had significantly higher behavioral/cognitive and overall engagement than those who worked by themselves. Comparisons of students' engagement between the two types of learning environments, however, could not be justified due to insufficient evidence to support consequential validity. As higher student engagement in a course has been found to lead to improved student learning outcomes, the relation between students' engagement in the activities and students' understanding was assessed through multiple linear regression. Results indicated that students who were more behaviorally/cognitively engaged in an activity scored higher on exam items related to the content covered on the activity, although emotional and social engagement were not found to be significant predictors.

Graphical Abstract

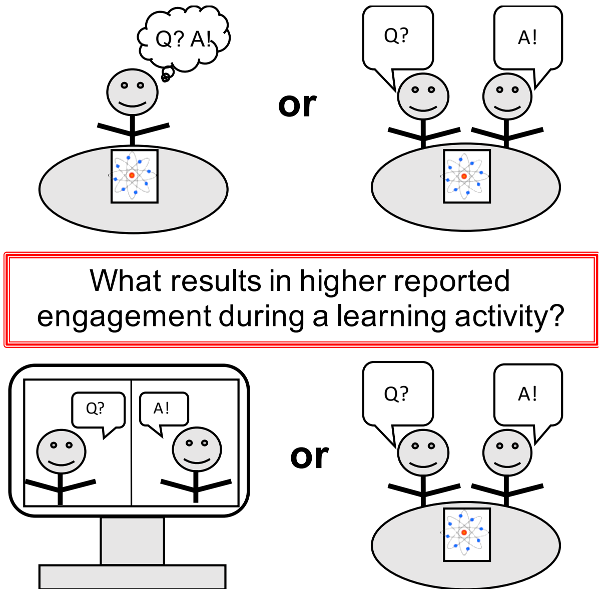

Q? A!
or
Q?
A!
What results in higher reported engagement during a learning activity?
Q?
A!
or
Q?
A!

Figure 9.1. Graphical abstract for Chapter 9.

Keywords

First-Year Undergraduate / General, Chemical Education Research, Collaborative / Cooperative Learning, Student-Centered Learning

Introduction

Incorporating active learning into a course has been found to lead to improved student learning outcomes over traditional lecture courses (Freeman et al., 2014; Rahman & Lewis, 2019). However, the extent of these benefits may vary not only across different types of active learning strategies (Rahman & Lewis, 2019), but also different courses that implement the same active learning strategy (Rahman & Lewis, 2019), as well as different student populations within the same course (Eddy & Hogan, 2014). As one component of active learning is engaging students to take part in the process of learning instead of simply being a passive observer (Freeman et al., 2014), variations in student engagement may contribute to the variations seen regarding student outcomes in active learning courses. Overall, higher student engagement in a class has been shown to lead to improved student learning outcomes (Handelsman et al., 2005; Skinner et al., 2017). However, while active learning activities provide students with the opportunity to engage with the material, whether or not a student does engage is entirely up to them (Cavanagh et al., 2016).

Engagement is considered to be a malleable state that can vary based on the environment and context (Fredricks et al., 2004; Furlong & Christenson, 2008). Therefore, there are multiple aspects which may influence a student's engagement in a particular learning activity and further exploring these could provide deeper insight into why variations are seen regarding the benefits of active learning. One possible influence on students' engagement is whether or not they worked on the activity with others. For example, in the Interactive-Constructive-Active-Passive (ICAP) framework of engagement (Chi & Wylie, 2014), only students who work collaboratively with others are capable of reaching the highest mode of engagement (i.e., interactive engagement). Additionally, other studies have found that engagement related to students' emotional state are positively influenced by student-student interactions in active collaborative learning (Molinillo et al., 2018). Therefore, students who work with others might show evidence of higher engagement than those who work alone. Another possible aspect that could influence students' engagement in active learning activities is the environment of the class, for instance, if the class is conducted remotely or in-person. One study found that students' perceptions of their engagement in class was influenced by the type of environment (online vs. in-person) and the difference in social interactions between the two (Jeffery & Bauer, 2020). As active learning activities encourage students to work with others, differences in social interactions due to the environment could affect students' engagement.

Student engagement is generally theorized to be a meta-construct composed of multiple related dimensions (Fredricks et al., 2004). These dimensions include behavioral engagement, cognitive engagement, and emotional engagement. Behavioral engagement in a specific activity can include aspects such as being prepared, doing the work, writing things down, etc. (Fredricks et al., 2004; Naibert et al., 2022; Sinatra et al., 2015) Cognitive engagement generally focuses around students' investment in their learning, such as trying to understand the material and connecting or applying what they learn (Fredricks et al., 2004; Naibert et al., 2022; Sinatra et al., 2015). It has been noted in the literature that the definitions of behavioral and cognitive engagement may be difficult to separate when focusing on students' effort and their investment in their learning (Fredricks et al., 2004; Sinatra et al., 2015). When considering engagement at the activity level, students have been found to conflate behavioral and cognitive engagement and perceive them similarly, especially in relation to asking questions, paying attention, thinking through how to solve problems, using resources, etc. (Naibert et al., 2022) Additionally, multiple studies have found support for a combined behavioral/cognitive dimension over individual behavioral and cognitive dimensions (Bathgate & Schunn, 2017; Ben-Eliyahu et al., 2018; Naibert & Barbera, 2022). Emotional engagement encompasses students' affective reactions (e.g., boredom, value, interest, etc.) to the interactions they have with the people and environment around them (Fredricks et al., 2004; Sinatra et al., 2015), which can include the learning activity itself. In addition to behavioral/cognitive and emotional engagement, other dimensions of engagement exist and may be relevant for certain environments and/or populations. For active learning activities in STEM courses, social engagement may also be an important dimension for students that work on the activity with others. Social engagement has been defined as students' interactions with others and their willingness to invest in these relationships while learning (Fredricks, Wang, et al., 2016; Wang et al., 2016). Social interactions have been found to be an important component of students' engagement when considering STEM students in higher-education (Brint et al., 2008). Additionally, ideas related to collaborative learning have been included in other engagement studies when considering active learning activities (Chi & Wylie, 2014). In higher-education chemistry courses, social engagement has recently been discovered as a dimension of student engagement in active learning activities in the classroom (Naibert & Barbera, 2022; Naibert et al., 2022).

To begin to explore students' engagement in active learning activities, the two versions of the Activity Engagement Survey (AcES) (see Naibert and Barbera (2022)) were used to evaluate students' engagement in worksheet activities in general chemistry. In the AcES development study (Naibert & Barbera, 2022), a bifactor model that included the individual dimensions of engagement (i.e., behavioral/cognitive, emotional, and social) as well as an overall engagement dimension, was found to be the most appropriate model. This model allows for overall engagement and the individual dimensions of engagement above and beyond students' overall engagement to be simultaneously assessed with the same set of items (Chen et al., 2012; Chen et al., 2006). The two versions of AcES differ with respect to the presence of social engagement items. The BC-E AcES includes 15 items related to students' overall engagement, with 10 of the items also related to students' behavioral/cognitive engagement and 5 items related to their emotional engagement. The BC-E-S AcES includes a total of 19 items to assess overall engagement; the same 15 items as the BC-E AcES in addition to 4 items related to students' social engagement. Both the BC-E and BC-E-S AcES have been previously evaluated in relation to students' engagement in general chemistry worksheet activities (Naibert & Barbera, 2022). Our investigation into students' engagement in the active learning activities incorporated in our general chemistry course included multiple research questions related to exploring differences in student engagement based on their social interactions (i.e., working alone vs. with others) and the type of environment (i.e., remote vs. in-person), as well as the association between students' engagement in the activities and their understanding of the material.

Research Questions

1) How does engagement differ between students who worked on the worksheet activities independently and students who worked on them with others?
2) How does student engagement in the learning activities differ between the remote environment and the in-person environment?
3) How does engagement relate to students' understanding of the material as assessed through subsequent exam questions?

Course Information

Data were collected from the year-long general chemistry (GC) series at Portland State University (PSU) over two academic years. During the first year (2020-2021), data were collected from two sections during each of the three terms of the academic year, which were taught by three instructors. During the second year (2021-2022), data were collected from one section of the Fall term course, which was taught by one of the same instructors as the previous year. As this study was completed during the COVID-19 pandemic, the course was taught remotely during the first year and in-person during the second year (see Table 9.1). The course included both activity and lecture days regardless of the environment. Activity days were centered around worksheets related to a single chemistry topic and included key questions, exercises, and problems for the students to respond and work through, similar in format to process-oriented guided-inquiry learning (POGIL) worksheets (Hanson et al., 2018).

Table 9.1. Course information for each term and section of general chemistry (GC).

Academic Year	Term (Course)	Environment	Instructor	Week-1 enrollment (N)
2020-2021	Fall (GCI)	Remote	A	297
			B	332
	Winter (GCII)	Remote	A	190
			B	266
	Spring (GCIII)	Remote	A	186
			C	156
2021-2022	Fall (GCI)	In-Person	A	242

Remote Environment

During the 2020-2021 academic year, all sections of the general chemistry course were taught in a synchronous remote format using the video platform software Zoom. For the first two activity days during Fall term, students were given the option to work on the worksheet independently in the main room or in a randomly assigned breakout room with

2-6 other students. For all remaining activities for Fall term and subsequent Winter and Spring terms, students were assigned to breakout rooms based on their selection of the type of breakout room they wanted to participate in for the term. The students were given the option to choose to 1) work with others with their video camera on, 2) work with others with their video camera off, 3) work independently in a breakout room with the ability to ask a learning assistant (LA) questions, or 4) work completely independently in the main room. Although students did not have to work in a breakout room, it was expected that all students completed the worksheet activity. Students that chose option 1 or 2 (i.e., to work with others), were assigned to the same breakout room for the entire term, although they could choose whether or not to join the breakout room for each activity. Each pre-assigned breakout room contained an average of 4 students. Undergraduate and graduate LAs were each assigned to rotate through two breakout rooms to answer questions and help facilitate the activity. The instructor and graduate teaching assistant (TA) facilitated the activity through check-ins with the LAs. All of the instructors over the year facilitated the worksheet activities similarly.

In-person Environment

During Fall 2021, students in the in-person environment were expected, but not required, to complete the activities during class. Students were encouraged to work with nearby students and generally worked in groups of 2 – 4, although some students opted to work alone. Undergraduate and graduate LAs, as well as the graduate TA and instructor, moved around the room during the activity to facilitate and answer questions. The instructor of the in-person class was one of the same instructors that taught the remote class the previous Fall term. Therefore, any comparisons between the remote and in-person environment only compared data collected in this instructor's course during the Fall terms (2020 and 2021).

Methods

All data collected for this study was approved by the Institutional Review Board (IRB) and Portland State University and appropriate consent was obtained as required by the IRB.

Survey Items

Engagement was assessed using the BC-E and BC-E-S versions of the Activity Engagement Survey (AcES) (Naibert & Barbera, 2022). The BC-E AcES measure includes 15 items related to students overall engagement, with 10 items of those items also related to behavioral/cognitive engagement and 5 items related to emotional engagement. The BC-E-S AcES includes 19 items related to students' overall engagement; the same 15 items from the BC-E AcES, as well as 4 items also related to social engagement. The survey asked students to respond to an initial item indicating what type of breakout room they participated in (for the remote environment) or if they worked with other students for any amount of time during the activity (for the in-person environment). Students who selected that they worked remotely in a LA question and answer breakout room (option 3) or in the main room (option 4), or who selected that they did not work with others during the in-person environment were presented with the 15 BC-E items only. Students who selected that they worked in a remote breakout room type 1 or 2 (i.e., worked with others with their camera on or their camera off), or that they worked with other students during the in-person environment were presented with the full set of 19 items from the BC-E-S AcES. All items were administered on a six-point Likert-type scale from *strongly disagree* (1) to *strongly agree* (6).

Survey Data Collection and Cleaning

Students over both years were notified of the survey through an in-class announcement on select activity days. An additional announcement that included a link to the Qualtrics survey was posted on the course's learning management site at the end of class. To reduce the potential for recall bias and encourage student responses to each specific activity, students only had 48 hours (i.e., the time before the next class period) to access and complete the survey. For the remote environment, students were given a nominal amount of extra credit for accessing the survey regardless of completion or consent. In the in-person environment, the survey was a required part of the course; however, only responses from students who consented to participate in the research study were included in the final data set. Additionally, students in the in-person environment were also asked for consent to retain their grades as part of this research study.

The responses for each activity were cleaned to remove responses from students who did not consent to the research project, incomplete responses, and any responses that included an incorrect response to a check item. Two types of check items were included in each survey; a topic-based check item and a response-based check item. The topic-based check item asked students to select the topic of the day's activity to screen for students who may not have participated in the activity. The response-based check item asked students to select a specific response (e.g., *somewhat agree*) and was used to remove responses from students who may not have read the survey items before responding. Details about the different data sets and subsets used to explore each research question are displayed in Figure 9.2. The data set from the 2020-2021 academic year was used to address RQ1 and included student responses over 9 activities completed in the remote environment. Responses from students who selected that they worked on an activity either in the main room (i.e., option 4) or the LA question and answer room (i.e., option 3) were combined into the *independent* group, whereas responses from students that selected they worked on an activity with other students in a breakout room either with their camera on or their camera off were combined into the *social* group. As the research question focused on differences in student engagement between those that worked with others and those that worked independently, the camera on (i.e., option 1) and camera off (i.e., option 2) groups were combined for this analysis. For RQ2, only data from students who worked with others during instructor A's Fall courses (2020 and 2021) were used. These datasets were used for the comparison of student engagement between the remote and in-person environments as they were collected from the same term during the academic year (i.e., Fall term) and were taught by the same instructor using the same activities. A subset of the Fall 2021 data was used for RQ3 and included students who worked with others on a single activity that covered the topic of *Solutions and Dilutions* and who consented to allow *both* their survey responses and exam grades to be collected as part of the research study.

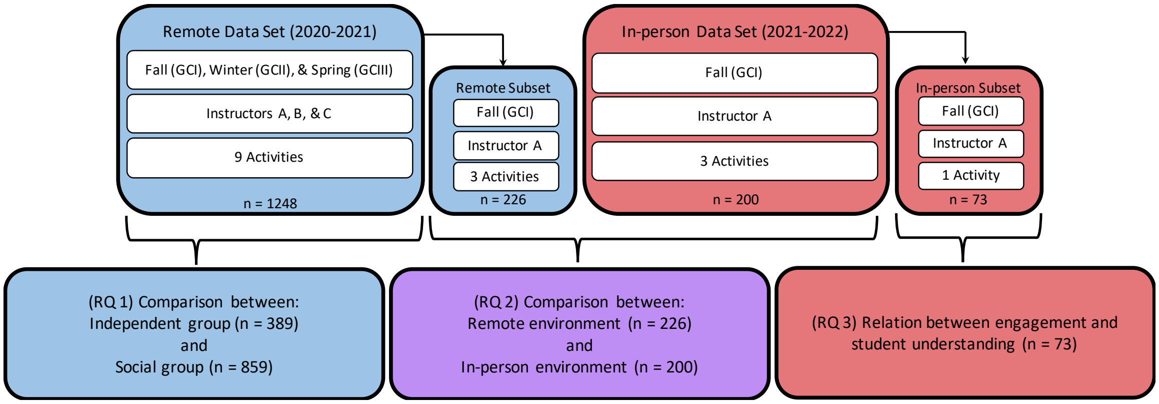

Remote Data Set (2020-2021)
Fall (GCI), Winter (GCII), & Spring (GCIII)
Instructors A, B, & C
9 Activities
n = 1248
Remote Subset
Fall (GCI)
Instructor A
3 Activities
n = 226
In-person Data Set (2021-2022)
Fall (GCI)
Instructor A
3 Activities
n = 200
In-person Subset
Fall (GCI)
Instructor A
1 Activity
n = 73
(RQ 1) Comparison between: Independent group (n = 389) and Social group (n = 859)
(RQ 2) Comparison between: Remote environment (n = 226) and In-person environment (n = 200)
(RQ 3) Relation between engagement and student understanding (n = 73)

Figure 9.2. Details about the data sets used to explore each research question (RQ). Sample sizes (n) indicate the total number of cleaned responses collected from students over the number of activities listed for each data set.

Exam grades

During Fall 2021, grades for each exam item were retained for students who consented to participate. There were a total of three exams over the term and all exam items were multiple-choice and written by the instructor of record. The first exam took place after students had completed the activity on the topic of *Solutions and Dilutions*. Four items from this exam were found to be related to the material covered on the *Solutions and Dilutions* activity by the primary researcher and were confirmed with two other researchers for alignment to the activity. These items asked students to calculate different aspects related to solutions and dilutions – such as concentration, moles, mass, and volume – and represented concepts that were emphasized throughout the entire *Solutions and Dilutions* activity. The other two surveyed activities from Fall 2021 were not included in this analysis due to a combination of low student response rates and less alignment between the topics covered by the activities and subsequent exam questions.

Internal Structure Validity

Internal structure validity evaluates whether the relation between the items and the construct match the hypothetical structure. Evidence in support of internal structure validity provides confidence that the items relate to the construct as expected (Arjoon et al., 2013; Knekta et al., 2019). The BC-E and BC-E-S AcES bifactor models relate each item to an overall engagement factor, as well as to the relevant behavioral/cognitive, emotional, or social engagement factor (see Naibert and Barbera (2022)). In a previous related study (Naibert & Barbera, 2022), evidence of internal structure validity was found for the data collected in the remote environment (2020-2021 academic year) (n = 1248) using the BC-E AcES bifactor model with a negative method factor. To provide evidence of internal structure validity for responses collected in the in-person environment (n = 200), the BC-E-S AcES bifactor model with a negative method factor was evaluated using confirmatory factor analysis (CFA). All factor analyses and structural models were tested using the lavaan package (version 0.6-5) in the statistical software R (version 3.6.2) using maximum likelihood estimation with Satorra-Bentler adjustment and robust standard errors to account for any non-normality of the data (Satorra & Bentler, 1988). Data-model fit was evaluated using suggested recommendations for good data-model fit given by Hu and Bentler (1999): CFI ≥ 0.95, TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08.

Consequential Validity

For research questions that addressed comparisons of student engagement between groups (i.e., RQ1 and RQ2), additional evidence of consequential validity was gathered through measurement invariance testing to support that the measure functioned similarly for both groups (Rocabado et al., 2020). Measurement invariance testing is completed through a series of sequential steps and an analysis of the change in data-model fit (Chen, 2007). Each step provides evidence of a “stricter” level of invariance between the groups (Rocabado et al., 2020). The first step of measurement invariance is to evaluate the unconstrained factor model for each group concurrently and is called configural invariance. The second step is metric invariance, which evaluates the factor model with constrained factor loadings across groups. Support for metric invariance provides confidence that the meaning of the factor and items are similar for both groups and allows for the next step of measurement invariance testing; however, evidence of configural and metric invariance does not yet provide supportive evidence for comparisons to be made between groups. The next step, scalar invariance, is tested through additionally constraining the item intercepts to be equal across groups. If evidence of scalar invariance is found, then comparison of the latent (i.e., factor) means of the groups using structural means modeling (SMM) (Rocabado et al., 2020), which allows for the determination of relative differences between factor means, is supported (Thompson & Green, 2013). As the factor loadings and intercepts are set to be equal across groups, the intercepts of the factors can be compared. If evidence of scalar invariance is found then the final step, conservative invariance, can be tested by additionally restraining residuals to be equal across groups. If there is sufficient evidence to support conservative invariance, then the observed scale scores of the groups can be compared (Rocabado et al., 2020). A comprehensive discussion of measurement invariance, along with sample data and analysis code (for both R and Mplus) can be found in a recent publication by Rocabado et al. (2020).

Group Comparisons

The presence of the bifactor model for the BC-E and BC-E-S AcES allow simultaneous comparisons of students' overall engagement, as well as the dimensions of engagement (e.g., behavioral/cognitive, emotional, and social) above and beyond students' overall engagement. Therefore, if evidence of scalar invariance is found, SMM can be used to compare latent means between groups using the full structural model (Rocabado et al., 2020). As SMM determines the relative difference in latent means, one of the groups was selected as the reference and the latent mean for that group was set to zero. The effect size of any differences was calculated as the absolute difference in the latent means divided by the square root of the pooled factor variances (Thompson & Green, 2013). Although similar to Cohen's d, where 0.20, 0.50, and 0.80 represent small, medium, and large effect sizes, respectively (Cohen, 1992), the guidelines for latent means differs slightly. As latent means are free from measurement error, the value of the effect size should be larger than the effect size of measured variables (Thompson & Green, 2013).

If conservative invariance is supported between two groups, observed mean scores can be compared. However, as the bifactor model associates items with both the overall factor and the individual dimensions (e.g., behavioral/cognitive, emotional, and social engagement), only the observed scores for the individual dimensions were calculated and compared. The unweighted mean scale scores of the groups were compared using one-way analysis of variance (ANOVA). ANOVAs were completed using the lessR package (version 3.9.2) in R. Cohen's f was used to determine the effect size, where 0.10, 0.25, and 0.40 are considered to represent small, medium, and large effects, respectively (Cohen, 1992).

Relation to Exam Grades

Student scores on the four exam items related to *Solutions and Dilutions* were transformed into z-scores using mean and standard deviation data from the entire class.

The relation between these scores and students' behavioral/cognitive, emotional, and social engagement on the activity was assessed using unweighted scale score means for each dimension. A multiple regression model with the behavioral/cognitive, emotional, and social engagement scale scores as predictors was evaluated using the lm function in R. The resulting effect size of the individual predictors was determined by calculating Cohen's f^2 with the semi-partial correlation coefficients of the predictors as determined using the ppcor package (version 1.1) in R. Values of Cohen's f^2 of 0.02, 0.15, and 0.35 indicate small, medium, and large effect sizes, respectively (Cohen, 1992).

Results and Discussion

How does engagement differ between students who worked on the worksheet activities independently and students who worked on them with others?

Only responses to the BC-E AcES were included when comparing students who worked with others and those that worked independently in the remote environment, as students who worked independently were not asked to respond to social engagement items. Evidence of internal structure validity for this data set was previously evaluated in Naibert and Barbera (2022) and found to show evidence of good data-model fit using recommended cutoffs from Hu and Bentler (1999) when evaluated with the BC-E bifactor model with a negative method factor (CFI = 0.974, TLI = 0.960, RMSEA = 0.053, SRMR = 0.030) (Naibert & Barbera, 2022).

Evidence of consequential validity was evaluated using measurement invariance testing and both scalar and conservative invariance were found to be supported between responses from students who worked alone (*independent group*, n = 389) and responses from students who worked with others in a breakout room (*social group*, n = 859) (see

Appendix F, Tables F.1 – F.2). As there was sufficient evidence of scalar invariance, SMM was used to compare latent means between the two groups using the bifactor model with a negative method factor. The presence of the bifactor model allowed the comparison of behavioral/cognitive, emotional, and overall engagement to be conducted simultaneously (Table 9.2). As comparing latent means through SMM only evaluates relative mean differences, the independent group (i.e., students who worked alone) was chosen as the reference, meaning that this group's latent means for each factor were set to zero. Therefore, positive latent mean differences indicate that the latent mean of the social group (i.e., students who worked with others) was higher than the latent mean of the independent group, while negative differences indicate the latent mean of the social group was lower compared to the independent group. When latent means between the groups were compared, it was found that students' emotional engagement was not significantly different, while students' behavioral/cognitive and overall engagement were significantly different with small effect sizes. Differences in the latent means for behavioral/cognitive and overall engagement were positive, which indicates that the social group had a higher overall engagement and was more behaviorally/cognitively engaged than the independent group.

Table 9.2. Latent mean difference between the independent group (n = 389) and the social group (n = 859).

Factor	**Latent mean difference[a]** **(ref = independent group)**	**Effect size**
Behavioral/cognitive	**0.079**	0.47
Emotional	-0.071	0.08
Overall Engagement	**0.215**	0.33

[a]Bold values indicate the difference was statistically significant at $p < 0.05$.

As conservative invariance was also supported between the two groups, unweighted observed mean scores were also compared. For this analysis, only the mean scores from the individual dimensions were evaluated since the same items were related to both the individual dimensions and the overall engagement factor. While the SMM analysis could account for this overlap due to the structure of the bifactor model, ANOVA cannot. Results from one-way ANOVAs of each individual dimension indicated that the social group scored higher on behavioral/cognitive and emotional engagement than the independent group, with medium and small effect sizes, respectively (Table 9.3).

Table 9.3. Observed unweighted mean scores and ANOVA results.

Factor	**Independent mean (SD) (n = 389)**	**Social mean (SD) (n = 859)**	**Mean difference[a] (social – independent)**	**Effect size (Cohen's f)**
Behavioral/cognitive	4.70 (0.68)	5.02 (0.66)	**0.32**	0.22
Emotional	4.21 (0.90)	4.43 (0.90)	**0.22**	0.11

[a]Bold values indicate the difference was statistically significant at $p < 0.05$.

While both SMM and observed score comparisons indicated that the social group had statistically higher behavioral/cognitive engagement than the independent group, results surrounding emotional engagement were conflicting. Although SMM found that the emotional engagement latent means of the groups were statistically equivalent, when observed scores were compared, a statistically significant difference was found. This variation could be due to the structure of the bifactor model and the nature of observed scale score comparisons. Whereas SMM, which relies on the bifactor model, only compares behavioral/cognitive and emotional engagement latent means that exist *above and beyond* what is captured by the overall engagement factor, observed scores do not parcel out variance due to the students' overall engagement. This could result in differences in outcomes when observed scores are used since these scores contain variance that would be accounted for by the overall engagement factor when the bifactor model is used with SMM.

The results from the latent means analysis suggest that students who chose to work with their peers in a breakout room had higher overall engagement than students who chose to work by themselves. Additionally, behavioral/cognitive engagement above and beyond general overall engagement was also found to be higher for students who worked with others. Higher levels of cognitive engagement in active learning activities has been conceptualized in other frameworks as being related to productive collaboration between students while working on a task or activity (Chi & Wylie, 2014). Therefore, increased interactions between the students who worked with others in a breakout room may have contributed to the higher behavioral/cognitive engagement found for these students. Since students were given the option to work with others on these activities, one possibility for the difference in overall engagement between the groups could be due to student buy-in to the activity. Student buy-in to active learning activities has been found to be positively related to student engagement (Cavanagh et al., 2016; Wang et al., 2021). It may be possible that students who sought the opportunity to work with others on the activity had a higher buy-in to the activity in general. Although no difference in emotional engagement above and beyond general overall engagement was found through SMM, there was a difference in emotional engagement between the groups when comparing overall scale scores (i.e., when overall engagement was not parceled out). This suggests that the difference between emotional engagement scale scores was likely due to differences in overall engagement between the two groups and that when this difference was taken into account, the two groups had similar emotional engagement.

How does student engagement in learning activities differ between the remote environment and the in-person environment?

As the BC-E-S AcES survey was given in a new environment (i.e., in-person) compared to our prior study (Naibert & Barbera, 2022), CFA was first used to evaluate the internal structure validity in this environment using the bifactor model with a negative method factor. Overall, there was evidence of good data-model fit (see Appendix F Table F.3), which provided support that even in the new environment, the items and the constructs were related in the expected manner. However, before engagement between the remote and in-person environments could be compared, evidence of measurement invariance had to be collected to provide confidence that the model functioned similarly between the groups. Although data from both environments (i.e., instructor A's Fall term courses) individually showed evidence of good data-model fit, only configural invariance could be supported between the two environments (see Appendix F Table F.4). The lack of support for metric (and higher levels) of invariance indicate that the meaning of the factors as related to the given items may not have been the same between the environments (Rocabado et al., 2020). Therefore, comparisons between students' engagement in the remote and in-person environments could not be supported as representing true differences between the groups and this research question currently remains unanswered. However, as evidence of internal structure was found for responses from the in-person environment, relations between engagement and students' understanding in the in-person environment could still be assessed.

How does engagement relate to students' understanding of the material as assessed through subsequent exam questions?

As internal structure validity was supported for the in-person environment using the BC-E-S bifactor model, the relation between engagement in the activity and subsequent understanding of the material was assessed using students' scores on relevant exam items. As the sample size for the *Solutions and Dilutions* activity was small (n = 73), multiple linear regression was chosen over structural equation modeling. The relation between students' engagement on an activity and their z-scores on related exam questions was evaluated using the unweighted mean scores of the students' behavioral/cognitive, emotional, and social engagement as predictors (Table 9.4). Overall model fit for the regression was found to be $R^2 = 0.10$ ($F_{3, 69} = 2.45$, $p = 0.07$). Behavioral/cognitive engagement was found to be a statistically significant predictor with a small to medium effect size and positively related to students' scores on relevant exam items. Emotional engagement was negatively related to students' scores, while social engagement was positively related, although these relations were not found to be statistically significant. All three predictors only accounted for about 10% of the variance in exam item scores.

Table 9.4. Regression analysis of engagement and exam item z-scores.

Predictor	**Exam items z-score**	
	Standardized regression coefficient[a] (SE)	**Effect size (Cohen's f^2)**
Behavioral/cognitive engagement	**0.315** (0.137)	0.08
Emotional engagement	-0.189 (0.126)	0.03
Social engagement	0.024 (0.137)	<0.01

[a]Bold values indicate result was statistically significant at $p < 0.05$.

Students' behavioral/cognitive engagement in the activity was the only dimension found to be a significant predictor of students' exam item scores. As emotional engagement was not found to be a significant predictor of students' grades, this suggests that students' affective reactions to the environment did not influence their performance on related exam material. Additionally, the nonsignificant relation between social engagement and exam item scores suggests that simply engaging with a group did not necessarily lead to improved grades if the students were not behaviorally/cognitively engaged. Other studies of active learning activities have found that students' who show higher levels of cognitive engagement have higher posttest grades (Chi & Wylie, 2014) and studies that have considered STEM courses as a whole have found that increased behavioral and emotional engagement can lead to higher overall course grades (Skinner et al., 2017). However, not all studies have found emotional engagement to predict student grades. One study of high school students found that students who were emotionally disengaged while still behaviorally and cognitively engaged had a similar GPA to students who were highly behaviorally, cognitively, and emotionally engaged (Wang & Peck, 2013) and another study of middle school students found that emotional engagement did not predict students' science course grades (Wang et al., 2016). However, even if emotional engagement is not a significant predictor of student outcomes related to exams or grades, it may be important when considering other student outcomes. For example, the study of high school students also discovered that students who were emotionally disengaged had lower educational aspirations, lower college enrollment rate, and higher depression than those that were engaged in all three dimensions (Wang & Peck, 2013), while the study of middle school students found that the emotional engagement dimension was the strongest predictor of students' future STEM career aspirations (Wang et al., 2016).

Conclusions

The first research question, *How does engagement differ between students who worked on the worksheet activities independently and students who worked on them with others?* was explored through comparing students' overall, behavioral/cognitive, and emotional engagement between students who worked with others and those that worked independently in the remote environment. Support for scalar and conservative invariance were found, which allowed for comparisons to be made both in a latent framework using SMM and using observed scores for the two dimensions of behavioral/cognitive and emotional engagement. Overall, students who chose to work with others on the activity had significantly higher overall and behavioral/cognitive engagement than those who chose to work by themselves when assessed with latent means. This suggests that working with others was related to a higher overall engagement and behavioral/cognitive engagement than students who chose to work on the activity independently; however, the two groups showed no difference in emotional engagement when latent means were compared. As SMM allowed for differences between the groups to be explored using the bifactor model, the nonsignificant difference in emotional engagement was only comparing emotional engagement that was not accounted for by the students' overall engagement.

When unweighted observed scale score means for behavioral/cognitive and emotional engagement were also compared between the two groups, students who worked with others were found to have higher behavioral/cognitive and emotional engagement than students who worked by themselves. The difference in results between the latent mean differences and the observed mean differences is likely due to variance from overall engagement still remaining in observed score means. Thus, although observed score results suggest that students who worked with others were more emotionally engaged, the results from SMM suggest that emotional engagement above and beyond the significant differences found in overall engagement were nonsignificant between the groups. This suggests that caution should be taken when comparing groups using observed mean scores, especially when a bifactor model is found to be the most appropriate model for the construct of interest.

As data were collected in both the remote environment and in-person environment, we sought to answer the second research question, *How does student engagement in learning activities differ between the remote environment and the in-person environment?* Before differences in engagement between the remote and in-person environments could be assessed in our study, steps were taken to ensure there was validity evidence to support the comparison between environments. While evidence of internal structure validity was found for responses collected in both environments using the BC-E-S AcES (Naibert & Barbera, 2022), which indicated that the data collected in each environment could be assessed independently with the BC-E-S bifactor model, evidence of consequential validity (i.e., measurement invariance) between the two environments was not found. Invariance beyond the configural level was not supported between the two groups, indicating that there was not enough evidence to justify that either latent or observed mean comparisons would represent true differences between the two groups (Rocabado et al., 2020). Therefore, no comparisons were made between responses collected in these two environments during this study. However, we encourage future studies that find evidence of scalar or conservative invariance in their data to explore comparisons between these two environments, as other studies completed at the beginning of the COVID-19 pandemic have found that shifting to a remote environment influenced students' perceptions of their engagement and peer interactions (Jeffery & Bauer, 2020; Wu & Teets, 2021).

The third research question, *How does engagement relate to students' understanding of the material as assessed through subsequent exam questions?* was explored through multiple regression using unweighted observed scale scores for behavioral/cognitive, emotional, and social engagement. Students' engagement in an activity was assessed in relation to their scores on four related questions on the subsequent exam. Overall, students who had higher behavioral/cognitive engagement on the activity were found to score higher on the exam questions, although emotional and social engagement in the activity were not found to be significantly related to students' scores. While this finding suggests that emotional engagement may not affect students' exam grades, other studies have found relations between students' emotional engagement and other student outcomes, such as career aspirations (Wang et al., 2016; Wang & Peck, 2013). Additionally, social engagement was not found to be a significant predictor of student exam scores, although behavioral/cognitive engagement was found to be significant. This suggests that simply engaging with a group without being behaviorally/cognitively engaged may not be sufficient to increase student understanding.

Limitations

This study was completed at a single institution and with specific worksheet-based activities. Therefore, it is unknown if the same results will be found in a different population and/or with different types of active learning activities. Additionally, the comparison between students who chose to work with others and those who chose to work independently was completed in a remote environment during the COVID-19 pandemic and facilitated through breakout rooms. It is unknown to what extent additional obligations, responsibilities, and stresses inherent to the pandemic influenced students' engagement in these activities. Although a difference was found in students' overall and behavioral/cognitive engagement between the two groups, this result cannot be automatically extended to in-person activities during non-pandemic times.

Although evidence of internal structure validity was found for data collected in both the remote and in-person environments, measurement invariance between the two environments was not supported. Therefore, differences in engagement between the remote and in-person environments were not explored in this study as these comparisons could not be justified. While the lack of measurement invariance may have been due to differences in how students in the two environments responded to the items, the relatively low sample size in both environments may have also contributed. Additionally, although active learning environments have been found to differentially influence performance outcomes of some demographic groups (Eddy & Hogan, 2014), the low sample size in this study restricted comparisons of student engagement between different demographic groups (e.g., gender identity, race/ethnicity, etc.).

Due to sample size limitations, the relation between engagement and exam grades was completed through the use of regression using observed scale score means. This meant that the assessment of behavioral/cognitive, emotional, and social engagement included variance from overall engagement in the observed scale scores and did not allow for the evaluation of how overall engagement was related to exam scores. With a larger sample size, the relation between student engagement and grades could be assessed using structural equation modeling, which would allow the relations to be explored within a latent means framework using the bifactor model. Additionally, students' grades on four related items on a subsequent exam were used as a proxy for students' understanding of the material. As this study was completed with minimal interruptions to the current class structure, no pretest was given. Future studies into the relation between engagement and student understanding are encouraged to assess students' change in understanding using a pretest/posttest design with items that assess different levels of understanding.

Implications for Research

The different results found when comparing emotional engagement between students who worked with others and those that worked independently when using latent means versus observed means suggest that caution should be taken when comparing observed means between two groups. Although there was a significant difference found when observed means were used for the comparison, the latent means comparison indicated that this was likely due to variance from students' overall engagement. Thus, researchers should be aware of this potential effect on results when using observed scale scores, especially when considering constructs modeled with a bifactor model.

There are many possibilities for future studies of student engagement with the AcES. Although a difference in overall and behavioral/cognitive engagement was found between students who worked with others and those that worked independently, this comparison was only completed in a remote environment during the COVID-19 pandemic. As many classes implement in-person active learning activities, it would be informative to explore whether this same difference exists when activities are conducted in an in-person environment. Additionally, if a larger sample size is collected and evidence of consequential validity is found to support comparisons within a latent means framework (Rocabado et al., 2020), differences of student engagement related to the type of environment (e.g., remote vs. in-person), type of activity, level of facilitation, or demographic groups (e.g., gender identity, race/ethnicity, etc.) could be explored. As remote learning has gained momentum in the last few years due to the COVID-19 pandemic, exploring differences in student engagement of active learning activities when conducted in a remote environment versus an in-person environment could provide useful information about how engagement may be affected when learning is moved online and provide valuable feedback to instructors who teach in remote environments.

Evaluating results from a larger sample size would also allow the relation between engagement in the activities and student understanding to be explored using structured equation modeling. This would allow for the simultaneous evaluation of students' overall engagement and the dimensions of engagement in a latent means framework with the bifactor model, as well as allow the influence of overall engagement to be parceled out from the individual dimensions of engagement. Additionally, a pretest/posttest design of previously evaluated content exam items would allow the change in student understanding due to their engagement in the activity to be more closely assessed.

Social engagement was not found to be a predictor of student performance in this study; however, the type and quality of group interactions and dynamics were not explored. Therefore, although students may have been socially engaged, they may not have been behaviorally/cognitively engaged in the activity. Although students are thought to be more cognitively engaged when they are interacting with other students to create new material and build off each other’s ideas (Chi & Wylie, 2014), one study found that student discussion more frequently centers around exchanging information instead of exchanging reasoning (Paine & Knight, 2020). Therefore, it would be informative for further research to explore how group dynamics and social interactions influence students’ overall, behavioral/cognitive, emotional, and social engagement.

Implications for Practice

Although this study focused on a specific population and type of active learning activity, the results related to the importance of working with others on students’ engagement may extend to other populations and activities. In a related qualitative study, students mentioned that although working with others was not necessary to be behaviorally engaged, working with a group helped them because they could share answers and find mistakes (Naibert et al., 2022). When considering engagement of STEM students, a study conducted by Brint et al. (2008) found that STEM generally encourages a “culture of engagement” that includes collaborative effort. Although group work is not a requirement of active learning, the results found in this study suggest that students who chose to work with others were more engaged, both overall and behaviorally/cognitively. Therefore, instructors may want to encourage more student buy-in to working with others during active learning activities.

This study found that students’ behavioral/cognitive engagement was significantly related to students’ grades on related exam items, while social and emotional engagement were not found to be significantly related. Therefore, instructors may want to provide opportunities and scaffolding to increase students’ behavioral/cognitive engagement in activities. Although some activities may be designed to encourage students to engage at certain levels of behavioral/cognitive engagement, that does not guarantee that students necessarily engage at the expected level throughout the activity (El-Mansy et al., 2022). Additionally, although emotional and social engagement were not found to be significantly related to students' grades on related exam items in this study, these aspects may still be important to students' learning and should not be overlooked when facilitating active learning in the classroom. For example, some studies have found students' social relationships, such as sense of belonging, to positively influence students' overall grade in a course (Edwards et al., 2021) and other studies focused on students' engagement in STEM courses as a whole found emotional engagement to be positively related to student performance outcomes (Skinner et al., 2017). Additionally, higher emotional engagement in a course has also been found to be related to higher educational and career aspirations (Wang et al., 2016; Wang & Peck, 2013).

Associated Content

Supporting Information

The Supporting Information is available in Appendix F and includes measurement invariance.

Author Information

Corresponding Author

*E-mail: jack.barbera@pdx.edu

Chapter 10: Conclusions and Implications

Conclusions

This dissertation research sought to explore the effects of active learning through the lens of student engagement and related variables. This was completed through three separate but related projects. The first project (see Chapters 4 and 5) was centered around investigating students' self-efficacy in flipped general chemistry courses at multiple institutions, as well as their interactions with and perceptions of pre-class materials used in these courses. The second project (see Chapter 6) focused on exploring student perceptions of two active learning environments incorporated into a principles of biology course – Deliberative Democracy activities and clicker question days. Finally, the third project (see Chapters 7 – 9) focused around evaluating engagement in worksheet-based active learning activities included in a general chemistry course through the development of a survey measure of behavioral/cognitive, emotional, and social engagement. The findings and results from each of these projects are addressed below with respect to the research questions posed for each in Chapter 1.

Project I: Students' Self-Efficacy and Interactions with Pre-Class Materials in Flipped Courses

The first phase of this project involved investigating students' interactions with and perceptions of pre-class materials in flipped courses (RQ 1.1). Data were collected through the use of focus groups and survey responses from five different institutions. The second phase focused around evaluating students' self-efficacy in three of these institutions through the use of survey responses and structured means modeling (RQs 1.2 – 1.3).

RQ 1.1: What are students' interactions with and perceptions of required pre-class materials in flipped courses?

Overall, a majority of students at all five institutions watched at least some of the assigned pre-class videos; however, there was some variation in how many and when students usually watched them. Students in courses where the predominate student behavior during the face-to-face (F2F) time consisted of solving worksheet problems in a group setting (i.e., Courses Four and Five) were more likely to report watching all the videos and doing so before coming to class compared to courses where students predominately responded to instructor-posed questions. This viewing behavior may be the result of students in Courses Four and Five finding more value in the videos, as students likely needed insights to the content covered in them to complete the worksheet problems during the F2F time. In the other courses, instructors conducted some level of material review guided by whole-class questioning, which likely reduced the need for students to fully engage with the video content in order to be successful during the F2F time. These results suggest a potential trend between the structure of F2F class time and students' video viewing habits, although further research is required to explore this possible connection as this study was not designed to determine the source of these differences. While there was a difference in how many and when students watched the videos, many students from all five courses reported re-watching the videos and engaging with them at higher cognitive levels.

Students' perceptions of the videos were fairly consistent across all five courses. Many students responded that *control of learning* (i.e., watch at their own pace, watch wherever and whenever) and *perceived usefulness* (e.g., easy to understand, help reinforce material, etc.) of the videos were helpful to their learning. Students also were relatively consistent in their responses to why the videos were not helpful, with the most common category being that the videos *do not meet learning expectations*, especially in regards to not having the opportunity to ask questions if they arose.

RQ 1.2: How does student self-efficacy change over the term in flipped courses?

Students' pre to post academic self-efficacy and chemistry self-efficacy were evaluated using structured means modeling (SMM) while controlling for students' time management and concentration. Results indicated that students' academic self-efficacy decreased over the term at all three institutions, while students' chemistry self-efficacy increased. The differences between the two could be due to the specificity of each measure. While the items used to assess chemistry self-efficacy were task-based, items for the academic self-efficacy scale were focused around general academics. As more specific self-efficacy measures have been found to be better predictors of performance compared to more general self-efficacy measures (Choi, 2005), it is likely that the structure of the students' chemistry course had a larger impact on their chemistry self-efficacy than their academic self-efficacy. Although an increase in chemistry self-efficacy was seen at all three institutions, the change at the Western institution was not found to be significant. As this institution was a second-term chemistry course, compared to first-term chemistry courses at the other two institutions, it is possible that students had more chemistry experience to inform their pre-term self-efficacy scores, which could have contributed to the nonsignificant increase.

RQ 1.3: How does student self-efficacy compare across flipped courses at different institutions?

Students' post-term chemistry self-efficacy scores were compared between institutions using SMM while controlling for their pre-term chemistry self-efficacy, time management, and concentration. Overall, students at the Southeastern institution were found to have significantly higher post-term chemistry self-efficacy when compared to students at the other two institutions (i.e., Western and Northwestern); however, there was no significant difference between the Western and Northwestern institutions. Although determining the source of these differences was not a design of the study, one possible explanation could be the F2F environment used at each of the institutions. When only considering the two institutions that were first-term general chemistry courses (i.e., Southeastern and Northwestern), the Southeastern institution primarily relied on instructor presentation of material followed by whole-class questioning during F2F time, whereas the Northwestern institution included primarily groupwork. The more guided and structured class time of the Southeastern institution may have contributed to the higher chemistry self-efficacy students reported, as other studies have found that including well-structured tasks can result in higher reports of student self-efficacy (Lodewyk & Winne, 2005).

Project II: Students' Perceptions of Different Active Learning Environments

The second project focused on evaluating a survey measure of students' perceptions of active learning activities with the student population and active learning environments at Portland State University (PSU). The original measure, the Assessing Student Engagement in Class Tool (ASPECT) was developed by Wiggins et al. (2017) to assess students' perceptions of their 'personal effort', 'value of group activity', and 'instructor contribution' of short- and long-activity days. Data collected with the measure in a general chemistry course and principles of biology course at PSU were used to answer the following research questions.

RQ 2.1: How well does an existing measure of student perceptions of active learning activities function in different active learning environments?

Data collected with the original ASPECT items in a first-term general chemistry course were evaluated using confirmatory factor analysis (CFA) with the structure proposed by Wiggins et al. (2017). The structure included three factors of 'personal effort', 'value of group activity', and 'instructor contribution'. Results from the CFA suggested model modifications related to item-item error correlations between three item pairs. While two of these pairs only included items from the 'value of group activity' factor, one pair included an item from the 'value of group activity' factor and an item from the 'personal effort' factor. Additionally, upon further inspection of the suggested modifications, it was noted that all items were related to some aspect of group function. These results suggested that there may have been another source of variance related to group function that was not accounted for by the original factor structure. However, adding an additional group function factor to the original structure was not ideal, as only two items would remain to assess 'personal effort'. This led to modifications to the measure, as addressed in RQ 2.2.

RQ 2.2: What modifications can be made to an existing measure in order to measure student perceptions of these environments?

Modifications to the original ASPECT survey were made with respect to the learning environment and the results from RQ 2.1 prior to the survey being distributed in a principles of biology course that incorporated both clicker question days and Deliberative Democracy activities. Since the course included two different types of active learning environments, minor wording modifications were made to the items such that most of the items could be administered in both environments. However, whenever an item referenced groupwork, the wording for the clicker question day focused around *discussions* with other students, whereas the wording for the Deliberative Democracy activity focused around *working* with other students. Additionally, as learning assistants (LAs) were present on Deliberative Democracy activity days, parallel LA-focused instructor contribution items were created for that environment. Modifications were also created to account for the possible group function factor discovered during RQ 2.1. This included eight new personal effort items to increase the items related to that factor and four additional “other-focused” items to increase the number of items that might potentially be related to group function. Overall, the modifications resulted in two survey measures, one for each environment. The mASPECT-C, which was designed for the clicker question day, included a total of 31 items. The mASPECT-DD, which was designed for the Deliberative Democracy activity day, included 35 items, where four of the items were the additional LA-focused instructor contribution items. The final items for both mASPECT-C and mASPECT-DD can be found in Tables 6.1 and 6.2 in Chapter 6.

RQ 2.3: How well does this modified measure function in different active learning environments?

Data collected with the modified ASPECT versions, mASPECT-DD and mASPECT-C, were evaluated using response process interviews and exploratory factor analysis (EFA). EFA does not require an *a priori* structure and allows items to load on multiple factors such that different factor structures can be explored. When data collected with mASPECT-DD was assessed, three factors related to 'personal effort', 'value of environment', and 'classroom support' were discovered. Data evaluated from mASPECT-C similarly showed evidence of factors related to 'personal effort', 'value of environment', and 'classroom support', in addition to a factor related to 'social influence'. However, the specific items related to each of the factors were different between the two environments; therefore, results between the two could not be compared. However, scale scores for each factor were evaluated within each environment and showed that students perceived their personal effort, value of the environment, and classroom support to be relatively high for both environments. Additionally, they also perceived social influence positively in the clicker day environment. Overall, although different factor structures were discovered for the mASPECT-DD and mASPECT-C, both of the modified measures were found to function well in each respective environment.

Project III: Students' Engagement in Worksheet Activities

This project focused on students' engagement in worksheet activities completed in a general chemistry course. A measure of engagement originally developed by Wang et al. (2016) for middle- and high-school science students was modified using data collected through interviews and surveys. Data collected with the modified measure, the Activity Engagement Survey (AcES), was used to evaluate differences between groups, as well as the association between engagement and student understanding.

RQ 3.1: How do students perceive engagement in worksheet activities?

Student interviews were analyzed to investigate how students perceived engagement with respect to the worksheet activities. Overall, student responses indicated that students generally perceived behavioral, cognitive, and emotional engagement similarly to literature definitions (Fredricks et al., 2004). However, results also indicated that students may have perceived behavioral and cognitive engagement to be similar when considering worksheet activities. Additionally, ideas related to social interactions were discovered when students were asked to describe behavioral, cognitive, and emotional engagement, suggesting that social engagement may also be an important component when considering engagement at the activity-level. Although there were some differences in specific codes found in the in-person and remote environments; overall, similar responses and trends were discovered in both.

RQ 3.2: What modifications can be made to an existing survey measure in order to measure student engagement in these activities?

Results from both response process interviews and factor analyses were used to inform modifications to the original survey developed by Wang et al. (2016). Many modifications consisted of minor wording changes to remove double-barreled items or to clarify the focus of the item. Some items were removed due to not being relevant to the higher education student population or the specific environment. Additionally, two new behaviorally-focused items were added to the behavioral engagement scale based on the definition of engagement by Fredricks et al. (2004). The final 19-item survey consisted of 5 items related to each behavioral, cognitive, and emotional engagement dimension, and 4 social engagement items.

RQ 3.3: How well does a modified measure of engagement function in this environment and student population?

The final survey was analyzed using CFA to explore the most appropriate factor structure and provide evidence that the survey functioned well in our environment and with our student population. Both the complete 19-item survey and a 15-item survey, without the social factor, were analyzed. In both cases, evidence was found to support combining the behavioral and cognitive factors into a single behavioral/cognitive factor due to high factor correlations between the two factors. Support was found for a correlated model with the individual engagement dimensions (e.g., behavioral/cognitive, emotional, and social); however, a bifactor model, where an overall engagement factor is included and related to each of the items in addition to the individual dimensions of engagement, was found to improve the data-model fit. Finally, the addition of a negative method factor, which was related to each negatively-worded item, accounted for response bias due simply to the wording of the items (i.e., positive or negative). Overall, the most appropriate models discovered for the 19-item and 15-item surveys were the BC-E-S and BC-E bifactor models with a negative method factor, respectively (see Figures 8.6 and 8.7 in Chapter 8). The evidence towards a combined behavioral/cognitive factor aligns with the possible conflation found between these two factors discovered during RQ 3.1. Additionally, other surveys that have focused on assessing engagement at the activity-level have had similar findings (Bathgate & Schunn, 2017; Ben-Eliyahu et al., 2018), suggesting that students may perceive these two dimensions of engagement to be very similar when asked about their task-level engagement. The presence of the bifactor model provides evidence of an overarching overall engagement factor, in addition to the separate dimensions of engagement above and beyond what is captured with overall engagement. This model allows for each of these factors to be measured simultaneously and accounts for variance due to overall engagement.

RQ 3.4: How does engagement in these activities differ across groups?

The goal of this research question was to use the AcES to compare engagement across groups for two comparisons of interest. The first comparison was between students that worked on an activity by themselves and those that worked on an activity with others. This was completed using the BC-E AcES survey, as students that worked by themselves did not respond to any social engagement items. As support for both scalar and conservative invariance was found for this comparison, differences between the groups were explored using both structural means modeling (SMM) with latent means and ANOVA with observed mean scale scores. Overall, students who worked with others were found to have higher overall and behavioral/cognitive engagement. When comparing latent means, both groups were statistically similar for emotional engagement above and beyond overall engagement. When the observed score differences were assessed, which did not parcel out overall engagement variance from the emotional engagement score, students who worked with others had significantly higher emotional engagement. This indicates that caution should be taken when comparing observed scale scores between groups when using the AcES or other measures modeled by a bifactor model.

The second comparison of interest was to explore engagement differences with the BC-E-S AcES when activities were completed in the remote environment versus the in-person environment. However, although internal structure validity was supported for each environment separately, consequential validity was not supported for this comparison due to the lack of evidence from measurement invariance. Therefore, this comparison could not be completed in this study. Although a lack of support for measurement invariance may have been due to differences in how the students in the two environments responded to the items, the low sample size of the groups may have also contributed.

RQ 3.5: How does engagement in these activities relate to students' understanding of the material?

Multiple linear regression between students' behavioral/cognitive, emotional, and social engagement scale scores and their grades on relevant exam items was used to explore the association between students' engagement in the activity and their understanding of the material covered on the activity. Overall, it was found that students' who had higher behavioral/cognitive engagement scored higher on relevant exam items, while students' emotional and social engagement in the activity were not statistically significantly related to exam item grades.

Implications for Research

Each project has its own implications for researchers and possibilities for future studies. For Project I, the variability in students' interactions with and perceptions of the pre-class materials in flipped courses could benefit from future studies designed to determine possible sources of the differences. Although there appeared to be a trend between the face-to-face structure of the course and how students interacted with the videos, further research using student focus groups, interviews, or tracking data could provide more support for this possible connection. With regards to self-efficacy, the differing pre to post trends found for academic and chemistry self-efficacy indicate that the specificity of self-efficacy items should be taken into consideration to ensure that the focus of the items align with the goals of the study. Additionally, although comparisons of students' self-efficacy between different institutions was completed, the sample size of the individual institutions did not allow for comparisons between demographic groups at the same institution. Therefore, future studies that have a sufficient sample size and find evidence of consequential validity are encouraged to explore comparisons of students' self-efficacy between different demographic groups within the same flipped course.

Results from Project II provided support for exploring response process and internal structure validity when data are collected with a measure in a new population and environment. Future studies that decide to use the mASPECT or ASPECT in another course, population, and/or active learning environment should collect additional validity evidence to support the use of the measure in their environment before outcomes from data collected with either measure is evaluated. Additionally, open-ended student interviews about their perceptions of an environment could provide information about the relevance of current mASPECT and ASPECT items and possibly direct the creation of additional items and/or factors to better represent students' perceptions. For example, although 'social influence' was not discovered as a factor for the structure of the mASPECT in the Deliberative Democracy environment (mASPECT-DD), it is possible that there were no items that adequately represented that factor in that environment. If evidence of validity is found to support the use of the same measure in multiple types of active learning environments, then the data collected could be used to investigate possible differences in how students perceive those environments. Additionally, if evidence of consequential validity is supported, comparisons between different demographic groups could be made.

Project III provided more insight into how students perceive engagement when considering worksheet activities and the Activities Engagement Survey (AcES) was developed as a possible tool for assessing student engagement in these types of activities. Future work with the AcES should first collect evidence of validity and reliability to support the use of AcES in a new environment and/or population. Although both qualitative and quantitative support was found for a combined behavioral/cognitive engagement factor, additional research should be completed to explore why this conflation exists and the possibility of it being due to the specificity of the measure (e.g., task-focused versus class-focused). The presence of a social engagement factor in both qualitative and quantitative results indicates that social engagement should be considered when evaluating engagement of activities that incorporate groupwork in general chemistry.

The best fitting model for the AcES was found to be a bifactor model, which has many potential uses in research, especially in relation to assessing students' overall engagement along with their individual dimensions of engagement above and beyond what is captured with the overall engagement factor. Results from the comparison of engagement between students who worked alone and those that worked with others were found to be different when using latent means with the bifactor model versus using observed scale score means. This suggests that comparisons of individual dimensions using observed scale scores may be influenced by additional variance from students' overall engagement, while this variance is parceled out when comparing latent means using the bifactor model. Therefore, when sample size allows, the use of latent mean comparisons with the AcES may provide more detailed information about students' engagement compared to using observed scale scores. Assuming evidence of consequential validity is found with the AcES, there are multiple other types of comparisons which could provide beneficial information about students' engagement for researchers and instructors. For example, although validity evidence did not support the comparison in this study, possible differences in students' engagement between an in-person environment and remote environment could provide information about how students' engagement may be influenced by remote activities. Additionally, differences between different demographic groups (e.g., gender identity, race, etc.) could be compared. Another use of the bifactor model would be to use structural equation modeling to explore the relation between students' engagement in the activities and their understanding of the material.

Implications for Practice

Overall, all three projects in this dissertation suggest that instructors should be cognizant of how different class environments may influence students' interactions, perceptions, and engagement. In investigating flipped courses, a positive trend between the amount of class time spent working in groups and the number of students who reported watching all of the videos before class was discovered. This trend may have been due to students finding more value in the content covered in the videos when the expectation during the F2F time was to complete worksheet problems with their peers. This suggests that if an instructor's expectation for a flipped class is that students watch the videos before coming to class, they may want to consider including more groupwork during class time. Although students in flipped courses generally found that the ability to watch pre-class videos in their own time and on their own pace to be helpful to their learning, one of the most common reasons students found the videos to not be helpful was the inability to ask questions. Therefore, instructors who use pre-class videos may want to consider how best to address this aspect of learning in their course; for example, by providing opportunities for students to ask questions in an online space or during class. When considering students' perceptions of active learning environments using the mASPECT, two different factor structures were discovered. This suggests that instructors should be cautious about using a survey measure developed in a different environment to directly compare their active learning environments or classes. However, student responses to individual items of the mASPECT and other similar measures could provide instructors with formative feedback about their active learning environments in order to inform modifications and changes to the implementation and facilitation of these environments. Results from investigating students' engagement in worksheet activities with the AcES measure found that students who chose to work with others on the activities were more engaged overall and behaviorally/cognitively than those who chose to work by themselves and that students who had higher behavioral/cognitive engagement were found to score higher on exam items related to the material covered in an activity. Therefore, instructors may want to consider informing students about the reasons for and benefits of groupwork, as well as structuring active learning activities to encourage more student buy-in to the activity and collaborative problem-solving.

References

Abeysekera, L., & Dawson, P. (2014). Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research. *Higher Education Research & Development, 34*(1), 1-14. https://doi.org/10.1080/07294360.2014.934336

Aceti, V. (2017). Perceptions of the effects of clicker technology on student learning and engagement: a study of freshmen Chemistry students. *Research in Learning Technology, 20*(2). https://doi.org/10.3402/rlt.v20i0.16150

Ahlfeldt, S., Mehta, S., & Sellnow, T. (2005). Measurement and analysis of student engagement in university classes where varying levels of PBL methods of instruction are in use. *Higher Education Research & Development, 24*(1), 5-20. https://doi.org/10.1080/0729436052000318541

Amaral, K. E., Shank, J. D., Shibley, I. A., & Shibley, L. R. (2013). Web-Enhanced General Chemistry Increases Student Completion Rates, Success, and Satisfaction. *Journal of Chemical Education, 90*(3), 296-302. https://doi.org/10.1021/ed200580q

American Phsychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Anderman, E. M., & Dawson, H. (2011). Learning and Motivation. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of Research on Learning and Instruction* (pp. 219-241). Routledge.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology, 44*(5), 427-445. https://doi.org/10.1016/j.jsp.2006.04.002

Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *Journal of Chemical Education, 90*(5), 536-545. https://doi.org/10.1021/ed3002013

Baeten, M., Dochy, F., & Struyven, K. (2013). The Effects of Different Learning Environments on Students' Motivation for Learning and Their Achievement. *British Journal of Educational Psychology, 83*(3), 484-501. https://doi.org/10.1111/j.2044-8279.2012.02076.x

Bancroft, S. F., Fowler, S. R., Jalaeian, M., & Patterson, K. (2019). Leveling the Field: Flipped Instruction as a Tool for Promoting Equity in General Chemistry. *Journal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.9b00381

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. Freeman.

Bandura, A. (2001). Social Cognitive Theory: An Agentic Perspective. *Annual review of psychology, 52*(1), 1-26. https://doi.org/10.1146/annurev.psych.52.1.1

Bandura, A. (2006). Toward a Psychology of Human Agency. *Perspectives on psychological science, 1*(2), 164-180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Barkley, E. F., & Major, C. H. (2020). *Student Engagement Techniques: A Handbook for College Faculty*. John Wiley & Sons.

Bathgate, M., & Schunn, C. (2017). The psychological characteristics of experiences that influence science motivation and content knowledge. *International Journal of Science Education, 39*(17), 2402-2432. https://doi.org/10.1080/09500693.2017.1386807

Bauer, C. F. (2008). Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts. *Journal of Chemical Education, 85*(10), 1440. https://doi.org/10.1021/ed085p1440

Ben-Eliyahu, A., Moore, D., Dorph, R., & Schunn, C. D. (2018). Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. *Contemporary Educational Psychology, 53*, 87-105. https://doi.org/10.1016/j.cedpsych.2018.01.002

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin, 112*(3), 400. https://doi.org/10.1037/0033-2909.112.3.400

Bergmann, J., & Sams, A. (2012). *Flip Your Classroom: Reach every student in every class every day*. International Society for Technology in Education.

Bernard, P., Broś, P., & Migdał-Mikuli, A. (2017). Influence of blended learning on outcomes of students attending a general chemistry course: summary of a five-year-long study. *Chemistry Education Research and Practice, 18*(4), 682-690. https://doi.org/10.1039/c7rp00040e

Blackburn, R. A. R. (2017). Write My Next Lecture: Prelecture Problem Classes and In-Lecture Discussion To Assist Case-Study Teaching of Synthesis. *Journal of Chemical Education, 95*(1), 104-107. https://doi.org/10.1021/acs.jchemed.7b00528

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Front Public Health, 6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bodner, G. M. (1992). Why changing the curriculum may not be enough. *Journal of Chemical Education, 69*(3). https://doi.org/10.1021/ed069p186

Bokosmaty, R., Bridgeman, A., & Muir, M. (2019). Using a Partially Flipped Learning Model To Teach First Year Undergraduate Chemistry. *Journal of Chemical Education, 96*(4), 629-639. https://doi.org/10.1021/acs.jchemed.8b00414

Boz, Y., Yerdelen-Damar, S., Aydemir, N., & Aydemir, M. (2016). Investigating the Relationships among Students' Self-efficacy Beliefs, Their Perceptions of Classroom Learning Environment, Gender, and Chemistry Achievement through Structural Equation Modeling. *Research in Science & Technological Education, 34*(3), 307-324. https://doi.org/10.1080/02635143.2016.1174931

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2020). Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research, 21*(1), 37-47. https://doi.org/10.1002/capr.12360

Braun, V., Clarke, V., Hayfield, N., & Terry, G. (2019). Thematic Analysis. In P. Liamputtong (Ed.), *Handbook of Research Methods in Health Social Sciences* (pp. 843-860). Springer Singapore. https://doi.org/10.1007/978-981-10-5251-4_103

Brazeal, K. R., & Couch, B. A. (2017). Student Buy-In Toward Formative Assessments: The Influence of Student Factors and Importance for Course Success. *Journal of Microbiology and Biology Education, 18*(1). https://doi.org/10.1128/jmbe.v18i1.1235

Brewe, E. (2008). Modeling Theory Applied: Modeling Instruction in Introductory Physics. *American Journal of Physics, 76*(12), 1155-1160. https://doi.org/10.1119/1.2983148

Brint, S., Cantwell, A. M., & Hanneman, R. A. (2008). The Two Cultures of Undergraduate Academic Engagement. *Research in Higher Education, 49*(5), 383-402. https://doi.org/10.1007/s11162-008-9090-y

Brooks, B. J., & Koretsky, M. D. (2011). The Influence of Group Discussion on Students' Responses and Confidence during Peer Instruction. *Journal of Chemical Education, 88*(11), 1477-1484. https://doi.org/10.1021/ed101066x

Brophy, J. (2008). Developing Students' Appreciation for What Is Taught in School. *Educational Psychologist, 43*(3), 132-141. https://doi.org/10.1080/00461520701756511

Brophy, J. (2010). *Motivating Students to Learn* (Third ed.). Routledge.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Brown, T. L., Brazeal, K. R., & Couch, B. A. (2017). First-Year and Non-First-Year Student Expectations Regarding In-Class and Out-of-Class Learning Activities in Introductory Biology. *Journal of Microbiology and Biology Education, 18*(1). https://doi.org/10.1128/jmbe.v18i1.1241

Bunce, D. M., Flens, E. A., & Neiles, K. Y. (2010). How Long Can Students Pay Attention in Class? A Study of Student Attention Decline Using Clickers. *Journal of Chemical Education, 87*(12), 1438-1443. https://doi.org/10.1021/ed100409p

Bunce, D. M., Komperda, R., Schroeder, M. J., Dillner, D. K., Lin, S., Teichert, M. A., & Hartman, J. R. (2017). Differential Use of Study Approaches by Students of Different Achievement Levels. *Journal of Chemical Education, 94*(10), 1415-1424. https://doi.org/10.1021/acs.jchemed.7b00202

Caldwell, J. E. (2007). Clickers in the large classroom: current research and best-practice tips. *CBE—Life Sciences Education, 6*(1), 9-20. https://doi.org/10.1187/cbe.06-12-0205

Campbell, C. M., & Cabrera, A. F. (2011). How Sound Is NSSE?: Investigating the Psychometric Properties of NSSE at a Public, Research-Extensive Institution. *The Review of Higher Education, 35*(1), 77-103. https://doi.org/10.1353/rhe.2011.0035

Canelas, D. A., Hill, J. L., & Novicki, A. (2017). Cooperative learning in organic chemistry increases student assessment of learning gains in key transferable skills. *Chemistry Education Research and Practice, 18*(3), 441-456. https://doi.org/10.1039/c7rp00014f

Cantwell, R. H., & Andrews, B. (2002). Cognitive and Psychological Factors Underlying Secondary School Students' Feelings Towards Group Work. *Educational Psychology, 22*(1), 75-91. https://doi.org/10.1080/01443410120101260

Casselman, M. D., Atit, K., Henbest, G., Guregyan, C., Mortezaei, K., & Eichler, J. F. (2019). Dissecting the Flipped Classroom: Using a Randomized Controlled Trial Experiment to Determine When Student Learning Occurs. *Journal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.9b00767

Cavanagh, A. J., Aragon, O. R., Chen, X., Couch, A., Durham, F., Bobrownicki, A., Hanauer, D. I., & Graham, M. J. (2016). Student Buy-In to Active Learning in a College Science Course. *CBE—Life Sciences Education, 15*(4). https://doi.org/10.1187/cbe.16-07-0212

Cavanagh, A. J., Chen, X., Bathgate, M., Frederick, J., Hanauer, D. I., & Graham, M. J. (2018). Trust, Growth Mindset, and Student Commitment to Active Learning in a College Science Course. *CBE—Life Sciences Education, 17*(1). https://doi.org/10.1187/cbe.17-06-0107

Chang, Y., & Brickman, P. (2018). When Group Work Doesn't Work: Insights from Students. *CBE—Life Sciences Education, 17*(3), ar42. https://doi.org/10.1187/cbe.17-09-0199

Chapman, K. J., & van Auken, S. (2016). Creating Positive Group Project Experiences: An Examination of the Role of the Instructor on Students' Perceptions of Group Projects. *Journal of Marketing Education, 23*(2), 117-127. https://doi.org/10.1177/0273475301232005

Chase, A., Pakhira, D., & Stains, M. (2013). Implementing Process-Oriented, Guided-Inquiry Learning for the First Time: Adaptations and Short-Term Impacts on Students' Attitude and Performance. *Journal of Chemical Education, 90*(4), 409-416. https://doi.org/10.1021/ed300181t

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural equation modeling: a multidisciplinary journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: a comparison of the bifactor model to other approaches. *Journal of Personality, 80*(1), 219-251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A Comparison of Bifactor and Second-Order Models of Quality of Life. *Multivariate Behavioral Research, 41*(2), 189-225. https://doi.org/10.1207/s15327906mbr4102_5

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist, 49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Choi, N. (2005). Self-Efficacy and Self-Concept as Predictors of College Students' Academic Performance. *Psychology in the Schools, 42*(2), 197-205. https://doi.org/10.1002/pits.20048

Christiansen, M. A. (2014). Inverted Teaching: Applying a New Pedagogy to a University Organic Chemistry Class. *Journal of Chemical Education, 91*(11), 1845-1850. https://doi.org/10.1021/ed400530z

Christiansen, M. A., Lambert, A. M., Nadelson, L. S., Dupree, K. M., & Kingsford, T. A. (2016). In-Class Versus At-Home Quizzes: Which is Better? A Flipped Learning Study in a Two-Site Synchronously Broadcast Organic Chemistry Course. *Journal of Chemical Education, 94*(2), 157-163. https://doi.org/10.1021/acs.jchemed.6b00370

Christiansen, M. A., Nadelson, L., Etchberger, L., Cuch, M., Kingsford, T. A., & Woodward, L. O. (2017). Flipped Learning in Synchronously-Delivered, Geographically-Dispersed General Chemistry Classrooms. *Journal of Chemical Education, 94*(5), 662-667. https://doi.org/10.1021/acs.jchemed.6b00763

Clark, R. C., Nguyen, F., & Sweller, J. (2005). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. Pfeiffer.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Dalgety, J., & Coll, R. K. (2006). Exploring First-Year Science Students' Chemistry Self-Efficacy. *International Journal of Science and Mathematics Education, 4*(1), 97-116. https://doi.org/10.1007/s10763-005-1080-3

Deci, E. L., & Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry, 11*(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01

Deci, E. L., & Ryan, R. M. (2012). Self-Determination Theory. In *Handbook of theories of social psychology* (pp. 416-436). Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n21

DiBenedetto, M. K., & Bembenutty, H. (2013). Within the Pipeline: Self-Regulated Learning, Self-Efficacy, and Socialization among College Students in Science Courses. *Learning and Individual Differences, 23*, 218-224. https://doi.org/10.1016/j.lindif.2012.09.015

Donnelly, J., & Hernández, F. E. (2018). Fusing a reversed and informal learning scheme and space: student perceptions of active learning in physical chemistry. *Chemistry Education Research and Practice, 19*(2), 520-532. https://doi.org/10.1039/c7rp00186j

Ealy, J. B. (2013). Development and Implementation of a First-Semester Hybrid Organic Chemistry Course: Yielding Advantages for Educators and Students. *Journal of Chemical Education, 90*(3), 303-307. https://doi.org/10.1021/ed200858p

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: how and for whom does increasing course structure work? *CBE—Life Sciences Education, 13*(3), 453-468. https://doi.org/10.1187/cbe.14-03-0050

Edwards, J. D., Barthelemy, R. S., & Frey, R. F. (2021). Relationship between Course-Level Social Belonging (Sense of Belonging and Belonging Uncertainty) and Academic Performance in General Chemistry 1. *Journal of Chemical Education, 99*(1), 71-82. https://doi.org/10.1021/acs.jchemed.1c00405

Eichler, J. F., & Peeples, J. (2016). Flipped classroom modules for large enrollment general chemistry courses: a low barrier approach to increase active learning and improve student grades. *Chemistry Education Research and Practice, 17*(1), 197-208. https://doi.org/10.1039/c5rp00159e

El-Mansy, S. Y., Barbera, J., & Hartig, A. J. (2022). Investigating small-group cognitive engagement in general chemistry learning activities using qualitative content analysis and the ICAP framework. *Chemistry Education Research and Practice*. https://doi.org/10.1039/d1rp00276g

Fautch, J. M. (2015). The flipped classroom for teaching organic chemistry in small classes: is it effective? *Chemistry Education Research and Practice, 16*(1), 179-186. https://doi.org/10.1039/c4rp00230j

Fenollar, P., Román, S., & Cuestas, P. J. (2007). University Students' Academic Performance: An Integrative Conceptual Framework and Empirical Analysis. *British Journal of Educational Psychology, 77*(4), 873-891. https://doi.org/10.1348/000709907X189118

Ferrell, B., & Barbera, J. (2015). Analysis of Students' Self-Efficacy, Interest, and Effort Beliefs in General Chemistry. *Chemistry Education Research and Practice, 16*(2), 318-337. https://doi.org/10.1039/C4RP00152D

Ferrell, B., Phillips, M. M., & Barbera, J. (2016). Connecting Achievement Motivation to Performance in General Chemistry. *Chemistry Education Research and Practice, 17*(4), 1054-1066. https://doi.org/10.1039/C6RP00148C

Fitzgerald, N., & Li, L. (2015). Using Presentation Software To Flip an Undergraduate Analytical Chemistry Course. *Journal of Chemical Education, 92*(9), 1559-1563. https://doi.org/10.1021/ed500667c

Flynn, A. B. (2015). Structure and evaluation of flipped chemistry courses: organic & spectroscopy, large and small, first to third year, English and French. *Chemistry Education Research and Practice, 16*(2), 198-211. https://doi.org/10.1039/c4rp00224e

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research, 74*(1), 59-109. https://doi.org/10.3102/00346543074001059

Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction, 43*, 1-4. https://doi.org/10.1016/j.learninstruc.2016.02.002

Fredricks, J. A., & McColskey, W. (2012). The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 763-782). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_37

Fredricks, J. A., Wang, M.-T., Schall Linn, J., Hofkens, T. L., Sung, H., Parr, A., & Allerton, J. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning and Instruction, 43*, 5-15. https://doi.org/10.1016/j.learninstruc.2016.01.009

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America, 111*(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education, 10*(2), 175-186. https://doi.org/10.1187/cbe.10-08-0105

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., Dirks, C., & Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE—Life Sciences Education, 6*(2), 132-139. https://doi.org/10.1187/cbe.06-09-0194

Furlong, M. J., & Christenson, S. L. (2008). Engaging students at school and with learning: A relevant construct forall students. *Psychology in the Schools, 45*(5), 365-368. https://doi.org/10.1002/pits.20302

Galyon, C. E., Blondin, C. A., Yaw, J. S., Nalls, M. L., & Williams, R. L. (2012). The Relationship of Academic Self-Efficacy to Class Participation and Exam Performance. *Social Psychology of Education, 15*(2), 233-249. https://doi.org/10.1007/s11218-011-9175-x

Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From Gatekeeping to Engagement: A Multicontextual, Mixed Method Study of Student Academic Engagement in Introductory STEM Courses. *Research in Higher Education, 53*(2), 229-261. https://doi.org/10.1007/s11162-011-9247-y

Gibbons, R. E., Laga, E. E., Leon, J., Villafañe, S. M., Stains, M., Murphy, K., & Raker, J. R. (2017). Chasm Crossed? Clicker Use in Postsecondary Chemistry Education. *Journal of Chemical Education, 94*(5), 549-557. https://doi.org/10.1021/acs.jchemed.6b00799

Graham, K. J., Bohn-Gettler, C. M., & Raigoza, A. F. (2019). Metacognitive Training in Chemistry Tutor Sessions Increases First Year Students' Self-Efficacy. *Journal of Chemical Education, 96*(8), 1539-1547. https://doi.org/10.1021/acs.jchemed.9b00170

Gregorius, R. M. (2017). Performance of underprepared students in traditional versus animation-based flipped-classroom settings. *Chemistry Education Research and Practice, 18*(4), 841-848. https://doi.org/10.1039/c7rp00130d

Gu, H., Wen, Z., & Fan, X. (2015). The impact of wording effect on reliability and validity of the Core Self-Evaluation Scale (CSES): A bi-factor perspective. *Personality and Individual Differences, 83*, 142-147. https://doi.org/10.1016/j.paid.2015.04.006

Hancock, G. R. (2004). Experimental, Quasi-Experimental, and Nonexperimental Design and Analysis with Latent Variables. In *The SAGE handbook of quantitative methodology for the social sciences* (pp. 317-334). SAGE.

Hancock, G. R., Mueller, R. O., & Stapleton, L. M. (2010). Factor Analysis: Exploratory and Confirmatory. In *The reviewer's guide to quantitative methods in the social sciences*. Routledge.

Handelsman, J., Miller, S., & Pfund, C. (2007). *Scientific teaching*. Macmillan.

Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A Measure of College Student Course Engagement. *Journal of Educational Research, 98*(3), 184-192. https://doi.org/10.3200/joer.98.3.184-192

Hanrahan, M. (1998). The Effect of Learning Environment Factors on Students' Motivation and Learning. *International Journal of Science Education, 20*(6), 737-753. https://doi.org/10.1080/0950069980200609

Hanson, D. M. (2007). *Foundations of chemistry: Applying POGIL principles*. Pacific Crest.

Hanson, D. M., Goodwin, J., & Phillips, M. (2018). *Foundations of Chemistry: Applying POGIL Principles* (5 ed.). Pacific Crest Publishing.

Harris, A. H., & Cox, M. F. (2003). Developing an Observation System to Capture Instructional Differences in Engineering Classrooms. *Journal of Engineering Education, 92*(4), 329-336. https://doi.org/10.1002/j.2168-9830.2003.tb00777.x

He, W., Holton, A., Farkas, G., & Warschauer, M. (2016). The effects of flipped instruction on out-of-class study time, exam performance, and student perceptions. *Learning and Instruction, 45*, 61-71. https://doi.org/10.1016/j.learninstruc.2016.07.001

He, W., Holton, A., Gu, H., Warschauer, M., & Farkas, G. (2019). Differentiated Impact of Flipped Instruction: When Would Flipped Instruction Work or Falter? *International Journal of Teaching and Learning in Higher Education, 31*, 32-49.

Hermida, R. (2015). The problem of allowing correlated errors in structural equation modeling: concerns and considerations. *Computational Methods in Social Sciences, 3*(1), 5-17.

Hibbard, L., Sung, S., & Wells, B. (2016). Examining the Effectiveness of a Semi-Self-Paced Flipped Learning Format in a College General Chemistry Sequence. *Journal of Chemical Education, 93*(1), 24-30. https://doi.org/10.1021/acs.jchemed.5b00592

Hill, D. J., Williams, O. F., Mizzy, D. P., Triumph, T. F., Brennan, C. R., Mason, D. C., & Lawrence, D. S. (2019). Introduction to Laboratory Safety for Graduate Students: An Active-Learning Endeavor. *Journal of Chemical Education, 96*(4), 652-659. https://doi.org/10.1021/acs.jchemed.8b00774

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational research methods, 1*(1), 104-121. https://doi.org/10.1177/109442819800100106

Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55. https://doi.org/10.1081/10705519909540118

Huff, J. D., & Nietfeld, J. L. (2009). Using Strategy Instruction and Confidence Judgments to Improve Metacognitive Monitoring. *Metacognition and Learning, 4*(2), 161-176. https://doi.org/10.1007/s11409-009-9042-8

Jeffery, K. A., & Bauer, C. F. (2020). Students' Responses to Emergency Remote Online Teaching Reveal Critical Factors for All Teaching. *Journal of Chemical Education, 97*(9), 2472-2485. https://doi.org/10.1021/acs.jchemed.0c00736

Jin, Y. (2020). A Note on the Cutoff Values of Alternative Fit Indices to Evaluate Measurement Invariance for ESEM Models. *International Journal of Behavioral Development, 44*(2), 166-174. https://doi.org/https://doi.org/10.1177/0165025419866911

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education, 38*(5), 758-773. https://doi.org/10.1080/03075079.2011.598505

Kavanagh, L., Reidsema, C., McCredden, J., & Smith, N. (2017). In C. Reidsema, L. Kavanagh, R. Hadgraft, & N. Smith (Eds.), *The Flipped Classroom: Practice and Practices in Higher Education* (pp. 15-35). Springer Nature.

King, A. (1993). From Sage on the Stage to Guide on the Side. *College teaching, 41*(1), 30-35. https://doi.org/https://doi.org/10.1080/87567555.1993.9926781

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Knekta, E., Runyon, C., & Eddy, S. (2019). One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research. *CBE—Life Sciences Education, 18*(1), rm1. https://doi.org/10.1187/cbe.18-04-0064

Komarraju, M., & Nadler, D. (2013). Self-Efficacy and Academic Achievement: Why do Implicit Beliefs, Goals, and Effort Regulation Matter? *Learning and Individual Differences, 25*, 67-72. https://doi.org/10.1016/j.lindif.2013.01.005

Komperda, R., Barbera, J., Shortlidge, E. E., & Shusterman, G. P. (2018). Connecting chemistry to community with deliberative democracy. In *Citizens first! Democracy, social responsibility and chemistry* (pp. 81-98). ACS Publications.

Komperda, R., Hosbein, K. N., & Barbera, J. (2018). Evaluation of the Influence of Wording Changes and Course Type on Motivation Instrument Functioning in Chemistry. *Chemistry Education Research and Practice, 19*(1), 184-198. https://doi.org/10.1039/C7RP00181A

Komperda, R., Hosbein, K. N., Phillips, M. M., & Barbera, J. (2020). Investigation of Evidence for the Internal Structure of a Modified Science Motivation Questionnaire II (mSMQ II): a Failed Attempt to Improve Instrument Functioning across Course, Subject, and Wording Variants. *Chemistry Education Research and Practice*. https://doi.org/10.1039/D0RP00029A

Komperda, R., Pentecost, T. C., & Barbera, J. (2018). Moving beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for Quantitative Chemistry Education Research. *Journal of Chemical Education, 95*(9), 1477-1491. https://doi.org/10.1021/acs.jchemed.8b00220

Krieger, J. (1990). Winds of Revolution Sweep through Science Education. *Chemical and Engineering News, 68*(24), 27-43. https://doi.org/10.1021/cen-v068n024.p027

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Lane, E. S., & Harris, S. E. (2015). A New Tool for Measuring Student Behavioral Engagement in Large University Classes. *Journal of College Science Teaching, 44*(6), 83-91. http://www.jstor.org/stable/43632000

Lavy, S. (2016). Who Benefits from Group Work in Higher Education? An Attachment Theory Perspective. *Higher Education, 73*(2), 175-187. https://doi.org/10.1007/s10734-016-0006-z

Lawson, A. E., Banks, D. L., & Logvin, M. (2007). Self-Efficacy, Reasoning Ability, and Achievement in College Biology. *Journal of Research in Science Teaching, 44*(5), 706-724. https://doi.org/10.1002/tea.20172

Lenczewski, M. S. (2016). Scaffolded Semi-Flipped General Chemistry Designed To Support Rural Students' Learning. *Journal of Chemical Education, 93*(12), 1999-2003. https://doi.org/10.1021/acs.jchemed.6b00320

Liu, Y., Ferrell, B., Barbera, J., & Lewis, J. E. (2017). Development and Evaluation of a Chemistry-Specific Version of the Academic Motivation Scale (AMS-Chemistry). *Chemistry Education Research and Practice, 18*(1), 191-213. https://doi.org/10.1039/c6rp00200e

Liu, Y., Raker, J. R., & Lewis, J. E. (2018). Evaluating student motivation in organic chemistry courses: moving from a lecture-based to a flipped approach with peer-led team learning. *Chemistry Education Research and Practice, 19*(1), 251-264. https://doi.org/10.1039/c7rp00153c

Livingstone, D., & Lynch, K. (2002). Group Project Work and Student-centred Active Learning: two different experiences. *Journal of Geography in Higher Education, 26*(2), 217-237. https://doi.org/10.1080/03098260220144748

Lodewyk, K. R., & Winne, P. H. (2005). Relations Among the Structure of Learning Tasks, Achievement, and Changes in Self-Efficacy in Secondary Students. *Journal of Educational Psychology, 97*(1), 3-12. https://doi.org/10.1037/0022-0663.97.1.3

Lund, T. J., Pilarz, M., Velasco, J. B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE—Life Sciences Education, 14*(2). https://doi.org/10.1187/cbe.14-10-0168

Maldonado, P., & Leontyev, A. (2018). *Using a meta-analysis to assess the effectiveness of flipped learning in chemistry* Abstracts of Papers, New Orleans, LA, March 18-22, 2018.

Maroco, J., Maroco, A. L., Campos, J. A. D. B., & Fredricks, J. A. (2016). University student's engagement: development of the University Student Engagement Inventory (USEI). *Psicologia: Reflexão e Crítica, 29*(1). https://doi.org/10.1186/s41155-016-0042-8

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural equation modeling: a multidisciplinary journal, 11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. Pearson Higher Ed.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates, Inc.

McNeal, K. S., Zhong, M., Soltis, N. A., Doukopoulos, L., Johnson, E. T., Courtney, S., Alwan, A., & Porch, M. (2020). Biosensors Show Promise as a Measure of Student Engagement in a Large Introductory Biology Course. *CBE—Life Sciences Education, 19*(4), ar50. https://doi.org/10.1187/cbe.19-08-0158

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment, 100*(1), 43-52. https://doi.org/10.1080/00223891.2017.1281286

Merriam, S. B., Caffarella, R. S., & Baumgartner, L. M. (2007). *Learning in Adulthood: A Comprehensive Guide*. Jossey-Bass.

Michael, J., & Modell, H. I. (2003). *Active Learning in Secondary and College Science Classrooms: A Working Model for Helping the Learner to Learn*. Routledge.

Molinillo, S., Aguilar-Illescas, R., Anaya-Sánchez, R., & Vallespín-Arán, M. (2018). Exploring the impacts of interactions, social presence and emotional engagement on active collaborative learning in a social web-based environment. *Computers & Education, 123*, 41-52. https://doi.org/10.1016/j.compedu.2018.04.012

Moog, R. S., & Spencer, J. N. (2008). *POGIL: An overview*. ACS Publications.

Mooring, S. R., Mitchell, C. E., & Burrows, N. L. (2016). Evaluation of a Flipped, Large-Enrollment Organic Chemistry Course on Student Attitude and Achievement. *Journal of Chemical Education, 93*(12), 1972-1983. https://doi.org/10.1021/acs.jchemed.6b00367

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*(2), 358-368. https://doi.org/10.1037/0022-0663.91.2.358

Morgan, G., Hodge, K., Wells, K., & Watkins, M. (2015). Are Fit Indices Biased in Favor of Bi-Factor Models in Cognitive Ability Research?: A Comparison of Fit in Correlated Factors, Higher-Order, and Bi-Factor Models via Monte Carlo Simulations. *Journal of Intelligence, 3*(1), 2-20. https://doi.org/10.3390/jintelligence3010002

Mueller, R. O., & Hancock, G. R. (2008). Best Practices in Structural Equation Modeling. In *Best Practices in Quantitative Methods* (pp. 488-508). https://doi.org/10.4135/9781412995627.d38

Murphy, P. K., Wilkinson, I. A., & Soter, A. O. (2011). Instruction Based on Discussion. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of Research on Learning and Instruction*. Routledge.

Naibert, N., & Barbera, J. (2022). Development and Evaluation of a Survey to Measure Student Engagement at the Activity Level in General Chemistry. *Journal of Chemical Education, 99*(3), 1410-1419. https://doi.org/10.1021/acs.jchemed.1c01145

Naibert, N., Geye, E., Phillips, M. M., & Barbera, J. (2020). Multicourse Comparative Study of the Core Aspects for Flipped Learning: Investigating In-Class Structure and Student Use of Video Resources. *Journal of Chemical Education, 97*(10), 3490-3505. https://doi.org/10.1021/acs.jchemed.0c00399

Naibert, N., Vaughan, E. B., Brevick, K., & Barbera, J. (2022). Exploring Student Perceptions of Behavioral, Cognitive, and Emotional Engagement at the Activity Level in General Chemistry. *Journal of Chemical Education, 99*(3), 1358-1367. https://doi.org/10.1021/acs.jchemed.1c01051

National Research Council. (1999). *How People Learn: Bridging Research and Practice*. National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards: A Guide for Teaching and Learning*. National Academy Press.

National Research Council. (2012). *Discipline-based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*. National Academy Press.

National Survey of Student Engagement. (2018). *NSSE Conceptual Framework (2018).* http://nsse.indiana.edu/html/psychometric_portfolio.cfm

Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education, 7*(2), 133-144. https://doi.org/10.1177/1477878509104318

O'Connor, C., & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods, 19*. https://doi.org/10.1177/1609406919899220

Oakley, B., Felder, R. M., Brent, R., & Elhajj, I. (2004). Turning Student Groups into Effective Teams. *Journal of Student Centered Learning, 2*(1), 9-34.

Paine, A. R., & Knight, J. K. (2020). Student Behaviors and Interactions Influence Group Discussions in an Introductory Biology Lab Setting. *CBE Life Sci Educ, 19*(4), ar58. https://doi.org/10.1187/cbe.20-03-0054

Pajares, F., & Urdan, T. (2006). Self-Efficacy Beliefs of Adolescents. In *Adolescence and Education* (Vol. 5). Information Age Publishing.

Parsons, A. F. (2019). Flipping Introductory Retrosynthetic Analysis: An Exemplar Course To Get the Ball Rolling. *Journal of Chemical Education, 96*(4), 819-822. https://doi.org/10.1021/acs.jchemed.8b00946

Pearson, R. J. (2017). Tailoring Clicker Technology to Problem-Based Learning: What's the Best Approach? *Journal of Chemical Education, 94*(12), 1866-1872. https://doi.org/10.1021/acs.jchemed.7b00270

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In *Handbook of research on student engagement* (pp. 259-282). Springer.

Pienta, N. J. (2019). Introductory Chemistry Using the "Flipped" Environment: An Update. *Journal of Chemical Education, 96*(6), 1053-1054. https://doi.org/10.1021/acs.jchemed.9b00458

Pilcher, S. C. (2017). Hybrid Course Design: A Different Type of Polymer Blend. *Journal of Chemical Education, 94*(11), 1696-1701. https://doi.org/10.1021/acs.jchemed.6b00809

Pintrich, P. R. (2003). Motivation and Classroom Learning. *Handbook of psychology*, 103-122. https://doi.org/https://doi.org/10.1002/0471264385.wei0706

Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational psychology review, 16*(4), 385-407. https://doi.org/1040-726X/04/1200-0385/0

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond Cold Conceptual Change: The Role of Motivational Beliefs and Classroom Contextual Factors in the Process of Conceptual Change. *Review of Educational Research, 63*(2), 167-199. https://doi.org/10.3102/00346543063002167

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*.

President's Council of Advisors on Sciences and Technology (PCAST). (2012). *Engaged to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*.

Pressley, M., & Ghatala, E. S. (1990). Self-Regulated Learning: Monitoring Learning From Text. *Educational Psychologist, 25*(1), 19-33. https://doi.org/10.1207/s15326985ep2501_3

Rahman, T., & Lewis, S. E. (2019). Evaluating the evidence base for evidence-based instructional practices in chemistry through meta-analysis. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21610

Rain-Griffith, L., Sheghwei, S., Shusterman, G. P., Barbera, J., & Shortlidge, E. E. (2020). Deliberative Democracy: Investigating the Longitudinal Impacts of Democratic Activities in Introductory Biology Courses. *The American Biology Teacher, 82*(7), 453-462. https://doi.org/https://doi.org/10.1525/abt.2020.82.7.453

Raker, J. R., Dood, A. J., Srinivasan, S., & Murphy, K. L. (2020). Pedagogies of engagement use in postsecondary chemistry education in the United States: results from a national survey. *Chemistry Education Research and Practice*. https://doi.org/10.1039/d0rp00125b

Ramnarain, U., & Ramaila, S. (2018). The Relationship between Chemistry Self-Efficacy of South African First Year University Students and their Academic Performance. *Chemistry Education Research and Practice, 19*(1), 60-67. https://doi.org/10.1039/C7RP00110J

Ranga, J. S. (2017). Customized Videos on a YouTube Channel: A Beyond the Classroom Teaching and Learning Platform for General Chemistry Courses. *Journal of Chemical Education, 94*(7), 867-872. https://doi.org/10.1021/acs.jchemed.6b00774

Rau, M. A., Kennedy, K., Oxtoby, L., Bollom, M., & Moore, J. W. (2017). Unpacking "Active Learning": A Combination of Flipped Classroom and Collaboration Support Is More Effective but Collaboration Support Alone Is Not. *Journal of Chemical Education, 94*(10), 1406-1414. https://doi.org/10.1021/acs.jchemed.7b00240

Reardon, R. F., Traverse, M. A., Feakes, D. A., Gibbs, K. A., & Rohde, R. E. (2010). Discovering the Determinants of Chemistry Course Perceptions in Undergraduate Students. *Journal of Chemical Education, 87*(6), 643-646. https://doi.org/10.1021/ed100198r

Reeve, J., & Tseng, C.-M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology, 36*(4), 257-267. https://doi.org/10.1016/j.cedpsych.2011.05.002

Reid, S. A. (2016). A flipped classroom redesign in general chemistry. *Chemistry Education Research and Practice, 17*(4), 914-922. https://doi.org/10.1039/c6rp00129g

Rein, K. S., & Brookes, D. T. (2015). Student Response to a Partial Inversion of an Organic Chemistry Course for Non-Chemistry Majors. *Journal of Chemical Education, 92*(5), 797-802. https://doi.org/10.1021/ed500537b

Reisner, B. A., Pate, C. L., Kinkaid, M. M., Paunovic, D. M., Pratt, J. M., Stewart, J. L., Raker, J. R., Bentley, A. K., Lin, S., & Smith, S. R. (2020). I've Been Given COPUS (Classroom Observation Protocol for Undergraduate STEM) Data on My Chemistry Class... Now What? *Journal of Chemical Education, 97*(4), 1181-1189. https://doi.org/10.1021/acs.jchemed.9b01066

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological Correlates of University Students' Academic Performance. *Psychological Bulletin, 138*(2), 353-387. https://doi.org/10.1037/a0026838

Ridley, D. S., Schutz, P. A., Glanz, R. S., & Weinstein, C. E. (1992). Self-regulated learning: The interactive influence of metacognitive awareness and goal-setting. *The Journal of Experimental Education, 60*(4), 293-306. https://doi.org/10.1080/00220973.1992.9943867

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do Psychosocial and Study Skill Factors Predict College Outcomes? *Psychological Bulletin, 130*(2), 261-288. https://doi.org/10.1037/0033-2909.130.2.261

Robert, J., Lewis, S. E., Oueini, R., & Mapugay, A. (2016). Coordinated Implementation and Evaluation of Flipped Classes and Peer-Led Team Learning in General Chemistry. *Journal of Chemical Education, 93*(12), 1993-1998. https://doi.org/10.1021/acs.jchemed.6b00395

Rocabado, G. A., Komperda, R., Lewis, J. E., & Barbera, J. (2020). Addressing diversity and inclusion through group comparisons: a primer on measurement invariance testing. *Chemistry Education Research and Practice, 21*(3), 969-988. https://doi.org/10.1039/d0rp00025f

Rotgans, J. I., & Schmidt, H. G. (2011). Cognitive engagement in the problem-based learning classroom. *Advances in Health Science Education & Theory Practice, 16*(4), 465-479. https://doi.org/10.1007/s10459-011-9272-9

Ryan, M. D., & Reid, S. A. (2015). Impact of the Flipped Classroom on Student Performance and Retention: A Parallel Controlled Study in General Chemistry. *Journal of Chemical Education, 93*(1), 13-23. https://doi.org/10.1021/acs.jchemed.5b00717

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

Satorra, A., & Bentler, P. (1988). Scaling corrections for statistics in covariance structure analysis.

Sawtelle, V., Brewe, E., Goertzen, R. M., & Kramer, L. H. (2012). Identifying Events that Impact Self-Efficacy in Physics Learning. *Physical Review Special Topics - Physics Education Research, 8*(2). https://doi.org/10.1103/PhysRevSTPER.8.020111

Schunk, D. H., & Ertmer, P. A. (2000). Self-Regulation and Academic Learning: Self-Efficacy Enhancing Interventions. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 631-651). Academic Press.

Seery, M. K. (2015a). ConfChem Conference on Flipped Classroom: Student Engagement with Flipped Chemistry Lectures. *Journal of Chemical Education, 92*(9), 1566-1567. https://doi.org/10.1021/ed500919u

Seery, M. K. (2015b). Flipped learning in higher education chemistry: emerging trends and potential directions. *Chemistry Education Research and Practice, 16*(4), 758-768. https://doi.org/10.1039/c5rp00136f

Shadle, S. E., Marker, A., & Earl, B. (2017). Faculty drivers and barriers: laying the groundwork for undergraduate STEM education reform in academic departments. *International Journal of STEM Education, 4*(1), 8. https://doi.org/10.1186/s40594-017-0062-7

Shattuck, J. C. (2016). A Parallel Controlled Study of the Effectiveness of a Partially Flipped Organic Chemistry Course on Student Performance, Perceptions, and Course Completion. *Journal of Chemical Education, 93*(12), 1984-1992. https://doi.org/10.1021/acs.jchemed.6b00393

Shortlidge, E. E., Rain-Griffith, L., Shelby, C., Shusterman, G. P., & Barbera, J. (2019). Despite Similar Perceptions and Attitudes, Postbaccalaureate Students Outperform in Introductory Biology and Chemistry Courses. *CBE—Life Sciences Education, 18*(1), ar3. https://doi.org/10.1187/cbe.17-12-0289

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The Challenges of Defining and Measuring Student Engagement in Science. *Educational Psychologist, 50*(1), 1-13. https://doi.org/10.1080/00461520.2014.1002924

Skinner, E., Saxton, E., Currie, C., & Shusterman, G. (2017). A motivational account of the undergraduate experience in science: brief measures of students' self-system appraisals, engagement in coursework, and identity as a scientist. *International Journal of Science Education, 39*(17), 2433-2459. https://doi.org/10.1080/09500693.2017.1387946

Smith, J. D. (2013). Student attitudes toward flipping the general chemistry classroom. *Chemistry Education Research and Practice, 14*(4), 607-614. https://doi.org/10.1039/c3rp00083d

Smith, K. C., & Alonso, V. (2020). Measuring student engagement in the undergraduate general chemistry laboratory. *Chemistry Education Research and Practice, 21*(1), 399-411. https://doi.org/10.1039/c8rp00167g

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. *CBE—Life Sciences Education, 12*(4), 618-627. https://doi.org/10.1187/cbe.13-08-0154

Snow, R. E. (1996). Self-Regulation as Meta-Conation? *Learning and Individual Differences, 8*(3), 261-267. https://doi.org/10.1016/S1041-6080(96)90017-5

Stains, M., & Harshman, J. *COPUS Analyzer*. Retrieved January 21, 2020 from http://www.copusprofiles.org

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Jr., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., Rodela, T. M., Sanders, E. R., Schimpf, N. G., Schulte, P. M., Smith, M. K., Stetzer, M., Van Valkenburgh, B., Vinson, E., Weir, L. K., Wendel, P. J., Wheeler, L. B., & Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science, 359*(6383), 1468-1470. https://doi.org/10.1126/science.aap8892

Stains, M., & Vickrey, T. (2017). Fidelity of Implementation: An Overlooked Yet Critical Construct to Establish Effectiveness of Evidence-Based Instructional Practices. *CBE—Life Sciences Education, 16*(1). https://doi.org/10.1187/cbe.16-03-0113

Stanford, C., Moon, A., Towns, M., & Cole, R. (2016). Analysis of Instructor Facilitation Strategies and Their Influences on Student Argumentation: A Case Study of a Process Oriented Guided Inquiry Learning Physical Chemistry Classroom. *Journal of Chemical Education, 93*(9), 1501-1513. https://doi.org/10.1021/acs.jchemed.5b00993

Stanich, C. A., Pelch, M. A., Theobald, E. J., & Freeman, S. (2018). A New Approach to Supplementary Instruction Narrows Achievement and Affect Gaps for Underrepresented Minorities, First-Generation Students, and Women. *Chemistry Education Research and Practice, 19*(3), 846-866. https://doi.org/10.1039/C8RP00044A

Steiger, J. H. (1980). *Statistically Based Tests for the Number of Common Factors* The Annual Meeting of the Psychometric Society, Iowa City, IA.

Sunny, C. E., Taasoobshirazi, G., Clark, L., & Marchand, G. (2016). Stereotype Threat and Gender Differences in Chemistry. *Instructional Science, 45*(2), 157-175. https://doi.org/10.1007/s11251-016-9395-8

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312. https://doi.org/10.1016/0959-4752(94)90003-5

Tabak, F., Nguyen, N., Basuray, T., & Darrow, W. (2009). Exploring the Impact of Personality on Performance: How Time-on-Task Moderates the Mediation by Self-Efficacy. *Personality and Individual Differences, 47*(8), 823-828. https://doi.org/https://doi.org/10.1016/j.paid.2009.06.027

The POGIL Project. *POGIL*. Retrieved 2020 from https://pogil.org/about-pogil/what-is-pogil

Thompson, M. S., & Green, S. B. (2013). Evaluating Between-Group Differences in Latent Variable Means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 163-218). Information Age Publishing.

Villafañe, S. M., Garcia, C. A., & Lewis, J. E. (2014). Exploring Diverse Students' Trends in Chemistry Self-Efficacy throughout a Semester of College-Level Preparatory Chemistry. *Chemistry Education Research and Practice, 15*(2), 114-127. https://doi.org/10.1039/C3RP00141E

Villafañe, S. M., Xu, X., & Raker, J. R. (2016). Self-Efficacy and Academic Performance in First-Semester Organic Chemistry: Testing a Model of Reciprocal Causation. *Chemistry Education Research and Practice, 17*(4), 973-984. https://doi.org/10.1039/C6RP00119J

Vishnumolakala, V. R., Southam, D. C., Treagust, D. F., Mocerino, M., & Qureshi, S. (2017). Students' Attitudes, Self-Efficacy and Experiences in a Modified Process-Oriented Guided Inquiry Learning Undergraduate Chemistry Classroom. *Chemistry Education Research and Practice, 18*(2), 340-352. https://doi.org/10.1039/C6RP00233A

Walker, C. O., Greene, B. A., & Mansell, R. A. (2006). Identification with academics, intrinsic/extrinsic motivation, and self-efficacy as predictors of cognitive engagement. *Learning and Individual Differences, 16*(1), 1-12. https://doi.org/10.1016/j.lindif.2005.06.004

Wang, C., Cavanagh, A. J., Bauer, M., Reeves, P. M., Gill, J. C., Chen, X., Hanauer, D. I., & Graham, M. J. (2021). A Framework of College Student Buy-in to Evidence-Based Teaching Practices in STEM: The Roles of Trust and Growth Mindset. *CBE—Life Sciences Education, 20*(4), ar54. https://doi.org/10.1187/cbe.20-08-0185

Wang, J., & Wang, X. (2019). *Structural Equation Modeling: Applications using Mplus*. John Wiley & Sons.

Wang, M.-T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The Math and Science Engagement Scales: Scale development, validation, and psychometric properties. *Learning and Instruction, 43*, 16-26. https://doi.org/10.1016/j.learninstruc.2016.01.008

Wang, M. T., & Peck, S. C. (2013). Adolescent educational success and mental health vary across school engagement profiles. *Developmental Psychology, 49*(7), 1266-1276. https://doi.org/10.1037/a0030028

Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item Response Theory Models for Wording Effects in Mixed-Format Scales. *Educational and Psychological Measurement, 75*(1), 157-178. https://doi.org/10.1177/0013164414528209

Weasel, L. H., & Finkel, L. (2016). Deliberative Pedagogy in a Nonmajors Biology Course: Active Learning That Promotes Student Engagement With Science Policy and Research. *Journal of College Science Teaching, 45*(4). https://doi.org/10.2505/4/jcst16_045_04_38

Weaver, G. C., & Sturtevant, H. G. (2015). Design, Implementation, and Evaluation of a Flipped Format General Chemistry Course. *Journal of Chemical Education, 92*(9), 1437-1448. https://doi.org/10.1021/acs.jchemed.5b00316

Webber, D. M., & Flynn, A. B. (2018). How Are Students Solving Familiar and Unfamiliar Organic Chemistry Mechanism Questions in a New Curriculum? *Journal of Chemical Education, 95*(9), 1451-1467. https://doi.org/10.1021/acs.jchemed.8b00158

Weinstein, C. E., Palmer, D., & Schulte, A. C. (2002). *Learning and Study Strategies Inventory* (2nd ed.). H & H.

Wieman, C. E. (2014). Large-scale comparison of science teaching methods sends clear message. *Proceedings of the National Academy of Sciences of the United States of America, 111*(23), 8319-8320. https://doi.org/10.1073/pnas.1407304111

Wiggins, B. L., Eddy, S. L., Wener-Fligner, L., Freisem, K., Grunspan, D. Z., Theobald, E. J., Timbrook, J., & Crowe, A. J. (2017). ASPECT: A Survey to Assess Student Perspective of Engagement in an Active-Learning Classroom. *CBE—Life Sciences Education, 16*(2). https://doi.org/10.1187/cbe.16-08-0244

Wilde, J. (2012). The Relationship between Frustration Intolerance and Academic Achievement in College. *International Journal of Higher Education, 1*(2). https://doi.org/10.5430/ijhe.v1n2p1

Willis, G. B. (2005). *Cognitive Interviewing*. SAGE Publications, Inc. https://dx.doi.org/10.4135/9781412983655

Wiltbank, L. B., Williams, K. R., Marciniak, L., & Momsen, J. L. (2019). Contrasting Cases: Students' Experiences in an Active-Learning Biology Classroom. *CBE—Life Sciences Education, 18*(3), ar33. https://doi.org/10.1187/cbe.19-01-0006

Winne, P. H. (2004). Students’ Calibration of Knowledge and Learning Processes: Implications for Designing Powerful Software Learning Environments. *International Journal of Educational Research, 41*(6), 466-488. https://doi.org/https://doi.org/10.1016/j.ijer.2005.08.012

Wolters, C. A., Won, S., & Hussain, M. (2017). Examining the Relations of Time Management and Procrastination within a Model of Self-Regulated Learning. *Metacognition and Learning, 12*(3), 381-399. https://doi.org/10.1007/s11409-017-9174-1

Woodward, R. L., & Reid, C. S. (2019). You've Got Mail (and Homework): Simple Strategies for Promoting Student Engagement with Prelecture Videos. *Journal of Chemical Education, 96*(9), 2055-2058. https://doi.org/10.1021/acs.jchemed.9b00315

Wren, D., & Barbera, J. (2013). Gathering Evidence for Validity during the Design, Development, and Qualitative Evaluation of Thermochemistry Concept Inventory Items. *Journal of Chemical Education, 90*(12), 1590-1601. https://doi.org/10.1021/ed400384g

Wu, F., & Teets, T. S. (2021). Effects of the COVID-19 Pandemic on Student Engagement in a General Chemistry Course. *Journal of Chemical Education, 98*(12), 3633-3642. https://doi.org/10.1021/acs.jchemed.1c00665

Xu, X., & Lewis, J. E. (2011). Refinement of a Chemistry Attitude Measure for College Students. *Journal of Chemical Education, 88*(5), 561-568. https://doi.org/10.1021/ed900071q

Ye, F., & Wallace, T. L. (2013). Psychological Sense of School Membership Scale. *Journal of Psychoeducational Assessment, 32*(3), 202-215. https://doi.org/10.1177/0734282913504816

Young, A. M., Wendel, P. J., Esson, J. M., & Plank, K. M. (2018). Motivational Decline and Recovery in Higher Education STEM Courses. *International Journal of Science Education, 40*(9), 1016-1033. https://doi.org/https://doi.org/10.1080/09500693.2018.1460773

Zeldin, A. L., & Pajares, F. (2016). Against the Odds: Self-Efficacy Beliefs of Women in Mathematical, Scientific, and Technological Careers. *American Educational Research Journal, 37*(1), 215-246. https://doi.org/10.3102/00028312037001215

Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist, 25*(1), 3-17. https://doi.org/10.1207/s15326985ep2501_2

Zimmerman, B. J. (1995). Self-efficacy and educational development. In A. Bandura (Ed.), *Self-efficacy in changing societies*. Cambridge University Press.

Zimmerman, B. J. (1998). Developing Self-Fulfilling Cycles of Academic Regulation: An Analysis of Exemplary Instructional Models. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-Regulated Learning: From Teaching to Self-Reflective Practice* (pp. 1-19). Guilford Press.

Zimmerman, B. J. (2000). Attaining Self-Regulation: A Social Cognitive Perspective. In M. Boekaerts, P. R. Pintrick, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13-41). Academic Press.

Zumbrunn, S., McKim, C., Buhs, E., & Hawley, L. R. (2014). Support, belonging, motivation, and engagement in the college classroom: a mixed method study. *Instructional Science, 42*(5), 661-684. https://doi.org/10.1007/s11251-014-9310-0

Appendix A: Supporting Information for Chapter 4

COPUS Code Descriptions for Selected Codes

Student Codes

The listening (**L**) code was used across all courses to indicate any time students were expected to be listening to the instructor/presenter, with or without an expectation of note-taking, such as during lecture or follow-up activities, but also while the instructor presented administrative information.

The individual work (**Ind**) code was used for any type of independent work that was not an official test or quiz. In Course Two, this code was used when students conducted "speed drills” over prior material at the beginning of class, and in Course Three, this code referred to a pre-test and post-test worksheet related to the in-class game.

The clicker groupwork (**CG**) code was only used when a clicker question was posed to students and the students discussed the question among themselves in groups of two or more, whereas the worksheet group (**WG**) code was only used if the students had a given worksheet/workbook on which to work, whether that worksheet/workbook was physical as in Courses One and Four, or digital as in Course Five. In Course Three, the other groupwork (**OG**) code was used on day one to denote the game students played as groups, and on day two was used to denote groupwork where students were asked to answer a question that was neither part of a worksheet nor a clicker question.

The answer question (**AnQ**) student code was used across all courses when students answered non-rhetorical questions posed by the instructor with the rest of the class listening, whereas the student question (**SQ**) code was used when students posed questions to the instructor, whether subject-matter related or administrative, with the rest of the class listening. In Course Two, in addition to the usual sense, the SQ code was used while the instructor read and answered student questions submitted to the clicker input system out loud to the class.

The test or quiz (**TQ**) code was used any time students took a test or quiz that was handed in during class to be graded. As participating courses were observed only during non-exam weeks, this code only refers to an in-class quiz given in Course One. The waiting (**W**) student code was used when instructors had technical problems with their presentations and students were waiting for class to resume. The other (**O**) code was used to denote students coming into class late or leaving early, pointing out inconsistencies in the text that was presented on the board/screen, or giving general comments.

Instructor Codes

The lecture (**Lec**) code was used across all courses when the instructor presented subject-matter related information to the class that was not related to an example problem, whereas the follow-up (**FlUP**) code was used across all courses when the instructor explained and/or followed-up on a student activity, such as clicker questions, or group or individual work. The real-time writing (**RtW**) code was used any time the instructor wrote on a surface for the whole class to see, such as a whiteboard, smartboard, or document projector.

The pose question (**PQ**) instructor code was used across all courses when an instructor posed a non-rhetorical question to the whole class and gave space for individual students to answer with the whole class listening. In Course Two, this code was also used to denote when the instructor administered "speed drills," as they were non-rhetorical and non-clicker questions posed to the entire class that the students were expected to answer. The clicker question (**CQ**) code was used any time an instructor conducted whole-class polling, with or without clicker software. For example, instructors could ask students to raise their hand in a vote, such as in Course One, or to raise colored response cards, such as in Course Five. Clicker questions could be to assess course knowledge, such as in Courses Two, Three, and Five, or to poll students regarding administrative matters. For example, Course One used a hand-raising poll to vote on when homework assignments would be due and Course Two used clicker polling for students to ask questions of the instructor or to self-report how many "speed drill" questions they answered correctly. The answer question (**AnQ**) instructor code was used across all courses when instructors answered questions posed by the students, whether subject-matter related or administrative, with the rest of the class listening. In Course Two, in addition to the usual sense, AnQ was used while the instructor read and answered student questions submitted to the clicker input system out loud to the class.

The moving-and-guiding (**MG**) code was used across all courses whenever the instructor was moving around the room and guiding work, either group or individual. In courses with learning assistants (LAs), such as Courses One, Three, and Four, this code was used if the instructor or even one of the LAs were moving through and guiding groups in a given time block, rather than indicating that all LAs were currently circulating. Relatedly, the one-on-one (**1o1**) code was used to indicate a directed conversation between the instructor with a single student or group. In courses with LAs, this code was used if the instructor or even one of the LAs participated in a one-on-one in a given time block, rather than indicating that all LAs were occupied in one-on-ones.

The administration (**Adm**) code was used when the instructor gave class-wide announcements related to facilitating the class meeting (e.g., time left on an activity) or course reminders (e.g., test dates, school closures, etc.) In addition to the usual sense, in Course Three, Adm was used when the instructor gave instructions regarding the in-class game. In Course Five, the demo (**D**) code was used to indicate the instructor performing a lab experiment in class. In Course Three, the waiting (**W**) instructor code was used when there was an opportunity for the instructor to be interacting with or observing/listening to student or group activities and the instructor was not doing so. The other (**O**) instructor code was used to denote instructors conferencing with LAs or researchers, or leaving the room entirely. In Course Three, O was used during day one to indicate the LAs collecting and/or distributing materials related to the in-class game, usually at the same time the instructor was still giving instructions (Adm) at the front of the classroom.

F2F Observation Summaries and COPUS Timelines

The COPUS protocol contains 13 student and 12 instructor categories to utilize during each two-minute interval within a F2F environment (Smith et al., 2013). These categories are meant to capture the range of common behaviors that typically occur in courses. Of the 25 codes, 22 were observed to occur across the recorded F2F sessions

(Table A.1). The categories not observed included those associated with student presentations, predictions and whole-class discussion.

Table A.1. COPUS codes from observations, colors matched to course timelines shown in Figures A.1 – A.5.

Student	**Code**	**Description**
Receiving	L	Listening to instructor
Groupwork	WG	Working in groups on worksheet activity
	CG	Discussing clicker question
	OG	Other assigned group activity
Individual Work	Ind	Individual thinking/problem solving
Q&A	AnQ	Answering question posed by instructor
	SQ	Asking a question
Non-work	TQ	Test or quiz
	W	Waiting
	O	Other
Instructor	**Code**	**Description**
Presenting	Lec	Lecturing or presenting information
	RtW	Real-time writing
	D/V	Showing or conducting a demo, experiment, simulation, etc.
Engaging	FlUp	Follow-up/feedback on clicker question or activity
	MG	Moving through class guiding ongoing student work
	1o1	One-on-one extended discussion with individual students
Q&A	PQ	Posing non-clicker question to students (non-rhetorical)
	CQ	Asking clicker question (entire duration)
	AnQ	Listening to and answering student questions to entire class
Non-work	Adm	Administration (assign homework, return materials)
	W	Waiting
	O	Other

For each two-minute time-block, the timelines indicate which student and/or instructor codes were observed (Figures A.1 – A.5). For further interpretation, we also determined the percentage of time-blocks in which a specific code was observed. These values are presented at the end of each code's row in a given timeline.

Course One

Class sessions revolved around the discussion and completion of pages from an instructor-authored workbook. The instructor displayed pages of the workbook on a document camera and worked through each problem using a variety of methods including lecturing (Lec, 17% of day one and 40% of day two 2-minute time blocks shown in Figure A.1), whole-class questioning (PQ, 83% and 80%) with follow-ups (FlUp, 63% and 49%), and real-time writing (RtW, 69% and 74%). The temporal associations of these often-overlapping methods is shown in Figure A.1. Students were observed mainly listening (L, 94% each day) or responding to instructor-posed whole-class questions (AnQ, 83% and 80%). Groupwork (WG, 29% and 40%) was observed to be spread throughout the class time, typically lasting between 2 – 8 minutes. The instructor and/or learning assistants were observed moving through the class guiding each group's work (MG, 26% and 37%). When these timelines were analyzed with the COPUS Analyzer

(Stains & Harshman), the COPUS Profile matched cluster 6, representing a student-centered instructional style.

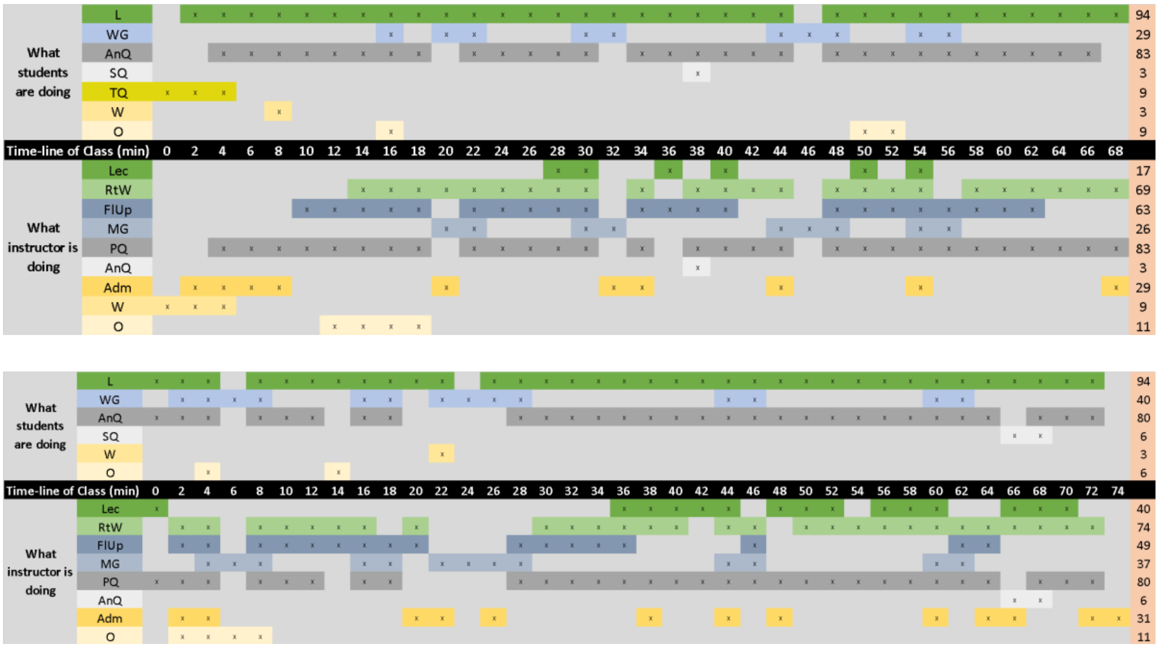


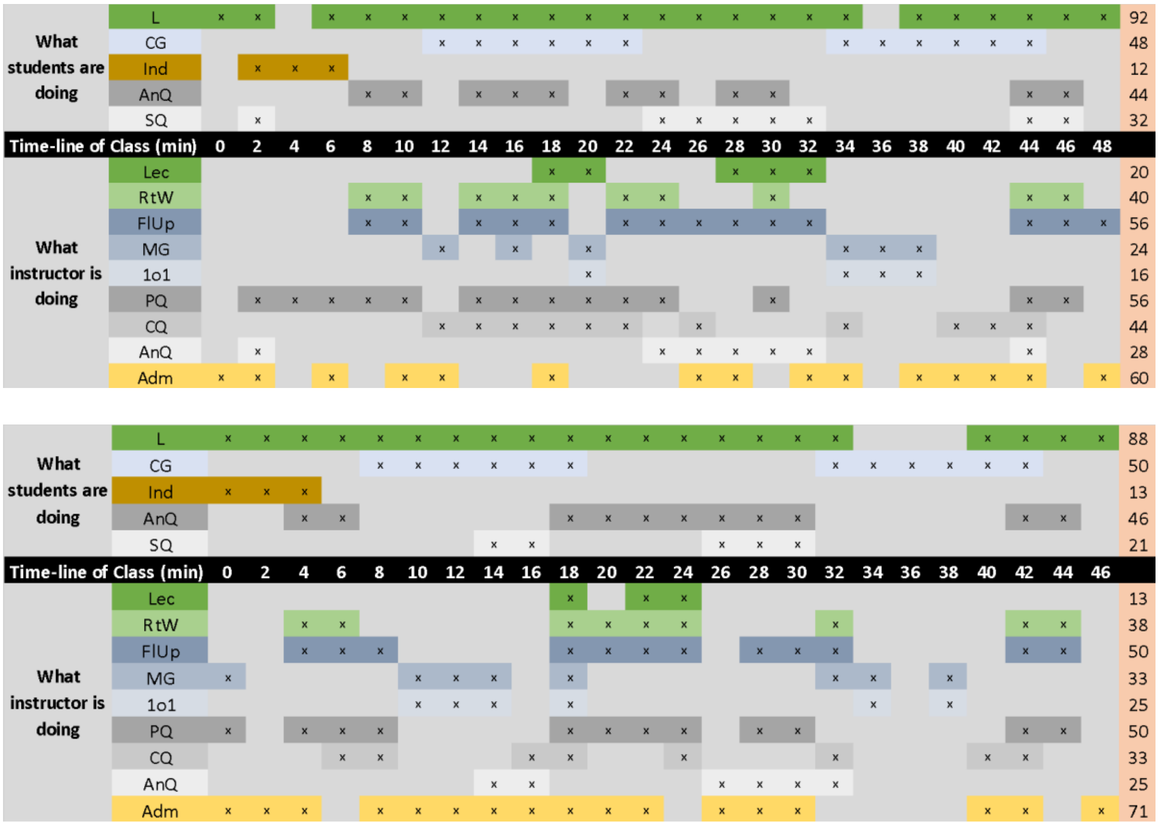
Figure A.1. COPUS code timelines for Course One. Observation day one on top and day two on bottom. Code abbreviations and colors are provided in Table A.1. Values at the end of each row are the percentage of 2-minute time blocks the code was observed.

The workbook used in Course One contained POGIL-style worksheets, by topic, that included models (often figures from the textbook) and text-like explanations of the content. Worksheet topics were aligned with the assigned pre-class videos. The worksheet from the day one observation contained models on the topic of periodic trends, models from day two covered ionic compound naming. Each model was followed by a series of questions. Initial questions asked students to extract and explain information from the model (e.g., key questions in the POGIL framework) (Hanson, 2007) to acquire new knowledge. These were followed by exercises requiring students to use the information presented in the model.

Course Two

Class sessions revolved around individual- and group-based questions posed by the instructor. Each class began with a set of questions about prior material. Students were observed to work independently (Ind, Figure A.2) during these "speed drills" and then entering their responses into the clicker application using their phone or tablet (12% and 13% of each day's two 2-minute time blocks). The instructor followed up on these questions with brief clarification as needed. The remainder of the class time moved back and forth between students responding to group-based clicker questions (CG, 48% and 50%), students posing questions (SQ, 32% and 21%), and the instructor following-up on questions (FlUp, 56% and 50%). During the group clicker questions, the instructor was observed to be guiding student/groupwork (MG, 24% and 33%) and answering questions

(1o1, 16% and 25%); no learning assistants were present. The majority of the class time was back-and-forth questioning by the instructor (PQ, 56% and 50%) and answering by the students (AnQ, 44% and 46%). The COPUS Profile for these timelines matched cluster 4, indicating an interactive lecture instructional style.



	Time-line of Class (min)	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	
What students are doing	L	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	92
	CG							x	x	x	x	x	x						x	x	x	x	x	x			48
	Ind		x	x	x																						12
	AnQ					x	x		x	x	x		x	x		x	x							x	x		44
	SQ		x											x	x	x	x	x						x	x		32
What instructor is doing	Lec										x	x				x	x	x									20
	RtW					x	x		x	x	x		x	x			x							x	x		40
	FlUp					x	x		x	x	x		x	x	x	x	x	x						x	x	x	56
	MG							x		x		x							x	x	x						24
	1o1											x							x	x	x						16
	PQ		x	x	x	x	x		x	x	x	x	x	x			x							x	x		56
	CQ							x	x	x	x	x	x		x				x			x	x	x			44
	AnQ		x											x	x	x	x	x						x			28
	Adm	x	x		x		x	x			x				x	x		x	x		x	x	x	x		x	60

	Time-line of Class (min)	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	
What students are doing	L	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x				x	x	x	x	88
	CG					x	x	x	x	x	x							x	x	x	x	x	x			50
	Ind	x	x	x																						13
	AnQ			x	x						x	x	x	x	x	x	x						x	x		46
	SQ								x	x					x	x	x									21
What instructor is doing	Lec										x		x	x												13
	RtW			x	x						x	x	x	x				x					x	x		38
	FlUp			x	x	x					x	x	x	x		x	x	x					x	x		50
	MG	x					x	x	x		x							x	x		x					33
	1o1						x	x	x		x								x		x					25
	PQ	x		x	x	x					x	x	x	x		x	x						x	x		50
	CQ				x	x				x	x			x				x				x	x			33
	AnQ								x	x					x	x	x	x								25
	Adm	x	x	x		x	x	x	x	x	x	x	x		x	x	x					x	x		x	71

Figure A.2. COPUS code timelines for Course Two. Observation day one on top and day two on bottom. Code abbreviations and colors are provided in Table A.1. Values at the end of each row are the percentage of 2-minute time blocks the code was observed.

Class sessions in Course Two entailed 10 – 15 clicker questions presented from a tablet, allowing the instructor to annotate responses as needed. The content of each class session was aligned with the assigned pre-class videos. Day one questions covered electron transfer reactions. Initial items included a series of reaction classification questions followed by a longer series regarding the determination of the oxidation number for a selected atom in a compound. Subsequent questions asked students to apply the concept to determining the oxidized and reduced species for a specified redox reaction. Initial items on day two covered oxidation number assignments followed by a series of reaction completion questions where students were given reactants and a table of cell potentials and asked to predict the products. A final set of questions asked students to identify the number of electrons transferred or the oxidizing agent for several redox reactions.

Course Three

Class sessions revolved around cycles of topic introduction, examples, and a topic review (Figure A.3). Each cycle was observed to begin with the instructor introducing a topic through lecture (Lec, 38% and 74% of each day's two 2-minute time blocks) and annotating the prepared slides (RtW, 3% and 51%). Following this introduction, students were presented with an example to work. Students were then observed discussing with their neighbors (CG, 21% and 26%), and submitting their individual responses into the clicker application using their phone or computer. During this time, the instructor and learning assistants were observed to be guiding student/groupwork (MG, 41% and 18%). Once responses were submitted, the instructor was observed reviewing the example and providing additional context as needed (FlUp, 13% and 31%). In addition to this typical in-class practice, day one included a non-typical group activity presented in the form of a game. Midway through the class session, students individually (Ind, 15%) completed and turned in a pre-test worksheet. They were then given instructions (coded as Adm) for the game, which was based on the topic of intermolecular forces. Students worked together to play the game (OG, 28%), racing to respond to questions from an arbiter. They then completed an individual post-test worksheet. This non-typical activity accounted for the variation in code percentages across observation days, especially in the case of the administration code (Adm, 49% and 8%), used to note the instructor's explanations of the game's rules and directions regarding the worksheets. The COPUS Profile for these timelines matched an interactive lecture instructional style represented by cluster 4.

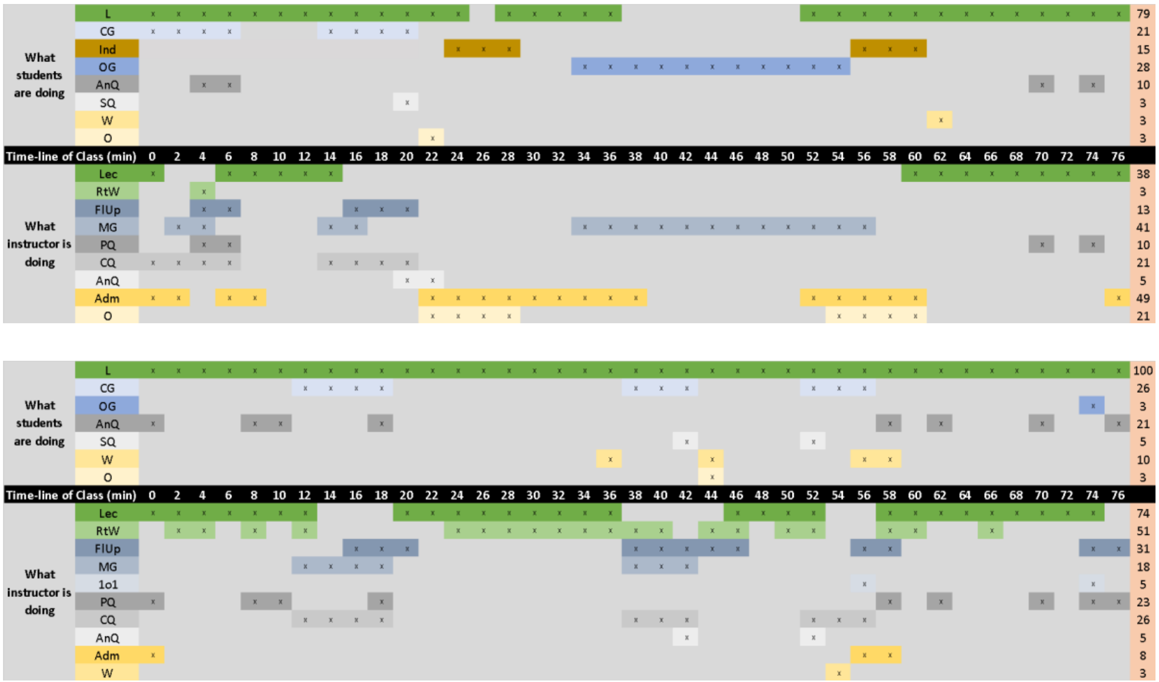


Figure A.3. COPUS code timelines for Course Three. Observation day one on top and day two on bottom. Code abbreviations and colors are provided in Table A.1. Values at the end of each row are the percentage of 2-minute time blocks the code was observed.

In Course Three, day one included two topic cycles around intermolecular forces (heats of vaporization/fusion and viscosity/surface tension/capillary action), day two included three topic cycles around phase changes (vapor pressure, heating curves, and phase diagrams). The instructor utilized prepared slides when introducing a topic, which included figures and graphs/plots from the text. These slides were annotated as needed. The examples presented to students during each cycle were calculation-based and, as indicated by the instructor, were either straight from or slight modifications (i.e., different numeric values) to those presented in the pre-class videos. The day one examples were one-step processes (e.g., energy change given an amount and standard heat value), some on day two included more multi-step calculations (e.g., energy change along multiple sections of a heating curve). For the day one non-typical activity, each group of three students were given a set of cards. Each card displayed a molecular structure and the molecule's boiling point. One student (the arbiter) selected two cards and formulated an intermolecular force question about them (e.g., Which has stronger forces?). The other two students raced to identify which molecule answered the question.

Course Four

Class sessions revolved around small-group completion of instructor-authored problem sets. Students were observed working in their assigned groups (WG, Figure A.4) during a majority of the time-blocks (79% of each day's two 2-minute time blocks). During these times the instructor and learning assistants were observed moving through the class guiding each group's work (MG, 74% and 68%) and responding to their questions (1o1, 77% and 68%). During the last ~20 minutes of class, the instructor concluded the activity by posing whole-class questions (PQ, 21% and 16%) on the focal points of the day's content. During this time, the instructor was observed providing clarifying information by lecturing (L, 23% and 24%), working a sample problem (RtW, 21% and 16%), and/or through follow-up questions (FlUp, 23% and 24%). During these wrap-up sessions, students were observed listening to the instructor (L, 23% and 29%) and answering questions (AnQ, 21% and 16%). These timelines matched a cluster 6 COPUS Profile indicating a student-centered instructional style.

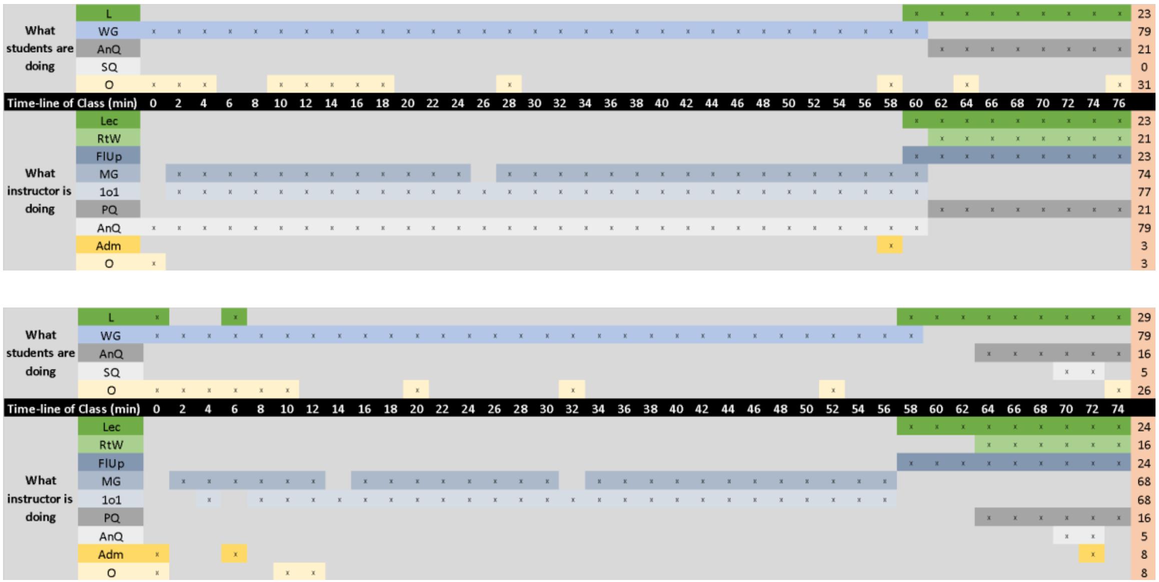


Figure A.4. COPUS code timelines for Course Four. Observation day one on top and day two on bottom. Code abbreviations and colors are provided in Table A.1. Values at the end of each row are the percentage of 2-minute time blocks the code was observed.

The problem sets in Course Four contained little to no guiding information and focused on a single topic from the assigned pre-class videos. The formal charge problems on day one contained a series of items asking students to first describe/explain the concept, then requiring them to apply it to a series of structures, and finally use the concept to explain the difference in atom connectivity for given formulas. On day two, the problems involved a series of molecular and ionic formulas for which students were asked to determine the electron pair and molecular geometries as well as to predict their bond angles.

Course Five

Class sessions revolved around small-group completion of an online instructor-authored worksheet. Students were observed working in their assigned groups (WG, Figure A.5) during almost all time-blocks (92% and 98% of each day's two 2-minute time blocks) and entering responses on a laptop. The instructor was observed to be consistently moving through the class guiding each group's work (MG, 75% and 85%) and responding to their questions (1o1, 86% and 98%); no learning assistants were present. Intermittently throughout the class time, the instructor was observed posing clicker questions (CQ, 14% and 15%); students were observed working on them individually (Ind, 14% and 15%) and presented their answer using colored response cards. The instructor was rarely observed presenting information (Lec, 0% and 2% and FlUp, 11% and 10%), with occurrences correlated to the clicker questions. Students were observed listening to the instructor (L, 42% and 27%) mainly when administrative details (Adm, 36% and 27%) were given or when check-in questions were being delivered. The COPUS Profile for these timelines matched cluster 5, which represents a student-centered instructional style.

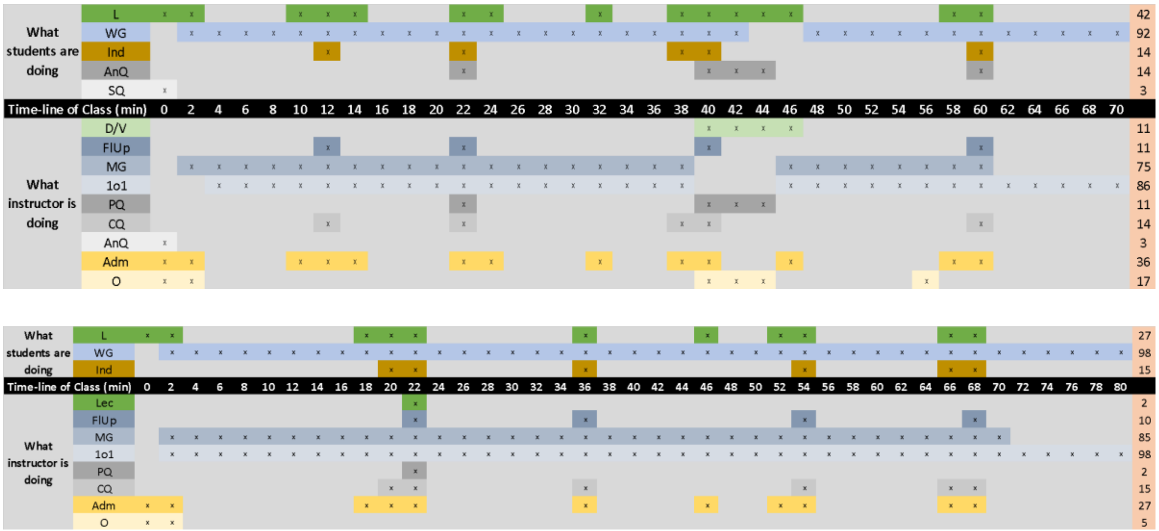


Figure A.5. COPUS code timelines for Course Five. Observation day one on top and day two on bottom. Code abbreviations and colors are provided in Table A.1. Values at the end of each row are the percentage of 2-minute time blocks the code was observed.

The online workbook used in Course Five contained POGIL-style worksheets. Each contained a number of models with a figure and limited text introducing the context of the figure. The class only met once per week and, as noted by the instructor, the worksheets were aligned with the more foundational or difficult topics covered in the video lectures assigned that week. The worksheets from the observed days covered the topics of electron configurations and periodic trends on day one and bonding on day two. Each model was followed by a series of questions. Initial questions asked students to extract and explain information from the model (i.e., key questions in the POGIL framework) (Hanson, 2007) to acquire new knowledge. These were followed by exercises requiring students to use the information presented in the model.

Pre-Class Material (PCM) Survey Development

The survey was developed through an iterative process that included two rounds of focus groups and one round of a pilot survey conducted in previous semesters of the courses. Brief descriptions of these processes are outlined in this section. The pilot version of the items can be found in supplemental Table A.4. Final item versions are embedded throughout Chapter 4 and Appendix A (in bold) as well within Chapter 4 Tables 4.5 – 4.7 and Appendix A Tables A.5 – A.9.

Focus Group Participants

During a developmental phase of the project, students were recruited from Courses Two, Four, and Five. An announcement regarding the focus groups was placed at the end of a pre-semester survey administered as part of the larger project. Students noted their interest in participating in a discussion group about their flipped course and provided contact information for scheduling. Prior to an on-campus visit, all students who expressed interest were notified of the focus group times and asked to respond indicating which group(s) fit their availability. Participating students were compensated with a $10 gift card.

Focus Group Data Collection and Analysis Protocols

Focus groups were conducted in person. Due to timing and other constraints, groups were not conducted with students from Courses One and Three. As researchers from the project were on each campus for a minimum of three days, a range of days and times were provided for the focus groups. For ease of access, each group met in a private location in or near the building where the flipped chemistry course was taught. Each group was conducted by two researchers. The first interviewer initiated the focus group questions, the follow-up questions, and managed discussions among participants. The second interviewer took notes on the discussion and provided additional follow-up questions or asked clarifying statements as needed. All focus groups were video recorded and these recordings were coded by one or two researchers.

The first round of focus groups were conducted with students from Course Four and were general in nature, asking about overall study habits with regard to the flipped course (Table A.2). Two coders conducted reviews of the video recordings from each group to develop a codebook and subsequently code each groups' responses. Independently, each coder reviewed three videos and documented the salient responses for each question, creating their own preliminary codebook. The coders then met to discuss their codebooks and generated a single codebook. The codebook was then independently used by each coder to review the remaining videos. This coding process produced Cohen's kappa scores >0.81 for each video, indicating near perfect inter-rater reliability (Cohen, 1960). The coded responses from these groups provided insights for the development of more formulated questions around students' use of the video resources for the pilot survey and an additional round of focus groups.

Table A.2. Pre-class preparation focus group questions.

As your instructor requests that you come to class prepared to engage in learning activities, the next set of questions is about your pre-class preparation.
1. On average, how much time do you spend preparing for class?
2. What do you typically do to prepare for class?
3. Are there other things you could/should be doing prior to class?
4. What are the limitations to doing these additional things?

The second round of focus groups, conducted with students in Courses Two and Five, focused on questions specific to the pre-class videos used in the flipped courses (Table A.3). A primary coder reviewed two focus group videos, one from each institution, and created an initial codebook. A secondary coder used the initial codebook to independently code the same two videos, following which they met to discuss discrepancies and make codebook modifications. Using the modified codebook, each reviewer independently coded two additional videos. This coding process produced Cohen's kappa scores >0.81 for each video, indicating near perfect inter-rater reliability (Cohen, 1960). All subsequent videos were coded by the primary coder using the modified codebook. Responses from these groups were used when developing the final version of the survey from the pilot version.

Table A.3. Pre-class preparation focus group questions.

1. Do you regularly watch the posted videos?
Follow-ups: For those that do
● How do you watch them? (e.g., from start-to-finish completely, skip around, fast forward through, etc.) ● What do you do when watching them? (e.g., nothing but watch and listen, take notes, work on quiz or homework questions, etc.) ● When do you typically watch them for the first time? (e.g., evening before class, end of week, etc.) ● Do you ever re-watch them? If so, when?
Follow-ups: For those that do not
● Have you ever accessed them? ● If yes, why do you not regularly watch them? ● If no, why do you not access them?
2. Do you find the videos helpful for learning the material? Please explain why or why not.
3. What do you like or dislike about the videos? Please explain.

Focus Group Results

A total of 56 students participated in focus groups, 10 from Course Two (7 groups), 24 from Course Four (7 groups), and 22 from Course Five (8 groups). Groups varied in size from a maximum of 4 to a minimum of 1. Due to scheduling issues and no-shows, some planned focus group meetings ended up including only one individual. While not ideal, we valued each student's time and input and therefore did not cancel smaller groups nor reject individual input.

From the first round of focus groups, conducted with students from Course Four, some general response themes emerged; such as, that students were watching the videos to prepare for class (79%) as well as re-watching the videos later (71%). Additionally, students reported general aspects that they liked about the videos (67%), such as being able to watch at their own pace, and also disliked (88%), such as not being able to ask questions. These general response themes were used to create specific questions around students' usage and perceptions of the pre-class videos, which were administered during the second round of focus groups and on the pilot survey.

The second round of focus groups, conducted with students from Courses Two and Five, provided similar and additional insights to students' use and perceptions of the pre-class videos. For example, with regard to how students watched the videos, only 37% reported that they watch straight through, with many reporting that they pause/rewind (87%), watch at a different pace (28%), or skip/fast-forward through (37%). Students expressed many positive perceptions of the videos such as being able watch at their own pace (64%) or whenever and/or wherever they want (45%), and commented positively about the structure (55%) and length (48%) of the videos. Their negative perceptions focused around not being able to ask questions (33%), that the videos did not keep their attention (21%), or that they did not feel there were enough problems or explanations provided (33%). In all, the responses from the two rounds of focus groups were informative in providing clarity to some of the pilot survey responses and to adjust the exact wording of items and responses for the final survey.

Pilot Survey Population and Administration

Students were recruited to participate in the pilot survey from Courses One, Two, and Five. Due to timing and other constraints, pilot survey data was not conducted in Courses Three and Four. Survey deployment in each course was coordinated to take place midway through the course during a non-exam week. The instructor was provided a brief script to make an initial in-class announcement regarding the survey. A note similar to the script was posted on the classroom management platform of each course. Students who were interested in participating clicked on a link to the Qualtrics survey that was part of the announcement note. Some instructors offered a nominal amount of extra-credit points for accessing the survey.

Pilot Survey Data Collection and Analysis Protocols

The pilot survey focused on questions related to the pre-class videos and contained a variety of item types including single-, multi-, and open-response formats (Table A.4). The survey flow contained logic steps that populated questions, and their associated follow-ups, based on a participant's prior responses. Therefore, the number of participants was not constant across items and not every individual was presented with each question or follow-up.

Response percentages for single- and multi-response item types were calculated based on the number of participants who were presented with the question. Open-response items were coded and response percentages per code were calculated based on the number of participants who were presented with the question. A primary coder reviewed an aggregated dataset, that contained the item-by-item responses from each institution, and created a codebook for each item. A secondary coder used the codebook to independently code all responses across items, following which they met to discuss any discrepancies. The coders discussed any noted discrepancies and came to consensus on the codes and their use.

Table A.4. Pre-class preparation pilot survey questions.

1. Do you regularly watch the posted videos? (Yes/No Radio Buttons)
(*If Yes*) ● Please explain how you watch them. (e.g., from start-to-finish completely, skip around, fast forward through, etc.) (Open-response) ● What do you do when watching them? (e.g., nothing but watch and listen, take notes, work on quiz or homework questions, etc.) (Open-response) ● When do you typically watch them for the first time? (Single-response question) ○ The evening before class ○ Within a few hours before class ○ Within a few hours after class ○ At the end of a week of class ○ Other (Please explain) (Open-response) ● Do you ever re-watch them? (Yes/No Radio Buttons) (*If Yes*) When do you re-watch the videos? (Multi-response) ○ At the end of a week of class ○ When studying for a quiz or exam ○ When working on the homework ○ Other (Please explain) (Open-response) (*If No*) Why do you not re-watch the videos? (Multi-response) ○ I don't find them useful to watch more than once ○ I watch other videos to get a different perspective than the ones posted ○ They typically don't match well to the quiz or exam material ○ Other (Please explain) (Open-response)
(*If No*) ● Have you ever accessed the posted videos? (Yes/No Radio Buttons) (*If Yes*) Why do you not regularly watch them? (Please explain) (Open-response) (*If No*) Why have you not accessed them? (Please explain) (Open-response)
2. [a]Do you find the videos helpful for learning the course material? (Yes/No/Sometimes Radio Buttons)
(*If Yes*) ● Please explain why you find the videos helpful for learning (Open-response)
(*If No*) ● Please explain why you do not find the videos helpful for learning (Open-response)
(*If Sometimes*) ● Please explain why you only sometimes find the videos helpful for learning (Open-response)
3. [a]What do you like about the videos? (Open-response)
4. [a]What do you dislike about the videos? (Open-response)

[a]Items 2 – 4 presented only if response to item 1 = yes

Pilot Survey Results and Survey Refinement

When presenting the results from the pilot survey, any resulting refinements for the final survey will also be noted. The pilot survey resulted in 263 responses from Course One, 27 from Course Two, and 22 from Course Five.

The first question, "**Do you regularly watch the posted videos?**", resulted in 94 (38%), 24 (89%), and 15 (68%) 'Yes' responses from each institution respectively. Students who responded with 'No' were not asked any of the additional follow-up questions about how they interacted with the videos. When this same question was asked of students in the second round of focus groups it was discovered that many students who did not *regularly* watch the posted videos still watched the videos occasionally or when they were confused about a specific topic. Therefore, the wording of this item was updated for the final survey to “**How many of the assigned videos have you watched?**” with options of ‘All’, ‘Most’, ‘Some’, or ‘None’. All students who selected ‘Some’, ‘Most’, or ‘All’ would be directed to answer the follow-up items about how they interacted with the videos. Students who responded ‘None’ would be directed to a follow-up item asking if they have ever accessed them, with appropriate follow-ups based on their response.

The 133 participants who responded that they did regularly watch the videos were presented with the first follow-up question “**How do you watch them?**”. Open-ended responses to this question generated several response categories: From start to finish (43%), pausing/rewinding (13%), playing at faster speed (4%), skipping around (15%), watch multiple times (6%), take notes (13%), and work problems (3%). For the responses (n = 133) from the second follow-up question “**What do you do when watching them?**”, the majority of students responded that they take notes (77%), while fewer reported just focusing on the video (32%) or working practice problems (25%). The response categories for these two follow-up questions were similar to those from the focus groups, therefore, each category was retained and multi-response options were generated for the final survey version.

The next question focused on when students typically watched the videos for the first time (as opposed to re-watching), with options worded around the timing of the class itself. While students at each institution used many of the provided categories, the ‘Other’ option was selected quite frequently, 15% each for Courses One and Two and 60% for Course Five students. Within the textbox provided, students’ explanations of when they watched were typically based on aspects such as when topics/material was being covered in class or around doing homework/studying for exams. These sentiments were also noted in the second round of focus group responses. Therefore, the response options for the final survey were modified to reflect these types of timings for viewing.

Next, the pilot survey included the ‘Yes’/’No’ item, “**Do you ever re-watch [the videos]?**”. Of the 143 responses, 64% noted that they did re-watch the videos. These students were administered the follow-up question “**When do you typically re-watch them?**”, reporting that they re-watched when studying (77%), doing homework (46%), or at the end of the week (13%). The 52 students who reported that they did not re-watch were asked why they did not and reported that they watched other videos instead (42%), found the videos were not useful the first time, or when selecting the ‘Other’ response (31%), wrote that they relied on the notes they wrote when watching the videos for first time. When the initial ‘Yes’/’No’ item was asked to students in the second-round focus groups, such as that they had not re-watched an entire video or that they would re-watch portions of the videos that they found confusing or contained a topic on which they needed clarification. Therefore, this item was modified in the final survey to read, “**Have you ever watched a video (or part of a video) more than once?**”

In addition to the items related to how the students interacted with the videos, the last items on the pilot survey and in the focus groups were about students’ perceptions of the videos. Students were asked “**Do you find the videos helpful for learning the course material?**” and what they liked and/or disliked about the videos. While the responses to these open-ended items were coded separately, their resulting response categories were similar. The categories of what students found helpful about the videos (i.e., ability to pace watching, easy to understand, reinforce material) map directly onto what students reported that they liked about the videos. Similar overlaps were found between why they reported that the videos were not helpful (i.e., not engaging/can't focus, too long or no time to watch, prefer other methods) and their reported dislikes. In the focus groups, students' likes and dislikes were expressed when responding to the questions about what was helpful and not helpful, often no additional insights were provided when they were asked about likes and dislikes. These outcomes revealed that both sets of items were not needed on the final survey. Students' reported likes/dislikes about the videos were combined with the themes that arose from what students found helpful/not helpful about the videos when generating response categories for the final survey items.

Supplementary Tables

Table A.5. Response percentages, by course, to survey item "Why have you not watched all of them [assigned videos]?"

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	57	29	117	109	8
[a]Why have you not watched all of them? (Select all that apply)					
Response *categories* and options	Percentage of students selecting an option				
General excuses					
Not enough time	53	45	48	32	50
I forget to watch them sometimes	47	45	32	39	25
They are too long***	16	7	15	35[b]	25
Not helpful					
They do not help my learning***	2	17	9	27	13
I only watch them when I need clarification on course material***	75[c]	34	39	29	38
I prefer other…					
videos than the ones my instructor posts	7	7	5	14	13
learning resources	26	24	23	35	13

[a]Survey item presented if 'Most' or 'Some' response options were selected for initial question, shown in Figure 4.3.

***$p < 0.001$. [b]Significant pairwise comparisons ($p < 0.05$) between Course Four and Courses Two (w = 0.25) and Three (w = 0.23). [c] Significant pairwise comparisons ($p < 0.05$) between Course One and Courses Two, Three, and Four.

Table A.6. Response percentages, by course, to survey item “Why have you never re-watched the videos?”

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	6	7	40	44	16
[a]Why have you never re-watched the videos? (Select all that apply)					
Response options	Percentage of student responses to item				
I refer to the notes I take the first time I watch**	33	57	70	59	100
I watch other videos to get a different perspective than the ones posted***	67	29	3	23	6
I typically understand the material after watching just once	33	14	30	36	56
They typically don’t match well to the quiz or exam material	0	29	13	20	0

[a]Survey item presented to everyone who selected ‘No’ to question in Figure 4.4. ** p < 0.01. *** p < 0.001.

Table A.7. Response percentages, by course, to survey item “When do you re-watch the videos?”

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	58	26	193	232	42
[a]When do you re-watch the videos? (Select all that apply)					
Response *categories* and options	Percentage of student responses to item				
I re-watch parts of a video when…					
…I have missed something the first time	78	65	78	72	76
…I need clarification at a later time (e.g., for homework or when completing a lab)*	88	69	81	72	79
…studying for an exam*	43	35	61	53	50
I re-watch an entire video…					
…to supplement my notes**	16	12	24	26	2
…when studying for an exam	28	15	28	32	14

[a]Survey item presented to everyone who selected ‘Yes’ to question in Figure 4.4. *p < 0.05. **p < 0.01.

Table A.8. Response percentages, by course, to survey item "When you watch the videos, how do you watch them?"

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	64	32	233	276	58
[a]When you watch the videos; how do you watch them? (Select all that apply)					
Response *categories* and options	Percentage of student responses to item				
Pacing of viewing					
From start-to-finish at normal speed**	66	59	58	49	34
From start-to-finish at varying (faster or slower) speed***	23	16	24	33	69
I pause and/or rewind while watching	73	63	61	64	52
I skip around or fast-forward through sections	23	16	27	20	19
Blocking of viewing					
I watch the assigned videos in one sitting***	25	13	32	55[b]	60[b]
I spread out watching the assigned videos throughout the day or week***	20	19	25	6	28

[a]Survey item presented to everyone except those who selected 'None' for initial question, shown in Figure 4.3. ** $p < 0.01$. *** $p < 0.001$. [b]Significant pairwise comparisons ($p < 0.05$) between Courses Four and Five and Courses One ($w = 0.24$ and $w = 0.36$, respectively), Two ($w = 0.26$ and $w = 0.46$, respectively), and Three ($w = 0.23$ and $w = 0.23$, respectively).

Table A.9. Response percentages, by course, to survey items "Were the videos NOT helpful to your learning? If so, in which ways were they NOT helpful?"

	Course One	**Course Two**	**Course Three**	**Course Four**	**Course Five**
Students, n	42	25	173	254	46
[a]Were the videos NOT helpful to your learning? If so, in which ways were they NOT helpful? (Select all that apply)					
Response *categories* and options	Percentage of student responses to item				
Do not meet learning expectations					
I am unable to ask questions or interact with the instructor***	40	40	28	62	59
They do not contain enough practice problems**	26	36	39	51	41
Not relevant to course					
They are too basic	14	20	21	26	13
The explanations are too difficult	7	16	10	12	2
They have a different focus than the class materials***	12	28	10	37	11
They contradict the class material***	0	0	1	9	4
Don't hold attention					
They are too long***	33	8	25	27	57
They are boring or not engaging***	17	12	29	40	17
The material presented is redundant	19	0	13	11	9
Poor quality/disorganized					
They are low quality making it difficult to see and/or hear***	5	0	0	27	9
They are confusing and/or disorganized**	0	4	7	14	4

[a]Survey items presented to everyone except those who selected 'None' for initial question, shown in Figure 4.3. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Appendix B: Supporting Information for Chapter 5

Individual scale analyses and modifications

For each individual scale, *a priori* single-factor models were investigated using Confirmatory Factor Analyses (CFA) on the full dataset. This step was undertaken to examine potential problematic items and to inform the need for modifications. After acceptable models were found for each scale, data-model fit was cross-validated at the institution level. Global and local data-model fit was assessed using the Comparative Fit Index (CFI) (Bentler, 1990), Root Mean Square Error of Approximation (RMSEA) (Steiger, 1980), and Standardized Root Mean Square Residual (SRMR) (Chen, 2007). For data-model fit, Hu and Bentler (1999) have suggested CFI values greater than 0.95, RMSEA values less than 0.06, and SRMR values less than 0.08 as evidence of good fit. However, McNeish and colleagues (2018) only suggest adherence to these aforementioned cutoff values for models that have items with similar properties to those in Hu and Bentler's (1999) simulation (i.e., all factor loadings approximately 0.7). McNeish et al. (2018) found that for models containing items with higher factor loadings (e.g., 0.9), that appropriate CFI values could be as low as 0.775 and RMSEA values could be as high as 0.20. This suggests that data could have appropriate data-model fit even when fit indices appear less ideal according to what Hu and Bentler (1999) originally found. Therefore, both item factor loadings and a range of fit indices were used when evaluating the data-model fit across all analyses in this study.

All initial models had poor RMSEA values. Table B.1 contains the summary data-model fit indices for the initial and final models using the full data sample. Individual scale modifications were made based upon modification indices and/or conceptual justifications (Wang & Wang, 2019). A discussion for each scale modification follows in the subsequent sections.

Table B.1. Fit Indices for initial and final versions of scales.

Initial Model[a]						Final Model[b]			
Measure		**CFI**	**SRMR**	**RMSEA**	**CI RMSEA**	**CFI**	**SRMR**	**RMSEA**	**CI RMSEA**
CSE	Pre	0.984	0.035	0.150	0.138-0.162	0.994	0.024	0.168	0.143-0.195
	Post	0.980	0.038	0.157	0.143-0.171	0.993	0.025	0.170	0.141-0.201
ASE	Pre	0.967	0.063	0.206	0.198-0.214	0.996	0.022	0.085	0.073-0.097
	Post	0.978	0.052	0.223	0.214-0.233	0.994	0.026	0.128	0.114-0.143
CON	Pre	0.959	0.052	0.138	0.130-0.147	0.987	0.030	0.105	0.093-0.117
	Post	0.949	0.058	0.161	0.152-0.171	0.988	0.031	0.106	0.092-0.121
TMT	Pre	0.940	0.070	0.143	0.135-0.152	0.982	0.036	0.149	0.133-0.166
	Post	0.928	0.081	0.162	0.153-0.172	0.977	0.043	0.183	0.164-0.202

[a]Initial models include all items. [b]Final models include a reduced set of items.

Chemistry Self-Efficacy (CSE)

Items 1 and 6 were removed to produce the final CSE measure. Item 1 was removed because it showed consistently high modification indices (correlated errors to other items). Item 6 was removed as it was deemed to not necessarily be specific to the lecture portion of each course in the sample.

Table B.2. Factor loadings for Chemistry Self-Efficacy scale models.

Item	Pre		Post	
	Initial	Final	Initial	Final
1. To what extent can you explain chemical laws and theories?	0.758	---	0.778	---
2. How well can you choose an appropriate formula to solve a chemistry problem?	0.812	0.784	0.775	0.745
3. How well can you describe the properties of elements by using the periodic table?	0.765	0.760	0.784	0.781
4. How well can you read the formulas of elements and compounds?	0.883	0.907	0.883	0.910
5. How well can you interpret chemical equations?	0.905	0.903	0.900	0.897
6. How well can you interpret graphs/charts related to chemistry?	0.754	---	0.745	---

Academic Self-Efficacy (ASE)

Items 2 and 8 were removed to produce the final ASE measure. Each was removed due to their consistently high modification indices. Additionally, each contained an aspect that may not have pertained to all courses (i.e., readings and assignments).

Table B.3. Factor loadings for Academic Self-Efficacy scale models.

Item	Pre		Post	
	Initial	Final	Initial	Final
1. I'm confident that I can understand the most complex material presented by the instructor in my courses.	0.821	0.729	0.928	0.834
2. I'm certain I can understand the most difficult material presented in the readings for my courses.	0.803	---	0.915	---
3. I believe I will receive excellent grades in my courses.	0.870	0.873	0.866	0.904
4. I'm confident I can understand the basic concepts taught in my courses.	0.801	0.824	0.800	0.819
5. I expect to do well in my courses.	0.878	0.896	0.837	0.857
6. Considering the difficulty of my courses, the instructor, and my skills, I think I will do well in my courses.	0.904	0.920	0.923	0.942
7. I'm certain I can master the skills being taught in my courses.	0.869	0.852	0.888	0.863
8. I'm confident I can do an excellent job on the assignments and tests in my courses.	0.911	---	0.926	---

Concentration (CON)

Items 7 and 8 were removed to produce the final CON measure. Item 7 was removed due its consistently low factor loadings. Item 8 was removed as it was a double-barreled item.

Table B.4. Factor loadings for Concentration scale models. REV indicates a reverse-coded item.

Item	Pre		Post	
	Initial	Final	Initial	Final
1. I concentrate fully when studying.	0.533	0.543	0.528	0.536
2. Because I don't listen carefully, I don't understand some course material. (REV)	0.598	0.569	0.638	0.618
3. I find it difficult to maintain my concentration while doing my coursework. (REV)	0.814	0.829	0.812	0.827
4. My mind wanders a lot when I study. (REV)	0.839	0.859	0.848	0.869
5. I find it hard to pay attention during lectures. (REV)	0.774	0.716	0.759	0.689
6. I am very easily distracted from my studies. (REV)	0.845	0.859	0.857	0.869
7. If I get distracted during class, I am able to refocus my attention.	0.247	---	0.110	---
8. I find that during lectures I think of other things and don't really listen to what is being said. (REV)	0.741	---	0.727	---

Time Management (TMT)

Items 2, 6, and 7 were removed to produce the final TMT measure. Each was removed due to their consistently low factor loadings. Additionally, correlated residuals were incorporated for items 5 and 8 based on their similarity in use of the word ‘cram’/‘cramming’.

Table B.5. Factor loadings for Time Management scale models.

Item	Pre		Post	
	Initial	Final	Initial	Final
1. I find it hard to stick to a study schedule.	0.688	0.697	0.704	0.705
2. When I decide to study, I set aside a specific length of time and stick to it.	0.346	---	0.286	---
3. When it comes to studying, procrastination is a problem for me.	0.844	0.865	0.862	0.870
4. I put off studying more than I should.	0.898	0.912	0.916	0.933
5. I spread out my study times so I do not have to "cram" for a test.	0.627	0.549	0.550	0.467
6. I do not have enough time to study because I spend too much time with my friends.	0.469	---	0.441	---
7. I set aside more time to study the subjects that are difficult for me.	0.376	---	0.195	---
8. I end up "cramming" for every test.	0.702	0.658	0.712	0.670

Final CFA models by institution

The RMSEA values were outside of the range as described by Hu and Bentler (1999), but are interpreted as being acceptable based on the findings and recommendations of McNeish and colleagues (2018). Within their simulation studies, McNeish and colleagues (2018) found that CFA models that included scales with excellent measurement quality (defined by high standardized factor loadings and McDonald’s omega values) showed a higher power to detect even trivial model misspecifications, thereby resulting in “seemingly unsatisfactory [data-model fit] values.” While they make a point to not recommend alternative acceptable values, they do note that under these conditions that SRMR values may exceed 0.14, RMSEA values may exceed 0.20, and that CFI values may fall below 0.775. Therefore, given that the majority of the factor loadings for our items were high (majority >0.70 for the final models, Tables B.2 – B.5) and that the McDonald's omega values of each scale were also high (all above 0.80, Table B.6), we believe that the data-model fit for each measure at each institution is acceptable (Table B.6).

Komperda and colleagues (2018) discuss various methods of estimating the single-administration reliability of scale data. If data from scale items do not fit parallel or tau-equivalence factor structures, alternatives to Cronbach's alpha are preferred (e.g., McDonald's omega). To assess the single-administration reliability of each scale, CFA models were therefore fit as congeneric with McDonald's omega values reported.

Table B.6. Data-model fit indices and single-administration reliability values (omega) for CFA final models by institution.

Scale	Time	Institution	df	χ^2	CFI	SRMR	RMSEA	90% CI	omega
CSE	Pre	Southeastern	2	30.368	0.995	0.024	0.160	0.113-0.213	0.91
		Western	2	9.245	0.995	0.023	0.131	0.055-0.222	0.86
		Northwestern	2	50.777	0.990	0.033	0.176	0.136-0.220	0.87
	Post	Southeastern	2	8.138	0.998	0.085	0.103	0.037-0.181	0.91
		Western	2	23.126	0.984	0.044	0.222	0.147-0.308	0.83
		Northwestern	2	45.929	0.994	0.033	0.217	0.109-0.345	0.88
ASE	Pre	Southeastern	9	56.371	0.996	0.025	0.098	0.074-0.123	0.95
		Western	9	36.783	0.993	0.033	0.122	0.082-0.164	0.92
		Northwestern	9	68.842	0.994	0.029	0.092	0.072-0.113	0.92
	Post	Southeastern	9	74.510	0.994	0.030	0.159	0.127-0.193	0.95
		Western	9	25.225	0.997	0.022	0.091	0.050-0.134	0.93
		Northwestern	9	113.979	0.998	0.018	0.114	0.086-0.145	0.93
CON	Pre	Southeastern	9	73.373	0.984	0.035	0.115	0.091-0.140	0.87
		Western	9	52.195	0.967	0.049	0.153	0.115-0.195	0.84
		Northwestern	9	106.381	0.986	0.033	0.099	0.078-0.121	0.86
	Post	Southeastern	9	62.035	0.985	0.041	0.143	0.111-0.178	0.89
		Western	9	58.064	0.974	0.053	0.160	0.122-0.200	0.86
		Northwestern	9	70.475	0.986	0.033	0.099	0.078-0.121	0.86
TMT	Pre	Southeastern	4	14.746	0.998	0.016	0.070	0.034-0.110	0.88
		Western	4	7.181	0.997	0.023	0.062	0.000-0.135	0.84
		Northwestern	4	17.901	0.996	0.019	0.067	0.037-0.099	0.84
	Post	Southeastern	4	3.952	1.000	0.011	0.000	0.000-0.088	0.88
		Western	4	8.811	0.996	0.022	0.075	0.000-0.144	0.82
		Northwestern	4	16.374	0.997	0.017	0.067	0.035-0.102	0.84

Descriptive statistics for each measure

Descriptive statistics for each scale were calculated using the *psych* package (version 1.9.12) in R (Table B.7). All observed means were calculated as the average of the individual items retained in the CFAs. Descriptive statistics by institution are shown in Table B.8. While there is evidence of non-normality in the data, the individual items are also ordinal in nature. Therefore, in all subsequent analyses, the WLSMV estimator was chosen to appropriately account for these data structures.

Table B.7. Descriptive statistics by scale and time point.

Scales	Time	Mean	Standard Deviation	Skew	Kurtosis
Chemistry Self-Efficacy (CSE)	Pre	3.17	0.90	-0.32	2.83
	Post	3.64	0.81	-0.43	3.41
Academic Self-Efficacy (ASE)	Pre	3.93	0.80	-1.08	4.44
	Post	3.44	1.03	-0.44	2.41
Concentration (CON)	Pre	3.23	0.87	0.06	2.44
	Post	3.20	0.88	0.13	2.43
Time Management (TMT)	Pre	2.89	0.92	0.19	2.53
	Post	2.81	0.91	0.25	2.64

Table B.8. Descriptive statistics for measures by institution.

			Aggregated	Southeastern	Western	Northwestern
CSE	Pre	n	1,559	554	211	794
		M (SD)	3.180 (0.789)	2.883 (0.925)	3.487 (0.787)	3.307 (0.804)
		Sk	-0.320	-0.048	-0.543	-0.350
		Ku	2.920	2.688	3.973	3.072
	Post	n	1,216	293	217	706
		M (SD)	3.638 (0.789)	3.892 (0.815)	3.610 (0.670)	3.540 (0.790)
		Sk	-0.381	-0.428	0.208	-0.547
		Ku	3.359	2.902	2.309	3.659
ASE	Pre	n	1,562	554	211	797
		M (SD)	3.899 (0.817)	4.148 (0.793)	3.748 (0.816)	3.765 (0.793)
		Sk	-1.021	-1.671	-0.692	-0.836
		Ku	4.162	6.703	3.491	3.607
	Post	n	1,221	293	219	709
		M (SD)	3.410 (1.046)	3.975 (0.940)	3.474 (0.922)	3.157 (1.030)
		Sk	-0.434	-1.022	-0.409	-0.300
		Ku	2.469	3.853	2.848	2.263
CON	Pre	n	1,562	554	211	797
		M (SD)	3.197 (0.856)	3.389 (0.874)	2.952 (0.826)	3.128 (0.825)
		Sk	0.066	-0.163	0.380	0.118
		Ku	2.441	2.538	2.694	2.471
	Post	n	1,220	293	219	708
		M (SD)	3.216 (0.874)	3.373 (0.958)	2.846 (0.790)	3.265 (0.830)
		Sk	0.089	-0.010	0.215	0.034
		Ku	2.417	2.158	2.894	2.417
TMT	Pre	n	1,560	554	210	796
		M (SD)	2.865 (0.906)	2.950 (0.968)	2.781 (0.866)	2.828 (0.753)
		Sk	0.211	0.060	0.283	0.288
		Ku	2.591	2.419	3.116	2.612
	Post	n	1,220	293	218	709
		M (SD)	2.823 (0.890)	2.950 (1.018)	2.679 (0.819)	2.814 (0.847)
		Sk	0.263	0.222	0.467	0.145
		Ku	2.771	2.236	3.162	2.816

Accounting for unused response categories

In this study some of the measurement invariance and structural means modeling analyzes required comparing institutional data. However, for some institutions, the data collected for the CSE scale did not include responses spanning the entire response scale. When conducting comparisons, response category thresholds cannot easily be removed from only a subset of the data. Therefore, a method was developed to account for these missing response categories when comparisons between institutions were conducted. To conduct these analyses, at least one response is required in each response category for each item in the scale at the institution level. Therefore, a single ‘dummy participant’ with a response pattern that included the missing response category was added to the data set as needed. For the remaining items on the scale, where students had used the full response scale, the dummy response pattern included the average value for that item. For example, a ‘dummy’ response pattern was added for the Western institution, which accounted for no students responding “strongly disagree” to Items 2, 4, and 5 on the post CSE scale. The effect of adding dummy response patterns was examined by first evaluating data-model fit statistics and latent means for only the institutions that included full response scale data (i.e., no dummy responses present). Then trial dummy response patterns were added to these institutions and the measurement invariance and latent means analysis was again examined and compared to the previous analysis that included only real data. The results from the ‘real’ and ‘real & dummy’ data were similar and no significant differences were detected. This suggested that adding these response patterns, in minimal quantities, for the institutions with missing response data would not significantly affect the outcome of the results. Based on this, a single dummy response pattern was added to the institution that was missing at least one response category, as needed.

Establishing measurement invariance

The focus of these analyses was to establish scalar invariance of the four measures, which involves setting factor loading and threshold response patterns equal across comparator groups. To address this, the CFI, SRMR, and RMSEA data-model fit values for both the configural and scalar models were evaluated and also compared, based upon the recommendations by Chen (2007) as well as Jin (2020). With respect to some items on the CSE scale, some of the institution’s data did not contain response for all categories (i.e., no students responded “strongly disagree”), which resulted in a different number of thresholds for these institutions. Since thresholds cannot be easily removed from only a subset of institutions, a ‘dummy’ response pattern was added. Finally, pre to post longitudinal invariance was assessed for all measures using the full sample. Syntax for the longitudinal invariance models was generated using the *measEq.syntax* feature within the *semTools* package (Version 0.5-3) in R.

While it is also recommended to evaluate the change in the fit indices when moving from the configural to the scalar model, this is not a requirement to establish invariance (Rocabado et al., 2020). We do, however, report the change values for each measurement invariance evaluation (Tables B.9-B.12) and note that while most fall into the recommended ranges (Chen, 2007; Jin, 2020), the RMSEA values of the *by gender* (Table B.11) and *by URM status* (Table 12) regularly fall outside of the range. However, given the model sensitivity issues noted by McNeish and colleagues (2018), we may not be able to use the recommended change values to conclude if the change is acceptable or unacceptable in a definitive fashion. Therefore, we support the invariance of each *by group* comparison based on the acceptable data-model fit to each of the scalar models.

Table B.9. Data-model fit indices for scalar longitudinal measurement invariance.

Model	df	χ^2	*p*-Value	CFI	SRMR	RMSEA	Δdf	$\Delta\chi^2$	ΔCFI	ΔSRMR	ΔRMSEA
CSE											
Configural	15	85.338	< 0.001	0.995	0.025	0.069	---	---	---	---	---
Scalar	26	99.550	0.057	0.995	0.025	0.054	11	19.249	0.000	0.000	0.015
ASE											
Configural	47	241.689	< 0.001	0.994	0.028	0.065	---	---	---	---	---
Scalar	64	351.229	< 0.001	0.990	0.029	0.068	17	116.39	0.004	-0.001	-0.003
CON											
Configural	47	275.771	< 0.001	0.983	0.038	0.071	---	---	---	---	---
Scalar	64	280.482	0.722	0.984	0.038	0.059	17	13.028	-0.001	0.000	0.012
TMT											
Configural	27	198.411	< 0.001	0.985	0.038	0.081	---	---	---	---	---
Scalar	41	215.803	0.036	0.984	0.038	0.066	14	24.811	0.001	0.000	0.015

Table B.10. Data-model fit indices for scalar measurement invariance by institution[a].

Model	df	χ^2	*p*-Value	CFI	SRMR	RMSEA	Δdf	$\Delta\chi^2$	ΔCFI	ΔSRMR	ΔRMSEA
CSE-Pre											
Configural	6	95.430	< 0.001	0.993	0.028	0.170	---	---	---	---	---
Scalar	34	167.886	0.001	0.989	0.030	0.087	28	58.361	-0.004	0.002	-0.083
CSE-Post											
Configural	6	76.827	< 0.001	0.993	0.028	0.172	---	---	---	---	---
Scalar	34	90.356	0.142	0.994	0.029	0.064	28	36.026	0.001	0.001	-0.108
ASE-Pre											
Configural	27	163.420	< 0.001	0.995	0.028	0.099	---	---	---	---	---
Scalar	71	290.979	0.001	0.992	0.028	0.078	44	91.403	-0.003	0.000	-0.021
ASE-Post											
Configural	27	213.658	< 0.001	0.994	0.030	0.131	---	---	---	---	---
Scalar	71	307.834	0.001	0.992	0.030	0.091	44	86.993	0.002	0.000	-0.040
CON-Pre											
Configural	27	227.406	< 0.001	0.981	0.039	0.121	---	---	---	---	---
Scalar	71	270.053	0.003	0.981	0.040	0.074	44	73.876	0.000	0.001	-0.047
CON-Post											
Configural	27	190.742	< 0.001	0.983	0.038	0.123	---	---	---	---	---
Scalar	71	283.853	0.001	0.978	0.039	0.087	44	92.904	-0.005	0.001	-0.036
TMT-Pre											
Configural	12	39.055	< 0.001	0.997	0.018	0.066	---	---	---	---	---
Scalar	48	99.225	0.061	0.994	0.021	0.046	36	49.933	-0.003	0.003	-0.020
TMT-Post											
Configural	12	29.444	< 0.001	0.998	0.017	0.060	---	---	---	---	---
Scalar	48	96.164	0.038	0.995	0.026	0.050	36	53.404	-0.003	0.009	-0.010

[a] Southeastern institution used as the reference category

Table B.11. Data-model fit indices for scalar measurement invariance by gender.[a] Values in italics outside of the recommended range noted by Chen (2007) and by Jin (2020).

Model	**df**	**χ^2**	***p*-Value**	**CFI**	**SRMR**	**RMSEA**	**Δdf**	**$\Delta\chi^2$**	**ΔCFI**	**ΔSRMR**	**ΔRMSEA**
CSE-Pre											
Configural	4	81.733	< 0.001	0.994	0.026	0.159	---	---	---	---	---
Scalar	18	79.652	0.702	0.995	0.026	0.067	14	10.799	-0.001	0.000	*0.092*
CSE-Post											
Configural	4	72.815	< 0.001	0.992	0.027	0.178	---	---	---	---	---
Scalar	18	94.682	0.005	0.991	0.028	0.089	14	31.552	0.001	-0.001	*0.089*
ASE-Pre											
Configural	18	129.028	< 0.001	0.996	0.025	0.090	---	---	---	---	---
Scalar	40	166.346	0.006	0.995	0.025	0.064	22	42.240	0.001	0.000	*0.026*
ASE-Post											
Configural	18	171.487	< 0.001	0.994	0.027	0.125	---	---	---	---	---
Scalar	40	204.986	0.003	0.994	0.027	0.087	22	44.416	0.000	0.000	*0.038*
CON-Pre											
Configural	18	206.902	< 0.001	0.982	0.036	0.118	---	---	---	---	---
Scalar	40	208.585	0.002	0.984	0.037	0.074	22	45.217	-0.002	-0.001	*0.044*
CON-Post											
Configural	18	165.400	< 0.001	0.984	0.036	0.123	---	---	---	---	---
Scalar	40	162.922	0.032	0.986	0.037	0.076	22	35.730	-0.002	-0.001	*0.047*
TMT-Pre											
Configural	8	36.976	< 0.001	0.997	0.017	0.069	---	---	---	---	---
Scalar	26	60.562	0.080	0.996	0.019	0.042	18	26.940	0.001	-0.002	*0.027*
TMT-Post											
Configural	8	33.507	< 0.001	0.997	0.018	0.077	---	---	---	---	---
Scalar	26	71.180	0.008	0.994	0.023	0.057	18	35.472	0.003	-0.005	*0.020*

[a]Male was used as the reference category.

Table B.12. Data-model fit indices for scalar measurement invariance by URM status.[a] Values in italics outside of the recommended range noted by Chen (2007) and by Jin (2020).

Model	**df**	χ^2	***p*-Value**	**CFI**	**SRMR**	**RMSEA**	**Δdf**	**Δ**χ^2	**ΔCFI**	**ΔSRMR**	**ΔRMSEA**
CSE-Pre											
Configural	4	82.973	< 0.001	0.994	0.026	0.160	---	---	---	---	---
Scalar	18	120.792	< 0.001	0.992	0.026	0.086	14	38.066	0.002	0.000	*0.074*
CSE-Post											
Configural	4	84.369	< 0.001	0.991	0.028	0.183	---	---	---	---	---
Scalar	18	101.540	0.002	0.991	0.029	0.088	14	33.921	0.000	-0.001	*0.095*
ASE-Pre											
Configural	18	150.359	< 0.001	0.995	0.027	0.098	---	---	---	---	---
Scalar	40	180.524	0.013	0.995	0.027	0.067	22	39.180	0.000	0.000	*0.031*
ASE-Post											
Configural	18	211.726	< 0.001	0.994	0.028	0.133	---	---	---	---	---
Scalar	40	241.346	0.002	0.993	0.028	0.091	22	45.287	0.001	0.000	*0.042*
CON-Pre											
Configural	18	208.937	< 0.001	0.982	0.036	0.118	---	---	---	---	---
Scalar	40	209.401	0.003	0.984	0.037	0.074	22	44.352	-0.002	-0.001	*0.044*
CON-Post											
Configural	18	197.575	< 0.001	0.982	0.036	0.129	---	---	---	---	---
Scalar	40	196.593	0.004	0.984	0.037	0.081	22	43.459	-0.002	-0.001	*0.048*
TMT-Pre											
Configural	8	39.812	< 0.001	0.996	0.017	0.072	---	---	---	---	---
Scalar	26	60.267	0.115	0.996	0.019	0.041	18	25.371	0.000	-0.002	*0.031*
TMT-Post											
Configural	8	31.170	< 0.001	0.997	0.016	0.070	---	---	---	---	---
Scalar	26	41.403	0.391	0.998	0.017	0.031	18	19.022	-0.001	-0.001	*0.039*

[a]non-URM was used as the reference category.

Supplemental structured means modeling tables

Table B.13 shows the results of the pre to post latent mean differences for the TMT and CON factors for each institution. As most of the differences were not significant, and the significant differences only represented small effect sizes, the decision was made to use the pre TMT and pre CON factors as controls in our larger CSE and ASE post comparisons between institutions.

Table B.13. Pre to post latent mean differences for each institution. Bolded values indicate the difference was statistically significant ($p < 0.05$).

Scale	Institution	Responses, n	Observed Pre Score[a]	Pre to Post Latent Mean Difference (Effect Size)
Time Management (TMT)	Southeastern	261	2.98	-0.03 (0.03)
	Western	162	2.83	**-0.16** (0.18)
	Northwestern	547	2.88	-0.01 (0.01)
Concentration (CON)	Southeastern	258	3.42	-0.04 (0.05)
	Western	163	3.02	-0.09 (0.15)
	Northwestern	547	3.15	**0.12** (0.19)

[a] Observed pre scale scores were calculated as an unweighted average of the items included in the final version of each scale.

Table B.14. Sample size of matched data set by gender and URM status shown by institution.

	Southeastern	Western	Northwestern	Aggregated
Male, n (%)[a]	97 (37)	62 (37)	185 (33)	344 (34)
Female, n (%)[a]	168 (63)	106 (63)	374 (67)	648 (66)
non-URM, n (%)[a]	26 (10)	81 (49)	417 (75)	524 (53)
URM, n (%)[a]	238 (90)	86 (51)	139 (25)	463 (47)

[a] Percentage of group responses within each data set

Table B.15 shows the pre to post latent mean differences for all four factors based on aggregated male and female groups. Results indicated that both male and female groups had an increase in CSE and decrease in ASE over the term. Only nonsignificant to small effects were seen for pre to post differences for TMT and CON.

Table B.15. Pre to post latent mean differences for male and female groups. Bolded values indicate the difference was statistically significant ($p < 0.05$).

Scale	Group	Responses, n	Observed Pre Score[a]	Pre to Post Latent Mean Difference (Effect Size)
Chemistry Self-efficacy (CSE)	Male	335	3.28	**0.62** (0.52)
	Female	632	3.23	**0.63** (0.53)
Academic Self-efficacy (ASE)	Male	331	3.94	**-0.30** (0.27)
	Female	637	3.85	**-0.62** (0.49)
Time Management (TMT)	Male	331	2.84	-0.03 (0.03)
	Female	623	2.93	-0.04 (0.04)
Concentration (CON)	Male	324	3.23	0.00 (0.01)
	Female	628	3.19	**0.07** (0.10)

[a] Observed pre scale scores were calculated as an unweighted average of the items included in the final version of each scale.

Table B.16 includes the post CSE differences between demographic groups from the aggregated data set. Comparisons between non-URM and URM groups showed that URM students had lower pre CSE compared to non-URM students but a higher post CSE when the pre latent means are controlled for. No differences were found between male and female groups.

Table B.16. Pairwise post chemistry self-efficacy (CSE) latent mean differences by demographic group with pre CSE, TMT, and CON factors as covariates. Each comparison is between two groups (i.e., non-URM vs. URM and male vs. female) while accounting for the pre latent means. Bolded values indicate the difference was statistically significant ($p < 0.05$).

Reference Group	Comparison Group	Pre Latent Mean Differences (Effect Size)		Post CSE Latent Mean Difference (Effect Size)
non-URM (n = 492)	URM (n = 444)	CSE	**-0.40** (0.28)	**0.60** (0.60)
		TMT	-0.03 (0.03)	
		CON	0.11 (0.14)	
Male (n = 324)	Female (n = 637)	CSE	-0.08 (0.06)	-0.06 (0.06)
		TMT	0.10 (0.08)	
		CON	-0.03 (0.05)	

Table B.17 shows the pre to post latent mean differences for the four factors based on aggregated non-URM and URM groups. Results indicated that both non-URM and URM groups had an increase in CSE and decrease in ASE over the term. Only nonsignificant to small effects were seen for pre to post differences for TMT and CON.

Table B.17. Pre to post latent mean differences for non-URM and URM groups. Bolded values indicate the difference was statistically significant (p < 0.05).

Scale	Group	Responses, n	Observed Pre Score[a]	Pre to Post Latent Mean Difference (Effect Size)
Chemistry Self-efficacy (CSE)	non-URM	509	3.35	**0.34** (0.27)
	URM	453	3.13	**0.89** (0.79)
Academic Self-efficacy (ASE)	non-URM	511	3.76	**-0.58** (0.51)
	URM	452	4.01	**-0.40** (0.31)
Time Management (TMT)	non-URM	502	2.91	-0.01 (0.01)
	URM	447	2.88	-0.05 (0.06)
Concentration (CON)	non-URM	502	3.16	**0.09** (0.14)
	URM	445	3.25	-0.01 (0.01)

[a] Observed pre scale scores were calculated as an unweighted average of the items included in the final version of each scale.

Table B.18 includes the post ASE differences between demographic groups from the aggregated data set. Comparisons between non-URM and URM groups showed that URM students had higher pre and post ASE compared to non-URM students. Comparisons between male and female groups found that female students had lower post ASE compared to male students.

Table B.18. Pairwise post academic self-efficacy (ASE) latent mean differences by demographic group with pre ASE, TMT, and CON factors as covariates. Each comparison is between two groups (i.e., non-URM vs. URM and male vs. female) while accounting for the pre latent means. Bolded values indicate the difference was statistically significant (p < 0.05).

Reference Group	Comparison Group	Pre Latent Mean Differences (Effect Size)		Post ASE Latent Mean Difference (Effect Size)
non-URM (n = 494)	URM (n = 443)	ASE	**0.38** (0.37)	**0.36** (0.28)
		TMT	-0.08 (0.06)	
		CON	0.10 (0.12)	
Male (n = 321)	Female (n = 616)	ASE	-0.13 (0.14)	**-0.46** (0.33)
		TMT	0.10 (0.09)	
		CON	-0.04 (0.05)	

Observations of the face-to-face environments

Observations of the face-to-face (F2F) environments were conducted at each of the institutions (Naibert et al., 2020) using the Classroom Observation Protocol in Undergraduate STEM (COPUS) (Smith et al., 2013). The protocol includes codes that are documented each time the instructor or student participates in a different behavior during the F2F time. For this study, only student codes for "groupwork" and "questioning" were examined (Figure B.1), the [1]full COPUS timelines can be found in our prior study (Naibert et al., 2020). "Groupwork" contains the COPUS codes for working on a worksheet activity (WG), discussing clicker questions (CG), and working on other groupwork (OG). "Questioning" includes COPUS codes for answering questions posed by the instructor (AnQ) and asking a question (SQ). Each code is documented if it occurs at least once within a 2-minute time-block and multiple codes can be coded for each of the time-blocks. Thus, the percentages may add up to more than 100%.

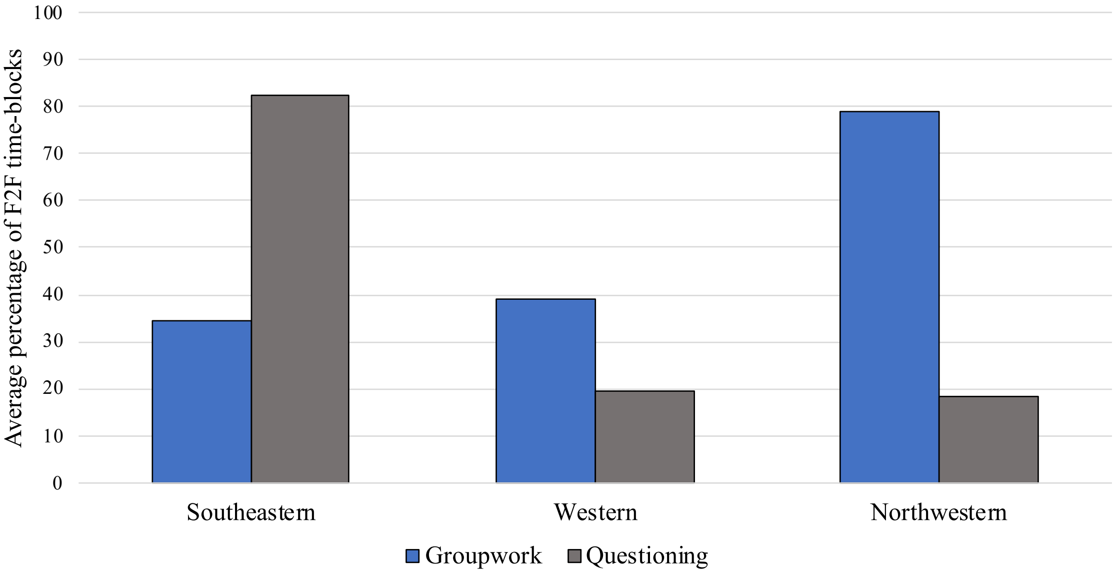


Figure B.1. Average percentage of F2F time-blocks students were observed participating in "groupwork" (blue) or "questioning" (gray) at each institution.

[1] In the prior study (Naibert et al., 2020), 'Course One' was from the Southeastern institution, 'Course Three' was from the Western institution, and 'Course Four' was from the Northwestern institution.

Appendix C: Supporting Information for Chapter 6

Initial analysis of the original ASPECT in our environments

Overview

Two PSU courses were surveyed over two separate terms within the same academic year. The first course was the first term of a three-term general chemistry series and the second course was the third term of the three-term introductory biology series. Students at PSU generally take these course series concurrently, providing a similar range of student demographics for this study (Table C.1). While different courses, terms, and active learning environments were included in this study, the data were only intended to evaluate the validity evidence of the measures administered at any one time, not to make comparisons on the outcomes between the different courses.

Table C.1. Self-reported study sample demographics.

	Course	
Category	**General Chemistry (%)** **n = 263[a]**	**Introductory Biology (%)** **n = 148[a]**
Gender		
Female	63	70
Male	35	27
Other	2	3
Race/ethnicity		
Non-URM	74	73
URM	26	27
Age bracket (years)		
18-22 (traditional)	69	61
23+ (nontraditional)	31	39
Major		
Biology	31	51
Chemistry	5	7
Other STEM	45	32
Non-STEM	12	5
None	7	5
University status		
Postbaccalaureate	14	10
Undergraduate	86	90
Transfer status		
Transfer from 2-year college	36	47

[a]only students who completed the demographics section of the survey are included

General chemistry – POGIL worksheet activities

The ASPECT was first evaluated in a first-term general chemistry course at PSU. The active learning strategy utilized in the general chemistry course was process-oriented guided inquiry learning (POGIL) worksheet activities (The POGIL Project). Specifically, students were given a worksheet containing key questions, exercises, and problems built around a model for a single chemistry topic. Each worksheet was designed to guide students through exploration of the focal topic to construct understanding. Students were encouraged, but not required, to work on the worksheet in groups of 2 – 4 and often chose to work with students that sat near them. The instructor of the class, a graduate teaching assistant (TA), and multiple undergraduate learning assistants (LAs) continuously moved throughout the room during the activity and assisted any groups that asked for help or appeared to be having difficulty working through the content. A classroom response system (i.e., clickers) was used during the activity to gauge students' understanding throughout; however, as clickers were also implemented during 'normal' lecture days, the main difference on these days was the presence of the worksheet. This course contained two sections taught by different instructors, however, the worksheet and facilitation of the groupwork was identical across sections.

Since the original ASPECT development resulted in a final 3-factor structure for the data (Wiggins et al., 2017), data from this study could be evaluated using confirmatory factor analysis (CFA). CFA allows for an *a priori* hypothesis of the factor structure to be assessed. If the analysis displayed good data-model fit, then there would be evidence to support the structural validity of the data collected in this environment. CFAs were completed using the statistical program R (version 3.6.2) with the package 'lavaan' (version 0.6-5). Maximum likelihood with the Satorra-Bentler adjustment and robust standard errors were used to account for any non-normality of the data (Satorra & Bentler, 1988). Since the survey was developed to measure students' perceptions of the active learning environments, all factors were correlated with each other in the analysis. Data-model fit was evaluated based on four fit indices and standard suggested values for good fit: CFI ≥ 0.95, TLI ≥ 0.95, RMSEA ≤ 0.06, and SRMR ≤ 0.08 (Hu & Bentler, 1999). Additionally, modification indices (MIs) were analyzed. MIs provide information about model specifications that can be made to improve data-model fit. As the MIs indicate an expected change in chi-square when the modification is included, the effect size of the suggested modifications were determined using Cohen's *w* (Cohen, 1992).

A total of 373 survey responses were collected across both general chemistry class sections. This was a 54% response rate based on the week-1 enrollment (N = 696). After cleaning, a sample size of 290 responses remained, with most of the removed responses due to students selecting the incorrect check item (*i.e., somewhat agree*). Although data for all 19 survey items were collected (see Table 6.1 in Chapter 6), only data from the 16 items that contributed to the original ASPECT factor structure (Wiggins et al., 2017) were analyzed using CFA based on the *a priori* model (Figure C.1). This included 3 items related to 'personal effort, 9 items related to 'value of the activity', and 4 items related to 'instructor contribution'. Factor loadings for the items can be found in Table C.2. The resulting fit statistics are included in Table C.3 and indicated overall poor data-model fit. To determine if there were possible improvements that could be made to the model, MIs were examined. Suggested modifications that had a medium effect size

(~0.3) or larger based on Cohen's *w* (Cohen, 1992) were flagged and evaluated. The high MI values suggested error correlations between three item pairs (Table C.4). Upon review, it was noted that the items in these pairs had related meanings, which can often lead to dependent error terms.

The presence of this error structure brought to light some concerns with using the original survey in our environment. The first concern was related to the suggested error correlation for two of the items, "Explaining the material to my group improved my understanding of it" and "I made a valuable contribution to my group today". These items are from two different factors; 'value of group activity' and 'personal effort', respectively (Wiggins et al., 2017). The high MI for the error correlation between these items indicated that there may have been a source of variance contributing to both of these items that was not accounted for by either the 'personal effort' or 'value of group activity' factors. While the two other suggested item error correlations were between items on the 'value of group activity' factor only, they provided further evidence that an additional source of variance may have been unaccounted for by the original three factors. Although simply including error structures can improve data-model fit, adding them can also result in excluding the influence of a relevant variable (Hermida, 2015). Therefore, upon further inspection, it was noted that all five of these items (Table C.4) contained wording centered around group function. Additionally, a fourth 'group-related' factor was noted as a possibility by the ASPECT developers; however, it was not fully realized in their analysis and a three-factor structure was found to be ideal (Wiggins et al., 2017).

These results suggested the influence of a fourth group-related factor in our environment, although simply incorporating this additional factor into the current three-factor model was not ideal as only two items would have remained to assess 'personal effort'. Only having two items load on a single factor would mean that the reliability estimate for that factor could not be appropriately determined (Komperda, Pentecost, et al., 2018) as the model for 'personal effort' would be under-identified (Brown, 2015). Because of these concerns, the data collected in this environment with the original ASPECT survey was found to lack sufficient evidence of internal structure validity and led to the creation of the modified ASPECT versions.

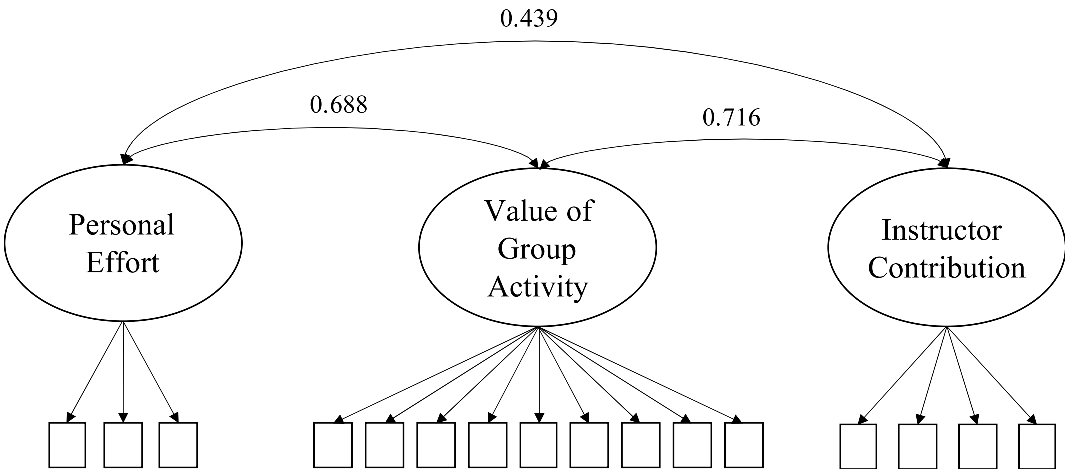


Figure C.1. CFA with the original ASPECT factor structure. Values obtained for factor correlations are included.

Table C.2. Factor loadings for the items with the original ASPECT factor structure.

Factor	Item	Factor Loading
Personal effort	#1 I made a valuable contribution to my group today.	0.720
	#2 I was focused during today's group activity.	0.841
	#3 I worked hard during today's group activity.	0.802
Value of group activity	#4 Explaining the material to my group improved my understanding of it.	0.773
	#5 Having the material explained to me by my group members improved my understanding of the material.	0.720
	#6 Group discussion during the activity contributed to my understanding of the course material.	0.821
	#7 Overall, the other members of my group made valuable contributions during the group activity.	0.679
	#8 I had fun during today's group activity.	0.771
	#9 I would prefer to take a class that includes this [topic] group activity over one that does not include this [topic] activity.	0.508
	#10 I am confident in my understanding of the material presented during today's group activity.	0.564
	#11 The group activity increased my understanding of the course material.	0.775
	#12 The group activity stimulated my interest in the course material.	0.762
Instructor contribution	#13 The instructor's enthusiasm made me more interested in the group activity.	0.763
	#14 The instructor put a good deal of effort into my learning for today's class.	0.892
	#15 The instructor seemed prepared for the group activity.	0.682
	#16 The instructor and TAs were available to answer questions during the group activity.	0.497

Table C.3. CFA fit indices with the original ASPECT factor structure (n = 290). Fit results that met the appropriate criteria for good fit are bolded.

Model	**χ^2**	**df**	**p-value**	**CFI ($\geq$ 0.95)[a]**	**TLI ($\geq$ 0.95)[a]**	**RMSEA ($\leq$ 0.06)[a]**	**RMSEA 90% CI**	**SRMR ($\leq$ 0.08)[a]**
ASPECT	217.315	101	< 0.001	0.926	0.912	0.078	0.063 – 0.092	**0.061**

[a]Suggested fit criteria based on Hu and Bentler (1999)

Table C.4. Modification indices with Cohen's *w* > 0.3 (medium effect size).

Suggested Modification	**Items[a]**	**Modification Index (MI)**	**Effect Size (Cohen's *w*)**
Item error correlation	Explaining the material to my group improved my understanding of it.	55.483	0.44
	I made a valuable contribution to my group today.		
Item error correlation	Having the material explained to me by my group members improved my understanding of the material.	36.080	0.35
	Group discussion during the activity contributed to my understanding of the course material.		
Item error correlation	Having the material explained to me by my group members improved my understanding of the material.	28.315	0.31
	Overall, the other members of my group made valuable contributions during the group activity.		

[a]Items from the original ASPECT (Wiggins et al., 2017).

Introductory biology – DD activity and clicker days

Before exploring the factor structure of the modified ASPECT surveys through EFA, CFA was first used to assess the original ASPECT structure in the DD activity and clicker day environments using a similar process as was completed with the general chemistry data. The CFAs with mASPECT-DD and mASPECT-C items were completed using only those items related to the original ASPECT (Items 1 – 16, Table 6.1 in Chapter 6) and the *a priori* original three-factor ASPECT model (i.e., 'personal effort', 'value of group activity', and 'instructor contribution') (Figure C.2). Factor loadings for both survey versions are included in Table C.5. Fit results for data collected during the DD activity day with mASPECT-DD, as well as data collected during the clicker day with mASPECT-C, suggested poor data-model fit for both environments (Table C.6).

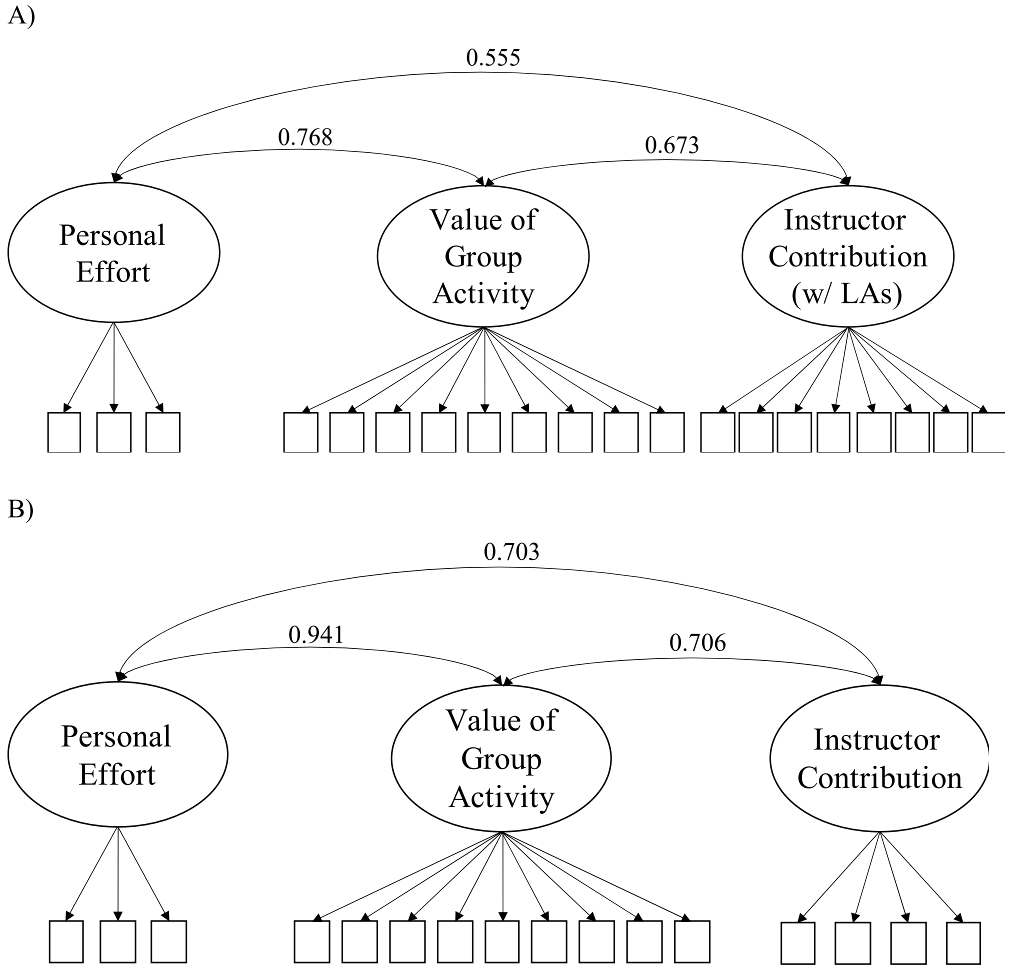


Figure C.2. Original ASPECT CFA models tested with data collected during the A) DD activity day and B) clicker day. Values obtained for the factor correlations are included.

Table C.5. Factor loadings for the DD activity and clicker day. Item wordings for the [DD activity day/clicker day] are included.

Factor	Item	Factor Loading	
		DD Activity	Clicker Day
Personal effort	#1 I was focused during today's class.	0.848	0.713
	#2 I worked hard during today's class.	0.691	0.694
	#3 I made valuable contributions when [working/having discussions] with other students during today's class.	0.555	0.595
Value of group activity	#4 Explaining the material to [my group members/other students] improved my understanding of it.	0.761	0.614
	#5 Having the material explained to me by [my group members/other students] improved my understanding of it.	0.787	0.518
	#6 [Working/Discussion] with other students during today's class contributed to my understanding of the material.	0.755	0.622
	#7 The students I [worked/had discussions] with made valuable contributions during today's class.	0.530	0.590
	#8 I had fun during today's class.	0.664	0.675
	#9 I would prefer to take a class that included today's [activity/clicker questions] over one that does not include [it/them].	0.620	0.311
	#10 I am confident in my understanding of the material presented during today's class.	0.399	0.427
	#11 Today's class increased my understanding of the material.	0.602	0.672
	#12 Today's class stimulated my interest in the course material.	0.689	0.654
Instructor contribution	#13 A The Professor/Teaching Assistant's enthusiasm made me more interested in today's class.	0.659	0.788
	#13 B The Learning Assistant's enthusiasm made me more interested in today's class.	0.702	---
	#14 A The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.	0.728	0.718
	#14 B The Learning Assistant put a good deal of effort into my learning for today's class.	0.765	---
	#15 A The Professor/Teaching Assistant seemed prepared for today's class.	0.733	0.684
	#15 B The Learning Assistant seemed prepared for today's class.	0.822	---
	#16 A The Professor/Teaching Assistant was available to answer questions during today's class.	0.719	0.676
	#16 B The Learning Assistant was available to answer questions during today's class.	0.802	---

Table C.6. CFA fit indices with the original ASPECT factor structure with mASPECT-DD and mASPECT-C Items 1 – 16 (see Table C.5).

Environment	Responses, n	χ^2	df	p-value	CFI (≥ 0.95)[a]	TLI (≥ 0.95)[a]	RMSEA (≤ 0.06)[a]	RMSEA 90% CI	SRMR (≤ 0.08)[a]
DD activity	149	406.517	167	<0.001	0.769	0.737	0.123	0.0108 – 0.0138	0.092
Clicker day	136	209.458	101	<0.001	0.795	0.756	0.115	0.093 – 0.137	0.099

[a]Suggested fit criteria based on Hu and Bentler (1999)

Survey modifications

Overview

Modifications were made to the original ASPECT survey to create two modified ASPECT (mASPECT) survey versions. These modifications included minor wording changes and the creation of new items based on the structure of the learning environments in the introductory biology class and the results from the general chemistry course. These two types of modifications are presented separately for clarity.

Modifications based on the learning environment

Because the introductory biology class included two different types of active learning environments, DD activity days and clicker days, modifications were made to the wording of the ASPECT items such that one version of most items could be administered in both environments.

Since clicker days were considered ‘normal’ classroom days, using the term “this group activity” would not make sense for that environment. As the DD activity took the entire class period on the days it was completed, changing the wording to “this class” would encompass the DD activity during those days, while also working for the clicker day version. Therefore, it was decided to reword items with the phrase “group activity” to “class”. For example, the original item “I was focused during today’s *group activity*” was changed to “I was focused during today’s *class*”. Although this change allowed for 11 out of the 19 items to be identical across environments, some of the original items included phrasing that was specific to a structured group activity, such as “group discussion” or “my group”. This wording matched the structure of the DD activity days. However, the clicker days were more flexible in that although students were strongly encouraged to discuss the questions with other students (i.e., ‘think-pair-share’), they were allowed to work with whomever they wanted and/or work alone. Thus, two versions of the remaining 8 items were created to account for the differences in the environments. For example, the original item “I made a valuable contribution to my group today” was changed to “I made a valuable contribution when *working with other students* during today’s class” for the DD activity, and to “I made a valuable contribution when *having discussions with other students* during today’s class”. The full set of changes can be seen in Table 6.1 in Chapter 6.

Additionally, all four ‘instructor contribution’ items were reworded to take into account the influence of the professor and graduate TA. These items were administered on both the DD activity and clicker days, as the professor and graduate TA were present for both. DD activity days also included the assistance of undergraduate LAs, whom were assigned to work with the same 4 – 5 small student groups during each DD activity day. Therefore, parallel items that differed only in distinguishing the LAs' influence from the professor/TA's influence were created. The LA-worded items were only included during the DD activity day survey, as LAs were not present during clicker days. The parallel versions (A and B) of each item (Items 13 – 16) can be seen in Table 6.1 in Chapter 6.

Modifications based on results from the first-term general chemistry course

In addition to the item modifications noted above, new items were created to allow for the exploration of different factor structures in the data. These were created to address the two major concerns found when analyzing the data from the general chemistry course: 1) the limited number of personal effort items and 2) the possibility of a group-related factor. A total of 8 new items were created to bolster the personal effort category. For example, "I did not make much of an effort during today's class". Additionally, to address the possibility of a group-related factor, 4 new "other-focused" items were created that were similarly worded to the "self-focused" items from the original ASPECT. For example, the item, "*I* worked hard during today's class" led to the creation of the mASPECT-DD item, "*The students I worked with* worked hard during today's class". All new items were worded such that they matched the wording changes made to the original items based on the learning environment. These new item additions are presented in Table 6.2 in Chapter 6.

Item descriptive statistics for mASPECT-DD and mASPECT-C

Table C.7. Item descriptive statistics for mASPECT-DD and mASPECT-C.

Item	mASPECT-DD				mASPECT-C			
	Mean	Standard Deviation	Skew	Kurtosis	Mean	Standard Deviation	Skew	Kurtosis
1	4.73	0.90	-0.84	0.88	4.82	0.94	-0.70	0.11
2	4.56	0.97	-0.92	1.18	4.70	1.06	-0.69	0.01
3	4.95	0.73	-0.45	0.73	4.93	0.85	-0.51	0.04
4	4.79	0.87	-0.69	0.78	5.11	0.80	-0.37	-0.93
5	4.68	1.10	-1.05	0.91	4.93	1.07	-1.06	1.05
6	4.64	1.09	-0.82	0.45	5.04	0.95	-0.85	0.33
7	4.86	0.94	-1.11	2.51	4.91	0.95	-0.87	0.82
8	4.21	1.15	-0.71	0.50	4.49	0.97	-0.70	0.91
9	3.91	1.33	-0.39	-0.4	4.98	1.24	-1.40	1.63
10	4.74	0.81	-0.33	0.44	4.43	1.04	-0.45	0.14
11	4.30	1.11	-0.83	0.49	5.15	0.82	-1.18	3.42
12	4.22	1.19	-0.57	0.03	4.93	0.90	-0.65	0.24
13A	4.55	1.12	-1.08	1.39	5.19	0.82	-0.60	-0.63
13B	4.46	1.25	-0.90	0.58	---	---	---	---
14A	5.02	0.89	-1.01	2.09	5.36	0.82	-1.30	1.58
14B	4.98	1.06	-1.74	3.98	---	---	---	---
15A	5.28	0.75	-1.35	3.40	5.62	0.58	-1.23	0.49
15B	5.23	0.83	-1.64	4.76	---	---	---	---
16A	5.11	0.89	-1.42	3.19	5.18	0.84	-0.78	-0.10
16B	5.26	0.84	-1.86	5.76	---	---	---	---
17	4.88	1.11	-1.42	2.40	5.04	0.96	-1.11	1.64
18	5.00	0.93	-1.50	3.75	5.07	0.78	-0.31	-0.78
19	2.86	1.39	0.58	-0.51	2.40	1.19	0.99	0.65
20	4.95	1.09	-1.26	1.37	4.24	1.33	-0.67	-0.33
21	4.81	1.03	-0.95	0.52	2.17	1.11	1.07	1.11
22	2.26	1.18	1.10	0.85	2.57	1.38	0.69	-0.38
23	2.04	1.01	1.27	1.86	3.35	1.56	0.13	-1.15
24	4.38	1.22	-0.70	-0.14	2.99	1.33	0.35	-0.75
25	4.63	1.10	-0.70	-0.17	4.86	1.00	-0.74	0.25
26	4.64	1.12	-0.96	0.41	2.27	1.14	0.89	0.50
27	4.72	1.03	-1.01	1.10	4.79	0.95	-0.51	-0.17
28	4.69	0.95	-1.18	2.25	4.82	0.90	-0.74	0.71
29	4.66	0.91	-0.76	0.37	4.76	1.02	-1.08	1.82
30	4.14	1.04	-0.64	0.65	4.35	1.01	-0.57	0.89
31	4.45	1.33	-0.86	-0.04	4.56	1.13	-0.81	0.29

Iterative EFA process details

mASPECT-DD results

The 30 items remaining after qualitative analysis were analyzed using an iterative EFA process; an EFA was conducted, items that were not functioning well were removed, and then the remaining items were analyzed with a subsequent EFA. This process continued until distinct, well-functioning items and factors were discovered. Items were removed based on low factor loadings, cross-loading between multiple factors, and loading on a factor that contained only one or two items.

The initial EFA with all 30 items consisted of a possible 7-factor solution based on the Kaiser criterion and the scree plot. Four items (Items 3, 17, 18, and 20) had low loadings (<0.4) on all factors and two items (Items 13A and 13B) cross-loaded onto multiple factors. Additionally, two factors consisted of only one (Item 19) or two (Items 7 and 31) items. These nine items were removed and a subsequent analysis of the remaining 21 items was completed.

From these 21 items, a 4-factor and 3-factor solution were suggested depending on the presence of two 'group-related' items (Items 28 and 29). In the 4-factor solution, the factor containing Items 28 and 29 also contained three other items; Item 1, 2, and 14B. Items 1 and 14B had cross-loadings to other factors. This suggested that this fourth factor was not a group-related factor, as Items 1, 2, and 14B were not related to working in a group but instead to personal effort (Items 1 and 2) and instructor contribution (Item 14). Although Items 28 and 29 and Items 1 and 2 were focused on different aspects of the activity (i.e., the student themselves vs. the students they worked with), it is possible that the presence of this fourth factor was a result of the similar wording between these items instead of a group-related factor. Therefore, it was decided that Items 28 and 29 should be removed to allow for a 3-factor solution with well-defined factors.

A 3-factor EFA with the 19 remaining items resulted in Items 4 and 5 cross-loading between two factors. These items were related to explaining the material to or having the material explained by other students. During response process interviews, instead of focusing on explanations of the class material between students, which these items were meant to address, students would often mention how everyone in their group brought different opinions to the discussion. As such, these two items were removed.

mASPECT-C results

Iterative EFAs were completed with the 28 mASPECT-C items that remained after qualitative analysis in a similar fashion to the process completed for the mASPECT-DD items. The initial EFA consisted of a possible 7-factor solution, however, Items 9, 19, and 31 had low loadings (<0.4) and one factor only contained one item (Item 25). After these four items were removed, a possible 5-factor solution was found. This solution contained one factor with only two items (Items 22 and 23). These items were removed and a four-factor EFA with the remaining 22 items was completed. In this solution, Item 2 cross-loaded between two factors and was removed.

Final survey structure comparison (mASPECT vs. ASPECT)

Some similarities were noted between the mASPECT versions and the original ASPECT items and factor structure (Table C.8). The ASPECT and mASPECT-DD both contained three similarly named factors, however, the items that comprised these factors were different. New ‘personal effort’ items were included in mASPECT-DD that were not in the original ASPECT and some of the original ASPECT items were found to be irrelevant to the DD activity. In regards to the mASPECT-C, the three factors of ‘personal effort’, ‘value of environment’, and ‘classroom support’ were present, but the data collected required an additional fourth factor in order to best describe its structure. Although most of the possible group-related items loaded on the ‘value of group activity’ factor during development of the original ASPECT (Wiggins et al., 2017), the versions of these items on the mASPECT-C were found to contribute to a fourth factor, ‘social influence’. This supported the results from the general chemistry course, where the same items were found to share a similar source of variance beyond the original three factors. In addition to these items, two new items, Items 28 and 29, also contributed to ‘social influence’, as well as Item 17. Item 17 was an original ASPECT item; however, it did not contribute to any of the three factors in the original study (Wiggins et al., 2017).

Table C.8. Full comparison of the items and factor structures for the original ASPECT[a] and modified versions, mASPECT-DD and mASPECT-C.

mASPECT-DD			Original ASPECT Items[a]	mASPECT-C		
Items		Factors		Factors	Items	
Item 1	I was focused during today's class.	Personal Effort	I was focused during today's group activity.	Personal Effort	I was focused during today's class.	**Item 1**
Item 2	I worked hard during today's class.		I worked hard during today's group activity.		*I worked hard during today's class.*	*Item 2 - Removed*
Item 21	I did not make much of an effort during today's class.		*new item*		I did not make much of an effort during today's class.	**Item 21**
Item 24	I found it difficult to maintain my concentration during today's class.		*new item*		I found it difficult to maintain my concentration during today's class.	**Item 24**
Item 26	I was not very engaged in today's class.		*new item*		I was not very engaged in today's class.	**Item 26**
Item 27	I was fully engaged in today's class.		*new item*		I was fully engaged in today's class.	**Item 27**
Item 3 - Removed	*I made valuable contributions when working with other students during today's class.*		I made a valuable contribution during today's group activity.	Social Influence	I made valuable contributions when having discussions with other students during today's class.	**Item 3**
Item 4 - Removed	*Explaining the material to my group members improved my understanding of it.*	Value of Environment	Explaining the material to my group improved my understanding of it.		Explaining the material to other students improved my understanding of it.	**Item 4**
Item 5 - Removed	*Having the material explained to me by my group members improved my understanding of it.*		Having the material explained to me by my group members improved my understanding of the material.		Having the material explained to me by other students improved my understanding of it.	**Item 5**
Item 6	Working with other students during today's class contributed to my understanding of the material.		Group discussion during the activity contributed to my understanding of the course material.		Discussion with other students during today's class contributed to my understanding of the material.	**Item 6**
Item 7 - Removed	*The students I worked with made valuable contributions during today's class.*		Overall, the other members of my group made valuable contributions during the group activity.		The students I had discussions with made valuable contributions during today's class.	**Item 7**
Item 17 - Removed	*I felt comfortable working with other students during today's class.*		*I felt comfortable with my group.* **		I felt comfortable having discussions with other students during today's class.	**Item 17**
Item 28 - Removed	*The students I worked with were focused during today's class.*		*new item*		The students I had discussions with were focused during today's class.	**Item 28**
Item 29 - Removed	*The students I worked with worked hard during today's class.*		*new item*		The students I had discussions with worked hard during today's class.	**Item 29**

Table C.8 cont.

mASPECT-DD			Original ASPECT Items[a]	mASPECT-C		
Items		Factors		Factors	Items	
Item 8	I had fun during today's class.	Value of Environment (cont.)	I had fun during today's activity.	Value of Environment	I had fun during today's class.	**Item 8**
Item 9	I would prefer to take a class that included today's activity over one that does not include it.		I would prefer to take a class that includes this [topic] group activity over one that does not include this [topic] group activity.		*I would prefer to take a class that included today's clicker questions over one that does not include them.*	*Item 9 - Removed*
Item 10 - Removed	*I am confident in my understanding of the material presented during today's class.*		I am confident in my understanding of the material presented during today's group activity.		I am confident in my understanding of the material presented during today's class.	**Item 10**
Item 11	Today's class increased my understanding of the material.		The group activity increased my understanding of the course material.		Today's class increased my understanding of the material.	**Item 11**
Item 12	Today's class stimulated my interest in the course material.		The group activity stimulated my interest in the course material.		Today's class stimulated my interest in the course material.	**Item 12**
Item 13A - Removed	*The Professor/Teaching Assistant's enthusiasm made me more interested in today's class.*	Classroom Support (Instructor and LA)	The instructor's enthusiasm made me more interested in the group activity.	Classroom Support (Instructor only)	The Professor/Teaching Assistant's enthusiasm made me more interested in today's class.	**Item 13**
Item 13B - Removed	*The Learning Assistant's enthusiasm made me more interested in today's class.*					
Item 14A	The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.		The instructor put a good deal of effort into my learning for today's class.		The Professor/Teaching Assistant put a good deal of effort into my learning for today's class.	**Item 14**
Item 14B	The Learning Assistant put a good deal of effort into my learning for today's class.					
Item 15A	The Professor/Teaching Assistant seemed prepared for today's class.		The instructor seemed prepared to the group activity.		The Professor/Teaching Assistant seemed prepared for today's class.	**Item 15**
Item 15B	The Learning Assistant seemed prepared for today's class.					
Item 16A	The Professor/Teaching Assistant was available to answer questions during today's class.		The instructor and TAs were available to answer questions during the group activity.		The Professor/Teaching Assistant was available to answer questions during today's class.	**Item 16**
Item 16B	The Learning Assistant was available to answer questions during today's class.					

[a] Original ASPECT items and factor structure from Wiggins et al. (2017).

[**] Although this item was included in the original ASPECT survey development, it was not found to contribute to any of the factors in the original ASPECT.

Appendix D: Supporting Information for Chapter 7

Student engagement codes from the in-person environment

Tables D.1 – D.3 include the codes, descriptions, and example texts that comprised the codebook when coding the short-answer responses.

Table D.1. Behavioral engagement codes, descriptions, and example segments from short-answer responses. Key words and phrases are bolded.

Code	Description	Example from responses
Asked questions	asked questions in general can include: to TA, LA, or instructor must not include: to other group members	**Asking questions** when they don't understand.
Worked on worksheet	actively worked on worksheet did worksheet	**only doing…the worksheet**
Focused/paid attention	includes: not distracted	I would describe someone who was very engaged in the worksheet activities as **someone who was focused, present**
Tried to do worksheet	tried to do the worksheet put effort into doing the worksheet must not include: understanding, etc.	as a good student, who is **trying to do what is asked of the**
Completed worksheet		**completing the worksheet assignment.**
Was prepared		A student who is very engaged **comes to class prepared with the information by reading the appropriate chapters. They review the activity before coming to class**.
Worked with others	only includes: worked with others does not include: talked, discussed, etc.	Someone who…**works in groups**
Asked for group feedback	asked for feedback from group members asked questions to group members asked group members for help	The student…**seeks feedback from** the mentors and **peers**.
Participated		Someone who **actively participates**
Engaged with others		Someone who is active (asking questions, **engaging classmates**, and doing the presented work).
Talked to others (positive)	talked to group members but does not specify type of conversation only coded for responses to item related to a *very engaged* student	**talking with classmates**
Put in effort	put in effort but does not specify into what (e.g., doing the worksheet, understanding the material, etc.) also includes: tried, etc.	A very engaged student is one who stays completely focused on the task at hand and aims to understand the topics of the activity to the best of their ability **while giving full effort**..
Was on a non-class related device	was on a teach device (e.g., phone, laptop, etc.) not related to activity	**sits on phone the whole time**.
Didn't work on worksheet	didn't actively work on worksheet didn't do worksheet	**Not working on the worksheet**

Table D.1 cont.

Code	Description	Example from responses
Worked on other things	worked on a different assignment or class	**doing other activities**.
Distracted	includes: not focused, not paying attention	**distracted**
Didn't try to do worksheet	didn't put effort into doing worksheet didn't try to do worksheet must not include: not understanding, etc.	The student **does not attempt the worksheet**
Participated in off-topic conversations		They are…**chatting with their friends without doing the worksheet.**
Didn't work with others	didn't work with others didn't work in a group does not include: didn't talk, didn't discuss, etc.	A student who **does not…work on it with other students**.
Didn't ask questions	didn't ask questions in general can include: to TA, LA, or instructor must not include: to other group members	**do not ask questions**
Left class early		**Leaving class early**
Wasn't prepared	Includes: not bringing the worksheet to class	**not bringing a copy of the worksheet**
Talked to others (negative)	talked to group members but does not specify type of conversation only coded for responses to item related to a *not engaged* student	Not doing the worksheet and just doing other things or **talking**
Didn't put in effort	didn't put in effort but does not specify into what (e.g., doing the activity, understanding the material, etc.) also includes: didn't try, etc.	Someone who **doesn't give any effort to the activity**.
Didn't participate		Somebody who **doesn't…otherwise participate in the activity**.
Didn't complete worksheet		A student who **does not complete the worksheet**.
Copied answers from others	copied answers from group members	A student who is not engaged will **passively write answers their group members come up with**.

Table D.2. Cognitive engagement codes, descriptions, and example segments from short-answer responses. Key words and phrases are bolded.

Code	Description	Example from responses
Helped others	helped group members taught group members provided feedback to group members	A person that is **helping everyone around them with the worksheet**
Tried to understand	tried to understand put effort into understanding	**trying to understand the material**.
Discussed with others	discussion with group members communicated with group members includes: collaboration or talking with group members about the worksheet, sharing ideas with group members, etc.	A student who does all the exercises and **discusses them with the students around them.**
Interacted with worksheet	e.g., read through worksheet, wrote down notes, studied worksheet, etc.	Someone who…**took specific notes to later put on their sheet for the final**
Put effort into learning	put effort into learning the material tried to learn the material	Someone who…is **trying to learn the content**
Learnt from and/or corrected mistakes		Somebody who…**strives to correct any mistakes made on the paper**.
Tried to solve problems a different way		**trying to solve the problem a different way.**
Used resources	used resources to help with worksheet (can include: previous notes, internet, book, worksheet models, etc.)	**using phone to find information that they may need to complete the activity**
Connected or applied material	…to previous or future course material, to other classes, to real-life, etc.	they would rather take the time to complete the worksheet **to learn and understand how it can be applied to real-life scenarios or problems seen on the test** instead of completing the worksheet to get it done.
Tried their best/didn't give up	tried their best when the worksheet was difficult didn't give up when it was difficult	Also a student who seeks for help when stuck **instead of giving up**.
Did more than the minimum	i.e., went beyond simply doing the activity	**Going out of their way for activity**
Made sure everyone understood	made sure everyone had the answer and/or understood the material	who solved the worksheet with all members and **make sure everyone understand**
Didn't try to understand	didn't try to understand didn't put effort into understanding	**not trying to understand concepts and equations**
Just wrote down answers	i.e., filled out or did worksheet without trying to understand includes: just looked up the answers	**only writing down the answers**.
Didn't discuss with others	didn't discuss with group members includes: didn't collaborate, didn't contribute, didn't talk, etc. with group members	**The student doesn't** put a lot of efforts in doing or **collaborating with others** to complete the worksheets.
Only did the minimum required	only did the minimum needed for the class i.e., only did clicker questions	Someone who…**is just there to get clicker participation credit.**

Table D.2 cont.

Code	Description	Example from responses
Didn't try their best/gave up	didn't try their best when the worksheet was difficult gave up when it was difficult	**giving up on the problem just to move on to the next one**
Didn't put effort into learning	didn't put effort into learning the material didn't try to learn the material	Someone that **does not…put any effort in learning the materials**.

Table D.3. Emotional engagement codes, descriptions, and example segments from short-answer responses. Key words and phrases are bolded.

Code	**Description**	**Example from responses**
Felt activity was beneficial	felt the activity was beneficial or useful for learning (i.e., valuable, etc.)	A person who **thinks the worksheets are important to understanding the concept**
Positive feelings	positive feelings in general (e.g., good)	Did all of the problems with a **positive attitude throughout the process**.
Liked/enjoyed the activity	liked the activity enjoyed the activity (e.g., had fun) must be specific to the worksheet or activity (not chemistry or science in general)	A student **who really likes the worksheets** would be engaged.
Interested in content	interested in specific content or topics covered on worksheet	someone **who finds the topic very interesting**
Looked forward to activity	looked forward to the activity excited about the activity	**those who actually look forward to activity days.**
Didn't feel frustrated	includes: not frustrated, not overwhelmed	**not feel overwhelmed or frustrated** by the assignment
Liked working with others	Includes: interested in working with others	**They are interested in teamwork.**
Felt activity wasn't beneficial	felt the activity was not beneficial or useful for learning (i.e., not valuable, etc.)	Someone who **views the worksheets as a waste of time or not worth even attempting to complete.**
Felt confused or discouraged	includes: had a hard time, struggling	Someone is not grasping the material fully **and is too confused to know where to start**
Negative feelings	negative feelings in general (e.g., bad)	**Not having good feelings towards worksheet**
Didn't like/enjoy activity	didn't like the activity didn't enjoy the activity must be specific to the worksheet or activity (not chemistry or science in general)	A student **who really dislikes the worksheets** would not be engaged.
Didn't care about activity	didn't care about doing the worksheet or activity	A student **who doesn't care about the worksheets**
Not interested in content	not interested in specific content or topics covered on worksheet	someone **who is not interested in the topic**
Felt frustrated	includes: frustrated, overwhelmed	**being frustrated** and not doing anything about it.
Didn't like chemistry/science	didn't like chemistry or science can also include: not excited, not interested, etc. must be directed toward the subject/field in general	**they just don't like chemistry as much as others.**
Didn't look forward to activity	also includes: not excited to do activity	**someone who doesn't…look forward to doing them in class**.
Didn't want to do activity		Someone who is **not wanting to work on them**
Didn't like working with others		Maybe **they don't like teamwork**

Student engagement codes from the remote environment

Tables D.4 – D.6 include the codes, descriptions, and example segments that comprised the codebook when coding the interview transcripts.

Table D.4. Behavioral engagement codes, descriptions, and example segments from interview transcripts. Key words and phrases are bolded.

Code	Description	Example from transcript
Wrote things down	includes: on paper or on computer	I feel like everyone that is engaging in it **writes down**…so I'll just like, if it's like a math one then I'll **write out the equation**…
Talked to/worked with others	must include: some kind of interaction with group members (e.g., talked to, used chat function, worked with, etc.) does not include: discussion or collaboration, etc. to solve problems on worksheet	...you're **talking and working on it with other people**.
Read question to self	read question out loud to self	I would say an engaged student would probably...they would just **read it out loud**.
Focused/paid attention	includes: not distracted	...you're **paying attention,** there's **no other distractions.**
Worked on worksheet	actively worked on worksheet did worksheet	One being that we're active, we're **doing the worksheet**...
Tried to do worksheet	tried to do the worksheet put effort into doing the worksheet must not include: understanding, etc.	...I just **try to work through** what I can...
Asked for group feedback	asked for feedback from group members asked questions to group members asked group members for help	I'll voice out my reasons, say, "oh **I need help**. You know, can someone help me." And that's just how I've been doing it. You know, I just **try to get feedback from other people**.
Led the group	took actions related to keeping the group on task, directing the group, etc.	...a lot of times my priority...is to **make sure that the group is on the same page**...I have tried to **ensure that we're all on, at least on the same point**, even if we're not all talking...
Was prepared		...what it really comes down to is...you got to **be prepared**.
Participated		...someone who is actively **participating**...
Asked questions	asked questions in general can include: to TA, LA, or instructor must not include: to other group members	…then you **ask your question**...
Shared screen	shared their screen to group	I would say a good example of that is probably someone or two...but usually one person **sharing their screen**.
Listened to others		...there was one time where I just got stuck and I think I just kinda stopped talking and **just listen**...

Table D.4 cont.

Code	Description	Example from transcript
Just "there"	e.g., just listened, not writing things down, not thinking about things, camera/mic off, etc.	They're **just there**. ...if I was not engaging physically or physically participating, I would probably **never write things down**...I would probably **never like even flip a page in my notebook**... ...I'm **just going to sit here and be here**...
Distracted	i.e., distracted actions (on phone, etc.) not including doing other work includes: not focused, not paying attention	I would be **doodling** on the sheet **daydreaming**. ...I'll be **on my phone**...
Worked on other things	worked on a different assignment or class	...you're actually **working on a different assignment** from a different class.
Didn’t try to do worksheet	must not include: not understanding, etc.	...I **don't try to do the activity**...
Didn't work on worksheet	didn't actively work on worksheet didn't do worksheet	...just **not doing it**.
Didn't ask questions	didn't ask questions in general	I would probably **never** even...**ask a question**, I guess, **to anybody**.
Wasn’t prepared		Probably **not bother reading any of the textbooks**...
Didn't talk to/work with others	didn't talk to group members (e.g., quiet) didn't work with group members didn't interact with group members (e.g., didn't use chat function, etc.)	Someone that's **quiet**, **not looking to talk to other people**.
Didn't lead the group	didn't take actions related to keeping the group on task, directing the group, etc.	...I might **not be exactly the one that's leading**...

Table D.5. Cognitive engagement codes, descriptions, and example segments from interview transcripts. Key words and phrases are bolded.

Code	**Description**	**Example from transcript**
Tried to understand	tried to understand put effort into understanding	...**making sure that I understand** the content fully.
Checked work/answers	checked answers (okay if checked with other group members) tried to find mistakes tried to understand mistakes	So that would just include specifically...going over what you've already done on it, **trying to find an error**. We'll **go over answers**.
Thought about how to solve problems		I suppose usually what I do is I try to **think about** kind of very briefly, like, what am I, **what's the general story** of the question, but very quickly I go to, what are the numbers that, what are they, which answer do they want? Uh, or I don't even want to put it like that. **What answer is being sought and what are the initial pieces of information that are actually pertinent to that?**
Discussed with others	includes: discussion, collaboration, sharing ideas, bouncing ideas, etc. with group members	I just try to **bounce back ideas back and forth** just to get a common understanding of what's going on. ...just **discussion, discussing with other people** about the question.
Helped others	helped group members understand provided feedback to group members answered group members questions	I **assist others** if they have trouble.
Interacted with worksheet	e.g., took down extra notes, wrote down extra details, etc.	I would just be **really detailed in my notation** to prove that I'm really **interacting with the material**.
Used resources	used resources when working on worksheets (can include: previous notes, internet, book, worksheet models, etc.)	So I usually **refer back to the notes** that we took on the lecture day...occasionally I'll **open my book**... I **look stuff up on the internet on the activities** and stuff…
Went through problems step-by-step		I like to like read the question out loud and then kind of **go step by step through it.**
Connected or applied material	…to previous or future course material, to other classes, to real-life, etc.	...I'm trying to **connect it to past topics** that we went over.
Just wrote down answers	i.e., filled out or did worksheet without trying to understand.	I could **write down other people's answers** if I wasn't engaged...like just **going through the worksheet not understanding.**
Gave up		…they might try to do one problem and then they **give up**…
Didn't discuss with others	didn't discuss, collaborate, share ideas, bounce ideas, etc. with group members	We **wouldn't be building a conversation** at all.
Didn't help others	didn't help group members didn't provide feedback to group members	...**not**...**trying to answer them [other's questions]**.

Table D.6. Emotional engagement codes, descriptions, and example segments from interview transcripts. Key words and phrases are bolded.

Code	Description	Example from transcript
Felt confident	confidence empowered emotions related to getting answers correct	...it feels good to...get things right, you know, and like positive feedback, if you get one thing right, you're **more confident** that you can get the next thing right.
Positive feelings	positive feelings in general (e.g., good)	...almost just like **a good feeling**, just like you're doing **something good and positive** and, you know, adding to yourself or what you're doing for that day.
Felt activity was beneficial	felt the activity was beneficial or useful for learning (i.e., valuable)	I do like when **it helps me understand the concept more**.
Wanted to learn		I **want to learn** and be able to absorb this stuff.
Wanted to/liked working with others	a desire to work with others or an enjoyment of working with others includes: liked working with others, wanted to work with others, interested in working with others, etc.	...I just **want to**, you know, **talk to everyone**...
Wanted to help others	wanted to help group members wanted to make sure all group members understood the material	I think an engaged person is someone that's **really eager to help others.**
Liked/enjoyed the activity	liked the activity enjoyed the activity (e.g., had fun) must be specific to the worksheet or activity (not chemistry or science in general)	...if someone's like, yeah, **this is fun. I like it**.
Liked chemistry/science	liked chemistry or science can also include: enjoyed, interested, etc. must be directed toward the subject/field in general	I actually **really enjoy learning about chemistry**.
Didn't feel frustrated	includes: not frustrated, not stressed	**Not** so much **stressed out**…
Excited about activity	excited/enthusiastic about activity can also include: being excited about learning through the activity	I think engagement is...kind of like **enthusiasm**, really.
Interested in content	interested in specific content or topics covered on worksheet	And **I think it's [the content] interesting** too...
Felt self-doubt	doubt not empowered emotions related to getting answers incorrect	I have like a feedback loop that happens, where your…thought process starts getting really negative and **self-doubting** and deprecating.
Negative feelings	negative feelings in general (e.g., bad)	...if you're feeling **negative**, **pessimistic**.
Felt frustrated	includes: frustrated, stressed	...I was extremely **frustrated**.

Table D.6 cont.

Code	Description	Example from transcript
Felt disconnected	felt disconnected in relation to the activity or material/content must not be related to working with others	…where I feel like I'm **untethered** and so I **don't know where to put that information** and I **don't know where that fits in** with the rest of it.
Didn't want to learn		...you **don't want to actually learn** how to do it, how to get the answer, how to get the right answer.
Didn't want to/like working with others	didn't want to work with others didn't like working with others can include: wanted to/liked to work on their own	They **don't want to work with others**.
Felt activity wasn't beneficial	felt the activity wasn't beneficial or useful for learning (i.e., not valuable)	...personally I feel, if **I feel like the learning is already**, I would say **sufficient**...I would probably...be more prone to not engage...
Felt left behind/rushed		...you definitely **feel a little bit rushed.**
Didn't like/enjoy the activity	didn't like the activity didn't enjoy the activity must be specific to the worksheet or activity (not chemistry or science in general)	…you **don't like** what you're doing [the activity]...
Didn't like chemistry/science	didn't like chemistry or science can also include: not excited, not interested, etc. must be directed toward the subject/field in general	So like with **chemistry**, I'm like, **it's not what I'm excited to learn** but I'm here.
Felt bored		Feeling **bored**, like a strong feeling of **boredom** probably.
Not interested in content	not interested in specific content or topics covered on worksheet	...that would probably just, that would include...**not being as interested** in why a particular answer is incorrect.

Coding results from the in-person environment

Tables D.7 – D.9 include the number of students whose responses aligned to each code when students were asked to describe VERY engaged and NOT engaged students in the context of the worksheet activities through short-answer responses.

Table D.7. Number of students that mentioned each behavioral engagement code when asked to describe students who were VERY engaged and NOT engaged in the worksheet activities relative to a specific definition of engagement (i.e., behavioral, cognitive, emotional).

Behavioral Code	Number of Students (%)		
	Engagement definitions given		
	Behavioral, n = 55	Cognitive, n = 57	Emotional, n = 58
Engagement			
Asked questions	21 (38.2)	21 (36.8)	11 (19.0)
Worked on worksheet	13 (23.6)	16 (28.1)	5 (8.6)
Focused/paid attention	8 (14.5)	9 (15.8)	10 (17.2)
Tried to do worksheet	8 (14.5)	8 (14.0)	10 (17.2)
Completed worksheet	8 (14.5)	8 (14.0)	6 (10.3)
Was prepared	8 (14.5)	2 (3.5)	0 (0.0)
Worked with others	6 (10.9)	2 (3.5)	5 (8.6)
Asked for group feedback	5 (9.1)	4 (7.0)	2 (3.4)
Participated	5 (9.1)	2 (3.5)	3 (5.2)
Engaged with others	3 (5.5)	1 (1.8)	2 (3.4)
Talked to others (positive)	2 (3.6)	3 (5.3)	2 (3.4)
Put in general effort	2 (3.6)	0 (0.0)	3 (5.2)
Disengagement			
Was on a non-class related device	19 (34.5)	15 (26.3)	8 (13.8)
Didn't work on worksheet	18 (32.7)	17 (29.8)	12 (20.7)
Worked on other things	8 (14.5)	11 (19.3)	4 (6.9)
Distracted	6 (10.9)	12 (21.1)	5 (8.6)
Didn't try to do worksheet	6 (10.9)	8 (14.0)	5 (8.6)
Participated in off-topic conversations	6 (10.9)	5 (8.8)	5 (8.6)
Didn't work with others	6 (10.9)	2 (3.5)	4 (6.9)
Didn't ask questions	4 (7.3)	3 (5.3)	0 (0.0)
Left class early	3 (5.5)	3 (5.3)	4 (6.9)
Wasn't prepared	3 (5.5)	3 (5.3)	0 (0.0)
Talked to others (negative)	3 (5.5)	1 (1.8)	0 (0.0)
Didn't put in general effort	2 (3.6)	6 (10.5)	8 (13.8)
Didn't participate	2 (3.6)	2 (3.5)	1 (1.7)
Didn't complete worksheet	2 (3.6)	2 (3.5)	1 (1.7)
Copied answers from others	2 (3.6)	2 (3.5)	1 (1.7)

Table D.8. Number of students that mentioned each cognitive engagement code when asked to describe students who were VERY engaged and NOT engaged in the worksheet activities relative to a specific definition of engagement (i.e., behavioral, cognitive, emotional).

Cognitive Code	**Number of Students (%)**		
	Engagement definitions given		
	Behavioral, n = 55	Cognitive, n = 57	Emotional, n = 58
Engagement			
Helped others	6 (10.9)	16 (28.1)	10 (17.2)
Tried to understand	3 (5.5)	15 (26.3)	7 (12.1)
Discussed with others	14 (25.5)	9 (15.8)	6 (10.3)
Interacted with worksheet	1 (1.8)	4 (7.0)	2 (3.4)
Put effort into learning	2 (3.6)	2 (3.5)	4 (6.9)
Learnt from and/or corrected mistakes	3 (5.5)	1 (1.8)	0 (0.0)
Tried to solve problems a different way	0 (0.0)	2 (3.5)	0 (0.0)
Used resources	1 (1.8)	1 (1.8)	0 (0.0)
Connected or applied material	0 (0.0)	1 (1.8)	2 (3.4)
Tried their best/didn't give up	0 (0.0)	1 (1.8)	2 (3.4)
Did more than the minimum	1 (1.8)	0 (0.0)	2 (3.4)
Made sure everyone understood	4 (7.3)	0 (0.0)	0 (0.0)
Disengagement			
Didn't try to understand	3 (5.5)	7 (12.3)	2 (3.4)
Just wrote down answers	2 (3.6)	5 (8.8)	4 (6.9)
Didn't discuss with others	2 (3.6)	4 (7.0)	2 (3.4)
Only did the minimum required	0 (0.0)	4 (7.0)	0 (0.0)
Didn't try their best/gave up	0 (0.0)	3 (5.3)	1 (1.7)
Didn't put effort into learning	2 (3.6)	0 (0.0)	2 (3.4)

Table D.9. Number of students that mentioned each emotional engagement code when asked to describe students who were VERY engaged and NOT engaged in the worksheet activities relative to a specific definition of engagement (i.e., behavioral, cognitive, emotional).

Emotional Code	Number of Students (%)		
	Engagement definitions given		
	Behavioral, n = 55	Cognitive, n = 57	Emotional, n = 58
Engagement			
Felt activity was beneficial	1 (1.8)	0 (0.0)	8 (13.8)
Positive feelings	0 (0.0)	0 (0.0)	6 (10.3)
Liked/enjoyed the activity	0 (0.0)	0 (0.0)	4 (6.9)
Interested in content	0 (0.0)	0 (0.0)	2 (3.4)
Looked forward to activity	0 (0.0)	0 (0.0)	2 (3.4)
Didn't feel frustrated	0 (0.0)	0 (0.0)	1 (1.7)
Liked working with others	1 (1.8)	0 (0.0)	0 (0.0)
Disengagement			
Felt activity wasn't beneficial	0 (0.0)	1 (1.8)	8 (13.8)
Felt confused or discouraged	1 (1.8)	2 (3.5)	7 (12.1)
Negative feelings	0 (0.0)	0 (0.0)	6 (10.3)
Didn't like/enjoy activity	0 (0.0)	0 (0.0)	5 (8.6)
Didn't care about activity	3 (5.5)	0 (0.0)	3 (5.2)
Not interested in content	1 (1.8)	0 (0.0)	3 (5.2)
Felt frustrated	0 (0.0)	0 (0.0)	3 (5.2)
Didn't like chemistry/science	0 (0.0)	0 (0.0)	1 (1.7)
Didn't look forward to activity	0 (0.0)	0 (0.0)	1 (1.7)
Didn't want to do activity	0 (0.0)	0 (0.0)	1 (1.7)
Didn't like working with others	1 (1.8)	0 (0.0)	0 (0.0)

Coding results from the remote environment

Tables D.10 – D.12 include the number of students that mentioned ideas related to each code when students were asked to describe engaged and not engaged students in the context of the worksheet activities during interviews.

Table D.10. Number of students that mentioned each behavioral engagement code when provided the specific definitions of engagement (i.e., behavioral, cognitive, emotional).

Behavioral Code	**Number of students (%), n = 14**			
		Engagement definitions given		
	Overall[a]	Behavioral	Cognitive	Emotional
Engagement				
Wrote things down	11 (78.6)	11 (78.6)	3 (21.4)	0 (0.0)
Talked to/worked with others	8 (57.1)	4 (28.6)	6 (42.9)	0 (0.0)
Read question to self	8 (57.1)	3 (21.4)	6 (42.9)	0 (0.0)
Focused/paid attention	6 (42.9)	5 (35.7)	2 (14.3)	0 (0.0)
Worked on worksheet	8 (57.1)	7 (50.0)	3 (21.4)	2 (14.3)
Tried to do worksheet	4 (28.6)	2 (14.3)	2 (14.3)	2 (14.3)
Asked for group feedback	5 (35.7)	2 (14.3)	4 (28.6)	0 (0.0)
Led the group	5 (35.7)	0 (0.0)	5 (35.7)	1 (7.1)
Was prepared	5 (35.7)	4 (28.6)	1 (7.1)	0 (0.0)
Participated	4 (28.6)	3 (21.4)	1 (7.1)	0 (0.0)
Asked questions	3 (21.4)	2 (14.3)	1 (7.1)	0 (0.0)
Shared screen	3 (21.4)	3 (21.4)	0 (0.0)	0 (0.0)
Listened to others	1 (7.1)	1 (7.1)	0 (0.0)	0 (0.0)
Disengagement				
Just “there”	12 (85.7)	10 (71.4)	6 (42.9)	0 (0.0)
Distracted	9 (64.3)	9 (64.3)	4 (28.6)	0 (0.0)
Worked on other things	10 (71.4)	6 (42.9)	4 (28.6)	3 (21.4)
Didn’t talk to/work with others	8 (57.1)	4 (28.6)	3 (21.4)	1 (7.1)
Didn’t try to do worksheet	7 (50.0)	5 (35.7)	1 (7.1)	1 (7.1)
Didn’t work on worksheet	6 (42.9)	3 (21.4)	2 (14.3)	2 (14.3)
Didn’t ask questions	3 (21.4)	1 (7.1)	2 (14.3)	0 (0.0)
Wasn’t prepared	2 (14.3)	1 (7.1)	1 (7.1)	0 (0.0)
Didn’t lead the group	1 (7.1)	0 (0.0)	1 (7.1)	0 (0.0)

[a]Number of students who mentioned code at least once during the three definitions.

Table D.11. Number of students that mentioned each cognitive engagement code when provided the specific definitions of engagement (i.e., behavioral, cognitive, emotional).

Cognitive Code	**Number of students (%), n = 14**			
		Engagement definition sections		
	Overall[a]	Behavioral	Cognitive	Emotional
Engagement				
Tried to understand	13 (92.9)	0 (0.0)	12 (85.7)	3 (21.4)
Checked work/answers	7 (50.0)	1 (7.1)	5 (35.7)	1 (7.1)
Thought about how to solve problems	7 (50.0)	3 (21.4)	4 (28.6)	0 (0.0)
Discussed with others	6 (42.9)	3 (21.4)	4 (28.6)	2 (14.3)
Helped others	5 (35.7)	2 (14.3)	5 (35.7)	0 (0.0)
Interacted with worksheet	5 (35.7)	3 (21.4)	2 (14.3)	0 (0.0)
Used resources	6 (42.9)	4 (28.6)	4 (28.6)	0 (0.0)
Went through problems step-by-step	4 (28.6)	1 (7.1)	2 (14.3)	1 (7.1)
Connected or applied material	4 (28.6)	1 (7.1)	3 (21.4)	1 (7.1)
Disengagement				
Just wrote down answers	6 (42.9)	1 (7.1)	6 (42.9)	1 (7.1)
Didn't discuss with others	3 (21.4)	2 (14.3)	1 (7.1)	0 (0.0)
Gave up	3 (21.4)	1 (7.1)	0 (0.0)	2 (14.3)
Didn't help others	1 (7.1)	0 (0.0)	1 (7.1)	0 (0.0)

[a] Number of students who mentioned code at least once during the three definitions.

Table D.12. Number of students that mentioned each emotional engagement code when provided the specific definitions of engagement (i.e., behavioral, cognitive, emotional).

Emotional Code	**Number of students (%), n = 14**			
		Engagement definitions given		
	Overall[a]	Behavioral	Cognitive	Emotional
Engagement				
Felt confident	10 (71.4)	0 (0.0)	0 (0.0)	10 (71.4)
Positive feelings	9 (64.3)	0 (0.0)	0 (0.0)	9 (64.3)
Felt activity was beneficial	8 (57.1)	1 (7.1)	3 (21.4)	7 (50.0)
Wanted to learn	4 (28.6)	1 (7.1)	1 (7.1)	3 (21.4)
Wanted to/liked working with others	5 (35.7)	1 (7.1)	2 (14.3)	3 (21.4)
Wanted to help others	4 (28.6)	1 (7.1)	4 (28.6)	0 (0.0)
Liked/enjoyed the activity	6 (42.9)	0 (0.0)	0 (0.0)	6 (42.9)
Liked chemistry/science	3 (21.4)	0 (0.0)	1 (7.1)	3 (21.4)
Didn’t feel frustrated	2 (14.3)	0 (0.0)	0 (0.0)	2 (14.3)
Excited about activity	4 (28.6)	0 (0.0)	1 (7.1)	4 (28.6)
Interested in content	2 (14.3)	0 (0.0)	2 (14.3)	0 (0.0)
Disengagement				
Felt self-doubt	6 (42.9)	0 (0.0)	0 (0.0)	6 (42.9)
Negative feelings	7 (50.0)	0 (0.0)	0 (0.0)	7 (50.0)
Felt frustrated	5 (35.7)	1 (7.1)	0 (0.0)	5 (35.7)
Felt disconnected	6 (42.9)	2 (14.3)	0 (0.0)	4 (28.6)
Didn’t want to learn	5 (35.7)	0 (0.0)	2 (14.3)	3 (21.4)
Didn’t want to/like working with others	5 (35.7)	1 (7.1)	3 (21.4)	1 (7.1)
Felt activity wasn’t beneficial	4 (28.6)	0 (0.0)	2 (14.3)	3 (21.4)
Felt left behind/rushed	2 (14.3)	1 (7.1)	0 (0.0)	1 (7.1)
Didn’t like/enjoy the activity	2 (14.3)	0 (0.0)	0 (0.0)	2 (14.3)
Didn’t like chemistry/science	2 (14.3)	0 (0.0)	0 (0.0)	2 (14.3)
Felt bored	1 (7.1)	0 (0.0)	0 (0.0)	1 (7.1)
Not interested in content	1 (7.1)	0 (0.0)	1 (7.1)	0 (0.0)

[a]Number of students who mentioned code at least once during the three definitions.

Appendix E: Supporting Information for Chapter 8

Demographics

Table E.1. Self-reported demographics data (compiled from Fall 2020 surveys) of students who participated in this study.

	Percentage of Students (n = 444)
Gender	
Female	53
Male	46
Other	1
Race/ethnicity[a]	
Non-URM	67
URM	33
Age bracket (years)	
18-22 (traditional)	67
23+ (nontraditional)	33
Major	
Biology	31
Chemistry	1
Biochemistry	7
Other Science[b]	41
Non-Science[c]	15
None	6
University status	
Postbaccalaureate	9
Undergraduate	90
Rather not say	1
Transfer status	
Transfer from 2-year college	39

[a] Non-URM consisted of individuals who identified either as non-Latino/a White or as Asian.
[b] E.g., physics, geology, etc.
[c] E.g., business, accounting, etc.

Preliminary Survey Item Modifications

Analysis of the preliminary survey items was completed in the general chemistry in-person lecture classes during the Fall and Winter terms of the 2019 – 2020 academic year. Preliminary items were adapted to be as close as possible to the items from the Wang et al. (2016) engagement survey and were minimally re-worded to focus on specific activities and be in past-tense. Throughout the two terms, quantitative survey responses were collected (n = 547 responses over both terms), along with response process data in the format of short-answer written responses (n = 27 – 44 responses per item) and interviews (n = 12). Items were modified, removed, or added based on the results from both quantitative and qualitative data (except for item S3, where technical difficulties resulted in no survey response data being collected). The preliminary and pilot items are included in Table E.2. Details about why items were modified, removed, or added to the engagement scales related to qualitative results are provided below. Any behavioral, cognitive, or emotional engagement items that were removed also displayed low loadings (λ <0.4) when survey data were analyzed with single-factor CFAs of the respective engagement dimensions.

Behavioral Engagement Items

Modified Items: Multiple items were slightly modified to clarify actions related to *doing* and *working* on the worksheet. Additionally, item B7 was modified as students were confused with to what "other things" was referring.

Removed Items: Three items, B4, B5, and B8, were removed due to being irrelevant to the class environment (i.e., not every class had assigned pre-work) and student population. When asked about B5, students would mention that they prefer to keep their home and school life separate. For example, one student wrote, *"I will help people who didn't understand and ask but I'm not going to chat about it with friends."* When asked about item B8, students would mention that they couldn't think of any reason to give up on the worksheet based on the prevalence of people (peers, LAs, etc.) who could help if they didn't understand something. For example, one student wrote, "*I asked questions to others in my group or to one of the people walking around helping in order for me to understand fully.*"

New Items: Two new items (B9 and B10) were created related to students' working on and trying to answer the worksheet problems. These items were created through discussion between the two authors and were focused on the behavioral definition of engagement by Fredricks et al. (2004).

Cognitive Engagement Items

Modified Items: Item C1 was reworded to focus on understanding the material instead of simply getting the answers right, since answers were provided to students throughout the activity and on a final posted answer key. Additionally, students were found to be responding to the item based on whether or not they reviewed their work instead of focusing on whether they tried to understand the material. For example, one student stated that, "*I didn't really look over it again but I felt like I understood the material so that's why I put 'somewhat'.*" Item C3 was clarified to specify connection back to previous concepts they had learned instead of vague "things". When asked about item C5, student responses were focused more on the structure of the activities, as most mentioned that they liked both working through the problems and getting the answers. Since answers were provided to students throughout the activity through check-ins, they were still expected to do the work. However, students had the ability to simply write the answers on their worksheet instead of doing and understanding the work, so item C5 was reworded to reflect this. Item C6 was slightly reworded to focus on how hard students thought when they got to challenging problems. As the worksheets were a combination of key questions, exercises, and problems, it was expected that students would not necessarily have to think hard on the initial key questions. Student responses reflected this variation in problem difficulty with one student saying, "*During today's activity, there were times where I was thinking hard, but there were times where I wasn't really thinking too hard because the answer was obvious.*"

Removed Items: Item C2 was removed as students interpreted the item differently than intended, with one student writing, "*I tried to think of other examples of limiting reactants in other scenarios, such as fruit in smoothies, plywood/nails while building a house, etc.*" These responses may have been due to the design of the worksheets, as they were meant to guide students through the process and methods needed to solve the problems instead of being open-ended. Item C7 was removed as students did not find it applicable to the structure of the worksheets, stating, "*I thought most of the problems were the same level of difficulty and you couldn't really skip one because they were all tied together*."

Emotional Engagement Items

Modified Items: Items E2 and E8 were slightly modified. "New things" was removed from item E2 since the material included in the activities was not necessarily new to the students. Item E8 was slightly reworded to focus on the activity itself instead of the content that the students felt they needed to learn to pass the class. Item E5 was modified to center the feeling of frustration around the activity instead of other sources.

Removed Items: When asked about item E3, students mentioned *needing* to understand the content because it was going to be on the test, whether or not the content was presented in lecture-format or through the activity. Thus, E3 was removed since it was not directly related to the activity itself. Item E9 was removed because multiple students expressed that they didn't really know what the item meant and that they associated "feeling down" with "being depressed". For example, one student said that, "*Down is something that I'm a little iffy on the wording of, just it makes it feel like this activity is depressing me…and like, kind of, but it just made me bored."* Item E10 was removed since it was found to be irrelevant to students who felt like 1) it wasn't new material, and 2) there was no reason to be nervous to do the worksheet activities. When asked a follow up question about if they were ever worried going into the activities, one student responded that, "*Not really, no. I read prior to it as well so I felt pretty okay about the ideas in general*."

Social Engagement Items

Removed Items: As students generally worked with whoever was seated nearby in the in-person environment, the item S3 was removed as it was not relevant to the classroom environment. As one student stated, "*I just work with who's around me*."

Table E.2. Preliminary items (slightly re-worded from Wang et al. (2016)) and pilot items based on preliminary CFAs and interview results.

Preliminary Items	Pilot Items
Behavioral Items	
B1. I stayed focused during today's activity.	I stayed focused during today's activity.
B2. I put effort into learning during today's activity.	I put effort into doing today's activity.
B3. I kept trying even if something was hard during today's activity.	I kept working on today's activity even if something was hard.
B4. I completed all the required pre-work for today's activity.	
B5. I plan to share with others what I learned during today's activity.	
B6. I didn't actively participate in today's activity.	I didn't actively participate in today's activity.
B7. I did other things when I was supposed to be working on today's activity.	I did things I was not supposed to be doing during today's activity.
B8. If I didn't understand something during today's activity, I gave up right away.	
B9. ---	I didn't do much work on today's activity.
B10. ---	I attempted to answer most of the items on today's activity.
Cognitive Items	
C1. I went through my work during today's activity and made sure it was right.	I made sure I understood my work on today's activity.
C2. I thought about different ways to solve problems during today's activity.	
C3. I tried to connect what I was learning during today's activity to things I have learned before.	I tried to connect what I was learning during today's activity to concepts I have learned before.
C4. I tried to understand my mistakes when I got something wrong during today's activity.	I tried to understand my mistakes when I got something wrong during today's activity.
C5. I would have preferred to have been given the answers rather than doing the work during today's activity.	I wrote down the answers to today's activity without trying to understand them.
C6. I didn't think very hard when working on today's activity.	I didn't think very hard when I came across a challenging problem on today's activity.
C7. I preferred to focus on the easier problems during today's activity.	
C8. I did just enough to get by during today's activity.	I did just enough to get by during today's activity.
Emotional Items	
E1. I looked forward to today's activity.	I looked forward to today's activity.
E2. I enjoyed learning new things during today's activity.	I enjoyed learning the class material during today's activity.
E3. I wanted to further understand what I learned about during today's activity.	
E4. I felt good during today's activity.	I felt good about today's activity.
E5. I often felt frustrated during today's activity.	Today's activity made me feel frustrated.
E6. I thought that today's activity was boring.	I thought that today's activity was boring.
E7. I didn't want to do today's activity.	I didn't want to do today's activity.
E8. I didn't care about learning during today's activity.	I didn't care about doing today's activity.
E9. I often felt down during today's activity.	
E10. I was nervous about learning new things during today's activity.	

Table E.2 cont.

Preliminary Items	Pilot Items
Social Items	
S1. I built on other students' ideas during today's activity.	I built on other students' ideas during today's activity.
S2. I tried to understand other students' ideas during today's activity.	I tried to understand other students' ideas during today's activity.
S3. I tried to work with other students who could help me during today's activity.	
S4. I tried to help other students who were struggling during today's activity.	I tried to help other students who were struggling during today's activity.
S5. I didn't care about other students' ideas during today's activity.	I didn't care about other students' ideas during today's activity.
S6. I didn't share ideas when working with other students during today's activity.	I didn't share ideas when working with other students during today's activity.
S7. I didn't like working with other students during today's activity.	I didn't like working with other students during today's activity.

Survey Scales and Item Modifications

Before the overall engagement survey was evaluated, the pilot items from the preliminary analysis were further analyzed to evaluate modified and additional items, as well as to ensure the scales of the individual dimensions functioned well in the new environment (i.e., remote instruction). Both quantitative analysis through the use of CFAs and qualitative analysis through response process interviews (n = 21) were used to assess the scales. Single-factor CFAs for each dimension were analyzed, as well as 3-factor (behavioral, cognitive, emotional) and 4-factor (behavioral, cognitive, emotional, social) correlated models. Additionally, a combined behavioral/cognitive single-factor CFA was evaluated. Final results for single-factor CFAs are included in Table E.3. Data-model fit was assessed through the use of the fit indices CFI, TLI, RMSEA, and SRMR. Recommended cutoffs for good data-model fit by Hu and Bentler (1999) were used to evaluate the scales, in addition to the joint criteria given by Mueller and Hancock (2008) of CFI ≥ 0.96 and SRMR ≤ 0.09. Modification indices (MIs), along with evidence from response process interviews, were used to direct potential modifications of the scales. Omega was calculated for the final scales to provide evidence of acceptable reliability. Details for each of the scales, including initial and final fit statistics and loadings, are provided below (Tables E.4 – E.7).

Table E.3. Fit indices and reliability statistics for single-factor CFAs of final scales for each dimension of engagement. Bolded values indicate results met the suggested criteria based on recommendations from Hu and Bentler (1999).

Scale	**n**	**χ^2 (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**	**omega[b]**
Behavioral	1262	22.770 (5)	<0.001	**0.988**	**0.976**	0.071 **[0.043 – 0.102]**	**0.021**	0.86
Cognitive	1271	19.307 (5)	0.002	**0.987**	**0.974**	**0.055 [0.031 – 0.082]**	**0.022**	0.77
Emotional[a]	1272	88.281 (5)	<0.001	**0.961**	0.922	0.135 [0.111 – 0.161]	**0.034**	0.87
Social	870	2.146 (2)	0.342	**1.000**	**0.999**	**0.011 [0.00 – 0.084]**	**0.011**	0.75
Behavioral/ Cognitive	1255	116.826 (35)	<0.001	**0.977**	**0.970**	**0.053 [0.043 – 0.064]**	**0.028**	0.89

[a] Emotional scale shows evidence of good-data model fit through the use of the joint criteria CFI ≥ 0.96 and SRMR ≤ 0.09 (Mueller & Hancock, 2008).

[b] Although there are no suggested cutoffs for omega, values above 0.7 are generally considered acceptable.

Behavioral Engagement Scale Development

The initial behavioral scale contained 7 items. Overall, two items were removed based on qualitative and quantitative results. Response process interviews indicated that students generally perceived the item, *I didn't actively participate in today's activity*, to be related to social engagement. For example, one student who worked with others said, "*I always actively participate. I usually have my camera on and my mic and I will ask other people for help or tell other people what I'm getting and ask if it's, if it's right or they got the same thing, so I was, I'm pretty active*." Additionally, a student who didn't work with others but did work through the activity responded that they only somewhat agreed with the item, "*Because I was not an active participant. I worked alone*." When this item was retained during analysis of the social-focused B-C-E-S correlated engagement survey, MIs with medium to large effect sizes were found for a possible item-item error correlation between this item and the social item, *I didn't share ideas when working with other students during today's activity*, (w = 0.46) and a possible cross-loading for this item on the social engagement factor (w = 0.33). The item, *I did things I was not supposed to be doing during today's activity*, was removed due to a suggested item-item error correlation with the item, *I stayed focused during today's activity,* through analyzing MIs (w = 0.23). Additionally, during response process interviews, some students found the wording of the item confusing. For example, one student said that, "*Well, I didn't really know what we're not supposed to do, but I assume we were just supposed to do the worksheet and we did the worksheet, so, I think, I think this one we got all the way through.*" The final behavioral engagement scale contained 5 items and the results showed evidence of good data-model fit and acceptable reliability.

Table E.4. Fit statistics and factor loadings for the initial and final behavioral engagement scale. Bolded values indicate results met the suggested criteria based on recommendations from Hu and Bentler (1999). Reverse coded items are noted with (rev).

Fit Statistics								
	n	**χ^2 (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**	**omega**
Initial Fit	1256	100.074 (14)	<0.001	**0.964**	0.946	0.085 [0.070 – 0.102]	**0.036**	--
Final Fit	1262	22.770 (5)	<0.001	**0.988**	**0.976**	0.071 **[0.043 – 0.102]**	**0.021**	0.86

Standardized Factor Loadings		
Item	**Initial**	**Final**
I stayed focused during today’s activity.	0.739	0.715
I put effort into doing today’s activity.	0.771	0.772
I kept working on today’s activity even if something was hard.	0.751	0.771
I didn’t actively participate in today’s activity. (rev)	0.619	--
I did things I was not supposed to be doing during today’s activity. (rev)	0.510	--
I didn’t do much work on today’s activity. (rev)	0.739	0.722
I attempted to answer most of the items on today’s activity.	0.726	0.748

Cognitive Engagement Scale Development

The initial cognitive scale contained 6 items. One item was removed due to qualitative results. The item, *I did just enough to get by during today’s activity*, was removed prior to CFA due to response process interview results. Multiple students indicated that they did not understand what this item was asking and some students responded to the item in a way that contradicted their explanation. Even though this was meant to be a negatively worded item, one student who *agreed* with the item provided an explanation where they listed things related to positive engagement, “*I mean, I sat there the whole class, I did try to work on it, I did every time he came back to talk things out and talk things through, I was here, I was listening, and I was taking my own notes and things like that.*” The final scale contained 5 items and showed evidence of good data-model fit and acceptable reliability.

Table E.5. Fit statistics and factor loadings for the final cognitive engagement scale. Bolded values indicate results met the suggested criteria based on recommendations from Hu and Bentler (1999). Reverse coded items are noted with (rev).

Fit Statistics								
	n	**χ^2 (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**	**omega**
Final Fit	1271	19.307 (5)	0.002	**0.987**	**0.974**	**0.055 [0.031 – 0.082]**	**0.022**	0.77

Standardized Item Loadings	
Item	**Final**
I made sure I understood my work on today’s activity.	0.732
I tried to connect what I was learning during today’s activity to concepts I have learned before.	0.631
I tried to understand my mistakes when I got something wrong during today’s activity.	0.696
I wrote down the answers to today’s activity without trying to understand them. (rev)	0.610
I didn’t think very hard when I came across a challenging problem on today’s activity. (rev)	0.529

Emotional Engagement Scale Development

The initial emotional engagement scale contained 7 items. Two items, *I felt good about today's activity,* and, *Today's activity made me feel frustrated*, were removed. Through CFA, the MIs suggested an item-item error correlation between these two items with a large effect size (*w* = 0.48). Additionally, the MIs suggested an item-item error correlation with the item, *I felt good about today's activity,* and, *I enjoyed learning the class material during today's activity* (*w* = 0.30). In response process interviews, students had varied perceptions of what "feeling good" meant with respect to the activities. Some students mentioned feeling accomplished, satisfied, or not frustrated. However, other students' responses centered around groupwork and how they had anxiety about the activity prior to coming to class but ended up feeling okay about it in the end. For example, one student who *strongly disagreed* with the item stated, "*When we're actually going through the worksheets and stuff, it's, it's fine. It's, it's actually not, not too bad. It just kind of feels like another, another class, you know what I mean…I mean, in terms of the content and stuff, I do, I do feel good about it when we're working on it, it's mainly just the, like, feeling leading up to it.*" Response process interviews about the other item, *Today's activity made me feel frustrated*, showed that not all students saw frustration as negative to their engagement, as was seen in this response from a student who said that they *agreed* with the item, "*So in terms of this activity, I didn't view the frustration as something negative, as something that was bad as a result of the activity. I just saw it as something that is necessary*." A concurrent study focused around students' perceptions of engagement in this environment also found that students' perceptions of frustration in relation to their engagement in the worksheet activities varied (Naibert et al., 2022). The final emotional scale contained 5 items and data collected with these items showed evidence of acceptable data-model fit through the use of joint criteria (Mueller & Hancock, 2008), as well as acceptable reliability.

Table E.6. Fit statistics and factor loadings for the initial and final emotional engagement scale. Bolded values indicate results met the suggested criteria based on recommendations from Hu and Bentler (1999). Reverse coded items are noted with (rev).

Fit Statistics								
	n	**χ^2 (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**	**omega**
Initial Fit	1269	443.838 (14)	<0.001	0.854	0.781	0.185 [0.171 – 0.200]	0.080	--
Final Fit[a]	1272	88.281 (5)	<0.001	**0.961**	0.922	0.135 [0.111 – 0.161]	**0.034**	0.87

Standardized Item Loadings		
Item	**Initial**	**Final**
I looked forward to today's activity.	0.793	0.794
I enjoyed learning the class material during today's activity.	0.786	0.735
I felt good about today's activity.	0.625	--
Today's activity made me feel frustrated. (rev)	0.470	--
I thought that today's activity was boring. (rev)	0.725	0.745
I didn't want to do today's activity. (rev)	0.792	0.822
I didn't care about doing today's activity. (rev)	0.657	0.684

[a]Joint criteria for good data-model fit: CFI ≥ 0.96 and SRMR ≤ 0.09 (Mueller & Hancock, 2008).

Social Scale Development

The initial social engagement scale contained 6 items. The item, *I tried to help other students who were struggling during today's activity*, was removed due to a suggested item-item error correlation with the item, *I didn't share ideas when working with other students during today's activity,* based on MIs (w = 0.42), as well as student responses to this item during response process interviews. When asked to explain their response to this item, many students mentioned that they didn't feel comfortable helping others when they felt they didn't understand the material itself. For example, one student said that, "*I'm not like the ultimately reference point or knowledge point on these things and so when somebody truly just does not understand something, I don't find it my place to really, really dive in with them and try to explain something that I am just now learning. Um, not as a selfish thing, more of like a, I don't want to screw you up, uh, cause I'm not a 100% on this either*." Another item, *I didn't care about other students' ideas during today's activity*, was also removed due to a suggested item-item error correlation with the item, *I didn't like working with other students during today's activity,* (w = 0.28), and student responses in interviews that the wording "I didn't care" seemed harsh. The final social engagement scale contained 4 items and data collected with these items showed evidence of good data-model fit and acceptable reliability.

Table E.7. Fit statistics and factor loadings for the initial and final social engagement scale. Bolded values indicate results met the suggested criteria based on recommendations from Hu and Bentler (1999). Reverse coded items are noted with (rev).

Fit Statistics								
	n	**χ^2 (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**	**omega**
Initial Fit	869	138.997 (9)	<0.001	0.881	0.801	0.161 [0.138 – 0.185]	**0.068**	--
Final Fit	870	2.146 (2)	0.342	**1.000**	**0.999**	**0.011 [0.00 – 0.084]**	**0.011**	0.75

Standardized Item Loadings		
Item	**Initial**	**Final**
I built on other students' ideas during today's activity.	0.697	0.756
I tried to understand other students' ideas during today's activity.	0.600	0.634
I tried to help other students who were struggling during today's activity.	0.763	--
I didn't care about other students' ideas during today's activity. (rev)	0.519	--
I didn't share ideas when working with other students during today's activity. (rev)	0.775	0.660
I didn't like working with other students during today's activity. (rev)	0.536	0.561

Table E.8. Activity Engagement Survey (AcES) descriptive item statistics for the entire aggregated data set (students who worked independently and socially), as well as the social-focused subset for students who worked in a breakout room with others.

Activity Engagement Survey (AcES) Scales and Items	**Aggregated (n = 1248)**				**Social (n = 853)**			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
Behavioral Items								
I stayed focused during today's activity.	4.64	1.04	-0.92	0.86	4.80	0.98	-0.99	1.17
I put effort into doing today's activity.	5.03	0.84	-1.13	2.71	5.13	0.81	-1.28	3.61
I kept working on today's activity even if something was hard.	4.97	0.89	-1.05	1.65	5.07	0.89	-1.18	1.91
I didn't do much work on today's activity.	2.02	0.98	1.07	1.17	1.89	0.93	1.19	1.55
I attempted to answer most of the items on today's activity.	5.14	0.87	-1.23	2.32	5.20	0.84	-1.15	1.57
Cognitive Items								
I made sure I understood my work on today's activity.	4.88	0.91	-0.80	0.99	4.94	0.90	-0.89	1.21
I tried to connect what I was learning during today's activity to concepts I have learned before.	4.83	0.98	-1.02	1.47	4.92	0.95	-1.04	1.69
I tried to understand my mistakes when I got something wrong during today's activity.	5.08	0.81	-0.91	1.37	5.18	0.79	-1.12	2.23
I wrote down the answers to today's activity without trying to understand them.	2.05	1.06	1.30	1.82	1.98	1.03	1.43	2.36
I didn't think very hard when I came across a challenging problem on today's activity.	2.27	1.06	0.83	0.43	2.11	1.01	0.98	0.97
Emotional Items								
I looked forward to today's activity.	3.79	1.20	-0.24	-0.33	3.81	1.24	-0.25	-0.41
I enjoyed learning the class material during today's activity.	4.35	1.06	-0.57	0.43	4.41	1.07	-0.57	0.39
I thought that today's activity was boring.	2.54	1.08	0.66	0.21	2.46	1.07	0.72	0.42
I didn't want to do today's activity.	2.67	1.24	0.61	-0.19	2.57	1.23	0.68	-0.06
I didn't care about doing today's activity.	2.10	1.00	1.04	1.11	2.01	0.99	1.15	1.46
Social Items								
I built on other students' ideas during today's activity.	--	--	--	--	4.65	1.07	-0.91	0.85
I tried to understand other students' ideas during today's activity.	--	--	--	--	5.11	0.83	-1.34	3.47
I didn't share ideas when working with other students during today's activity.	--	--	--	--	2.14	1.17	1.12	0.74
I didn't like working with other students during today's activity.	--	--	--	--	2.06	1.07	1.19	1.41

Appendix F: Supporting Information for Chapter 9

Measurement invariance

Before comparisons can be made between groups, evidence of consequential validity must be gathered through measurement invariance testing. Evidence of measurement invariance was gathered through the use of confirmatory factor analysis (CFA) using maximum likelihood with Satorra-Bentler adjustment and robust standard errors (Satorra & Bentler, 1988). All models were identified by setting the factor loading of one item equal to 1 for each factor while the factor variance was freely estimated. Each step of invariance was assessed by evaluating the change in fit statistics (ΔCFI, ΔRMSEA, ΔSRMR) between two subsequent levels of invariance. Suggested guidelines by Chen (2007) were used to determine if there was sufficient evidence of each level of invariance: ΔCFI ≤ 0.010, ΔRMSEA ≤ 0.015, and ΔSRMR ≤ 0.030 for configural and metric invariance and a smaller ΔSRMR (≤ 0.010) for scalar and conservative invariance. The first level of invariance is configural invariance, in which the same structural model is tested between the groups. Factor loadings between the groups are then set equal to test metric invariance. The next level of invariance is scalar invariance. This includes setting the factor loadings and the intercepts to be equal between the groups. The last level of invariance is conservative invariance, where the loadings, intercepts, and residuals are set equal between groups (Rocabado et al., 2020).

Comparison between independent and social groups

Measurement invariance testing was completed before engagement was compared between responses from students who worked on an activity alone (*independent group*) and those from students who worked on the activity with others (*social group*). The data for each group showed evidence of good data-model fit with the BC-E AcES (Naibert & Barbera, 2022) bifactor model with negative method factor (Table F.1). Measurement invariance results showed evidence of scalar and conservative invariance for group comparisons between the independent and social groups (Table F.2), providing support for comparisons to be made using both latent mean differences and observed score differences.

Table F.1. Data-model fit statistics by group. Bolded values indicate results were good based on recommendations from Hu and Bentler (1999).

Group	**n**	**$\chi 2$ (df)**	**p-value**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**
Independent	389	143.482 (69)	<0.001	**0.969**	**0.952**	**0.057** **[0.044 – 0.071]**	**0.040**
Social	859	187.292 (69)	<0.001	**0.971**	**0.957**	**0.054** **[0.045 – 0.063]**	**0.032**

Table F.2. Fit indices and change statistics for measurement invariance between independent (n = 389) and social (n = 859) groups. Bolded values indicate the change meets the criteria recommended by Chen (2007).

Model	χ2 (df)	p-value	CFI	RMSEA	SRMR	Δχ2 (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR
Configural	335.292 (138)	<0.001	0.971	0.055	0.032	--	--	--	--
Metric	375.663 (170)	<0.001	0.969	0.051	0.039	40.371 (32)	**-0.002**	**-0.004**	**0.007**
Scalar	404.355 (181)	<0.001	0.967	0.051	0.041	28.719 (11)	**-0.002**	**0.000**	**0.002**
Conservative	430.664 (196)	<0.001	0.963	0.052	0.042	26.309 (15)	**-0.004**	**0.001**	**0.001**

Comparison between remote and in-person environments

Measurement invariance testing was completed for responses from the remote and in-person environments during the Fall term. The data for each group showed evidence of good data-model fit with the BC-E-S AcES (Naibert & Barbera, 2022) bifactor model with negative method factor (Table F.3). However, measurement invariance results did not support metric invariance based on suggested recommendations for changes in CFI from Chen (2007) (Table F.4). Since metric invariance could not be supported, scalar and conservative invariance (i.e., higher levels of invariance with additional restrictions) were not tested. Therefore, group comparisons between the two environments could not be justified and were not evaluated (Rocabado et al., 2020).

Table F.3. Data-model fit statistics by group. Bolded values indicate results were good based on recommendations from Hu and Bentler (1999).

Group	n	χ2 (df)	p-value	CFI	TLI	RMSEA [90% CI]	SRMR
Remote Environment	226	174.452 (125)	0.002	**0.967**	**0.954**	**0.048 [0.030 – 0.064]**	**0.054**
In-person Environment	200	180.537 (125)	0.001	**0.950**	*0.932*	**0.055 [0.036 – 0.072]**	**0.061**

Table F.4. Fit indices and change statistics for measurement invariance between the remote environment (n = 226) and in-person environment (n = 200). Bolded values indicate the change meets the criteria recommended by Chen (2007).

Model	χ2 (df)	p-value	CFI	RMSEA	SRMR	Δχ2 (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR
Configural	335.057 (250)	<0.001	0.959	0.051	0.055	--	--	--	--
Metric	433.885 (291)	<0.001	0.944	0.056	0.078	98.828 (41)	-0.015	**0.005**	**0.023**