January 2021

Using Quantitative Methods to Investigate Student Attitudes Toward Chemistry: Women of Color Deserve the Spotlight

Guizella A. Rocabado Delgadillo
University of South Florida

Using Quantitative Methods to Investigate Student Attitudes toward Chemistry:

Women of Color Deserve the Spotlight

by

Guizella A. Rocabado Delgadillo

A dissertation submitted in the partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Chemistry
College of Arts and Sciences
University of South Florida

Major Professor: Jennifer E. Lewis, Ph.D.
Jeffrey R. Raker, Ph.D.
Scott E. Lewis, Ph.D.
Robert F. Dedrick, Ph.D.

Date of Approval:
December 14th, 2020

Keywords: Measurement invariance testing, cognitive, affective, organic chemistry

DEDICATION

"Always remember, you are braver than you believe, stronger than you seem, and smarter than you think." (A. A. Milne)

I dedicate this work to my friends, mentors, and family, particularly my parents. To my dad, who dedicated his life to science and who does not get to see the day his baby becomes a doctor, yet I am certain he is dancing with the angels. To my mom, who has been and continues to be my strength, my support, my biggest cheerleader, my hero, and my joy.

This is for you, with all my love.

ACKNOWLEDGEMENTS

I want to begin by acknowledging the propelling power and influence of a loving Heavenly Father who is merciful and patient with me. God's kindness has been and forever will be the rock of my life and my reason to be. Thank you, Father, for giving me the strength to endure and overcome.

I want to thank my incredible mentors, beginning with my advisor Dr. Jennifer Lewis. Jennifer, thank you for pushing me to learn in unconventional ways and times, for encouraging me to follow my passions, even if it meant talking about organic chemistry, for allowing me to have a voice, and most of all for believing in my talents and abilities even when I could not see them in myself. My committee members have been a vital part of my education. Dr. Dedrick, I want to thank you for patiently teaching me everything I know about measurement. Your love for this subject is transferable to your students, and none of the work that follows could have been done without your support. Dr. Scott Lewis, I cannot thank you enough for the millions of times I have dropped by your office for "a question" that turned out to be three or four questions. Your counsel and support have been foundational to my growth and success in this program. Dr. Jeff Raker, I cannot describe my gratitude for your support and counsel every time I needed it, and for providing a safe space for me to express my ideas and my struggles. Having this safe space has made all the difference for me, especially when things got tough. Honestly, I could not have asked for a better committee of mentors that have worked to help me get through this program with success. Much

like Isaac Newton, I feel like I have accomplished something great only because "I stand on the shoulders of giants."

To all my other teachers at USF, thank you for making my last years of school the best! Especial thanks to Dr. Sarah Kiefer for her passion in teaching and mentoring that go beyond class, and to Dr. Liliana Rodríguez-Campos for her unconditional support.

I also want to thank Dr. Mackay Steffensen and Dr. Ty Redd from Southern Utah University for igniting my passion for organic chemistry and for opening my mind and eyes to the career possibilities outside of medicine. Thank you for being outstanding teachers, mentors, leaders, and for giving me the opportunity later to become a colleague. Your confidence in me has made all the difference in my career path. Thank you for your support all these years. And to all of my SUU family, thank you for being part of my journey always!

I want to thank Bernard Batson in the Engineering department at USF for his relentless encouragement during my time at USF beginning with invitations and financial support to countless professional development workshops, conferences, and socials, as well as for supporting me with a two-year Florida-Georgia Louis Stokes Alliance for Minority Participation fellowship. I could not have continued my education without this support. Thank you, Mr. Batson!

I want to thank all the USF chemistry office staff, who work so hard to make sure all the grad students get paid, have assignments, are on track, etc. I want to thank you because your work is vital and I appreciate you. Especial thanks to Ryan Jahn and Kaitlyn Kroner who are so incredibly patient with me and are always willing to help me.

Thank you fellow USF CER and non-CER friends! To my group, Jacob and Stephanie, thank you for your support and the infinite times you've spent reading my manuscripts over and over. Thank you Ying Wang, Dr. Amber Dood, Dr. Rebecca Gibbons, Brandon Yik, Jamie

Nunziata, AJ Sona, Dr. Vanessa Ralph, Dr. Justin Pratt, Aaron Clark, Ayesha Farheen, JD Young, James Kingsepp, and Md Tawabur Rahman for making my years at USF fun and for begin a huge social support, from trivia nights, to random walks around campus, to all-you-can-eat sushi birthday celebrations, to quarantine movie nights via zoom. Thank you, friends, you have been a lifeline!

Thank you Dr. Sara Moulton for being an amazing friend, and for helping me figure out factor analysis. Thank you for listening to my presentations, reading my manuscripts, and offering valuable measurement advice. You are truly brilliant and I cannot thank you enough for your support. Thank you Lilian Montes for being a great friend and for agreeing to collaborate with me in an amazing project which has brought two countries together! Thank you to all my collaborators for your support, and for pushing me to be a better researcher and writer.

Thank you to my SACNAS chapter at USF, which has given me purpose as well as a fun and a meaningful outlet to connect with like-minded people from other disciplines. Michelle, Amber, Lilyanna, Dr. Cruz, and all SACNistas, thank you for going on this journey with me!

To my friends, Mary, Ashley, and Leila, thank you for being there for me throughout all my life! You are my rock forever! And thank you to so many others friends who have helped me and been there for me while I pursued my education.

Finally, I thank my family for their endless support. There are no words that can truly say how blessed I am to have a family that is full of love and encouragement. Matthew, Amy, Josué, Ammón, Sarah, Laura, Lizzy, and Emma, you are my reason to live and my motivation to continue to be a better person each day. The strength and wisdom you already have in your youth gives me confidence in a bright future. Susy, Paul, Pily, Scott, Marcia, and Dani, thank you for your endless wisdom and counsel. Thank you for being an example for me to follow and for loving me and

supporting me without conditions or reservations. Also, thank you for making me laugh and being the source of true happiness in my life.

Dad, although you are no longer here, I want to thank you for the legacy you have left me. Your love of science has permeated my life in so many ways. But more importantly, your love and dedication to people coupled with your impeccable integrity are the pillars of my life. I hope I am making you proud and I can continue your legacy. I love you eternally.

Mom, thank you for sharing this adventure with me. I can't thank you enough for taking care of all of my temporal needs so I could fully focus my time on school and research. Also, thank you for providing much needed rest from work to spend time together. But most importantly, thank you for teaching me since I was a little girl about social justice, self-advocacy and hard work. You have taught me to fight for what is right, to fight for justice, and to pursue my dreams even against overwhelming odds. You taught me the power of asking, and also the practice of leaving no stone unturned. You have helped me be strong and resilient. Truly, you have influenced every aspect of my being. Thank you for being the greatest example of love, sacrifice, and humanity. You are the best part of my life and I love you infinitely.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

ABSTRACT

The field of Chemistry Education Research (CER) has been interested in understanding the reasons why students struggle in organic chemistry courses. Reports show that students perceive the material as difficult and have trouble keeping pace with the volume of content taught within the course. Beyond these and other explanations of why students struggle in organic chemistry are affective factors, such as attitude toward chemistry, that influence students' success and retention in this course. Studies have shown that in many instances, underrepresented groups of students report less positive attitudes than their peers. Of greater concern are the students representing multiple marginalized identities, such as Women of Color, who may display even less positive attitudes, which in turn may influence their success and retention in STEM fields.

Investigating whether attitude trends observed in organic chemistry classrooms with an intervention (chapter 3) or without an intervention (chapter 5) extend to a group of Women of Color (*i.e.,* Black female students) is an important focus of the work presented herein. Evaluating attitude gains or losses over the course of the semester can help researchers and practitioners continue to improve pedagogies that influence both cognitive and affective domains of learning. Additionally, the focus on investigating the impact of these pedagogies on underrepresented groups of students, particularly Women of Color, is paramount to answer the call of increasing diversity, inclusion, and equity in STEM.

Studies included in this dissertation have shown that in traditional lecture organic chemistry courses, attitude toward chemistry tends to remain constant or decline over a semester (chapter 5). On the other hand, pedagogical interventions, such as flipped classroom, can produce a positive gain in attitude across a semester (chapter 3). In order to determine whether these attitude gains or losses extend to subgroups of students within the classroom, measurement invariance testing (chapter 4) was utilized to provide support for the desired comparisons. When quantitative studies are conducted with an effort to learn about the similarities or differences of groups within the same learning environment, strict measurement standards must be used in order to safeguard against threats to the validity of inferences that might favor one group over another. Chapter 4 provides a step-by-step tutorial on how to conduct measurement invariance testing when group comparisons or longitudinal comparisons are desired. This technique was utilized throughout this work to ensure comparisons were supported (chapters 3, 5, and 6).

Additionally, chapter 6 reports the process of refinement and development of a new instrument to measure attitude that includes an *emotional satisfaction* factor and a *utility* factor. This instrument was developed simultaneously in English and Spanish. It was administered in the U.S. and in Chile in order to demonstrate its function in both languages and in different countries. Evidence shows that the internal structure of the instrument holds in both contexts, and although comparisons are not supported, metric invariance was achieved indicating similar factor meaning across the two groups.

CHAPTER 1

INTRODUCTION

Researchers in the field of Chemistry Education Research (CER) have long been interested in investigating how students learn (Posner *et al.,* 1982; Bodner, 1986; diSessa, 1988; Johnstone, 1993; Bretz, 2001; Chi, 2005), how they solve problems (*e.g*., Bodner and Herron, 2002; Clair-Thompson, Overton, and Bugler, 2012; Bodner, 2015; Crandell *et al*., 2019; Dood *et al.,* 2019), how they make sense of chemistry content (*e.g*., Cooper, Williams and Underwood, 2015; Graulich, 2015; Wang and Lewis, 2020), how they feel about chemistry (*e.g*., Pekrun, 2006; Bauer, 2008; Spagnoli *et al*., 2017; Gibbons *et al*., 2018) and their place as consumers of knowledge in our classrooms (Entwistle, 1991; Carlone and Johnson, 2007; Hazari, Sadler and Sonnert, 2013; Fink, Frey and Solomon, 2020; Hosbein and Barbera, 2020). These research foci are all important as researchers, practitioners, administrators, and policy makers make critical decisions that impact students' experiences, attainment of knowledge and skills, and movement toward making a contribution to society in the STEM workforce and health professions. However, little is known of the impact these decisions have on subgroups of students.

A call for increasing the science, technology, engineering, and mathematics (STEM) academia and workforce in the U.S. was made by President Obama in 2010 indicating that this plan included diversifying STEM (Obama, 2010; Seadler, 2012). Recent work has attempted to

heed this call by focusing on assessing differential impacts of instruction on underrepresented groups (URG) of students of various backgrounds (*e.g.,* Ballen *et al*., 2017; Stanich et al., 2018). Many studies have attempted to investigate differential impacts with comparisons between a heterogenous URG against non-URG (*e.g*., Fink *et al*., 2018; Harris *et al*., 2020), with the expectation to learn about the 'gap' that needs closing between these two groups; yet, the problem of underrepresentation remains. With comparisons such as these where students of various diverse backgrounds are aggregated within the same group (*i.e.,* all URG), we come short of understanding differences that may exist between distinct intersections of identity (Crenshaw, 1989; Thomsen and Finley, 2019) such as Black male or Hispanic female students. Understanding these differences may become crucial in our decisions as educators, administrators, and policy makers when intending to create diverse and inclusive spaces for our students.

Furthermore, most of these efforts are initiated under the assumption that URGs are deficient in performance, affect, and retention; thus, our efforts should attempt to 'close the gap' (Harris et al, 2020). This deficit mindset, while vastly utilized in our educational system, propagates social injustice for Students of Color who are viewed as deficient and in need of "fixing" (Bourdieu and Passeron, 1977; Sullivan, 2001). A change in mindset is necessary for all researchers, practitioners, administrators, and policy makers to truly implement initiatives that will serve social justice (Yosso, 2005; García *et al.,* 2018; Gillborn *et al*., 2018) and increase diversity and inclusion for URGs.

With the purpose of increasing diversity and inclusion in STEM spaces, I have undertaken the work that will be described in this dissertation, with a special concern for Women of Color

navigating organic chemistry classrooms, motivated by my own experiences in chemistry courses. This special focus on Women of Color is due to the lack of studies in CER that examine gender and race/ethnicity intersectional identities which are vital in understanding how subgroups of students may experience our classrooms. Thus, this dissertation pays attention to Women of Color who have learned or are learning to navigate STEM spaces by bringing to light even a small aspect of their perceptions while in our chemistry classrooms. This work represents my evolution as a student, researcher, and future practitioner as I questioned and challenged some of my own biases and deficit mindset (and continue to do so), and looked for ways to better serve Women of Color in my work and in my future career.

The purposeful attention to Women of Color and other relevant subgroups according to the context of each study is vital and reveals the significance of centering these groups in our studies to investigate their experiences, strengths, and needs to better support them. Moreover, each study presented herein provides empirical evidence of the utilization of appropriate quantitative methods to be utilized for subgroup comparisons in future studies in CER. The methods shown in this work (discussed in chapter 2) fall within a Classical Test Theory (CTT) framework, which I have endeavored to make accessible to both researchers and practitioners.

I have centered my studies in measuring attitude toward chemistry. Although investigating student understanding and learning strategies (Posner *et al.,* 1982; Bodner, 1986; diSessa, 1988; Johnstone, 1993; Bretz, 2001; Chi, 2005) is important, I have chosen to investigate student affect in an effort to gain greater insight into an understudied aspect of learning, which has also been shown to relate to metrics of achievement and retention (Pekrun, 2006; Halpern *et al.*, 2007;

Gibbons *et al.*, 2018). Focusing on measuring attitude for subgroups of students (*i.e.*, Women of Color) can bring challenges, such as sample size issues, but is important to study because it allows us to learn about perceptions and feelings of these subgroups in the classroom and how these perceptions and feelings can impact their trajectories through our chemistry classrooms and beyond.

Attitude

Attitude is a construct that has been investigated for over a century. Attitude theories emerged early in the 1920's with the earliest recorded mention of the term 'attitude' in 1862 by Herbert Spencer (Allport, 1985), although there are some ancient Greek philosophies that allude to what we know as attitude now (Oskamp and Schultz, 2005). In the last century, many theorists have defined attitude in general as perceptions toward an attitude object, such as science. For example, Bern (1970) described attitude as "likes and dislikes" (p. 14). Eagly and Chaiken (1993) defined attitude with an emphasis on the role of evaluation of an attitude object with "some degree of favor or disfavor" (p.1). Fishbein and Ajzen (1975) highlighted the idea that attitude is a "learned predisposition to respond in a consistently favorable or unfavorable manner" toward an attitude object (p. 6). Relatedly, Allport (1935) offered a broad definition of attitude as "a mental or neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related" (p. 810). The consistent idea of attitude as an organization of mental processes, cognitive and emotional, with respect to certain aspects of a person's world or attitude objects (Krech,

Crutchfield, and Ballanchey, 1962), permeates all of the aforementioned definitions. These mental processes, have been repeatedly conceptualized as "evaluations" of an attitude object (Eagly and Chaiken, 1993; Ajzen, 2001), which are noted to surge spontaneously from experiences with the attitude object (Ajzen, 2001). Fazio (1986) proposed that attitudes are automatically formed as a function of exposure to stimuli that incite conscious evaluative mental processes. Chronic exposure to an attitude object can produce stable attitudes toward that object; however, certain contextual factors could influence a change in attitude over time (Ajzen and Sexton, 1999; Reid, 2006). Change in attitude over time is a concept which will be explored in and chapters 3, 5 and 6.

Many attitude theorists have hypothesized that attitude is comprised of affective, cognitive, and behavioral subcomponents (Krech, Crutchfield, and Ballanchey, 1962; McGuire, 1969, Bagozzi and Burnkrant, 1979; Breckler, 1984; Eagly and Chaiken, 2007; Rosenberg and Hovland, 1960). Decades of research have been dedicated to discuss whether these subcomponents are so highly correlated that one cannot measure them distinctly (McGuire, 1969), or whether the subcomponents are correlated yet discrete (Krech, Crutchfield, and Ballanchey, 1962; Bagozzi and Burnkrant, 1979). The latter view has been labeled the tripartite model of attitude and has been largely employed by theorists and researchers since the mid 1900's (*e.g.*, Bagozzi, Tybout, Craig and Sternthal, 1979; Breckler, 1984; Eagly, Mladinic and Otto, 1994; Huskinson and Haddock, 2004).

Figure 1.1a shows the tripartite model of attitude, but Figure 1.1b shows a more nuanced model that emerged from questioning the tripartite model. This more complex model indicates that while the affective and cognitive aspects of attitude are correlated, each contributes separately to behavioral intentions (Ajzen and Fishbein, 2000). Furthermore, the cognitive and affective

components can be present in different magnitudes according to the character of the attitude object (Kempf, 1999). For example, attitude toward a videogame can be dominated by the affective component of attitude with feelings of satisfaction and enjoyment. But attitude toward the hardware that allows the use of the videogame can be dominated by the cognitive aspect that portrays an acknowledgement of the complexity with which it was created as well as its utility. In other words, attitude is comprised of cognitive and affective components which can be present in different levels. These related aspects of attitude contribute to a person's intentions to behave in a certain manner toward the object of attitude; however, they are not the only factors that influence intentional behaviors or behavior itself. Thus, it is thought that behaviors involve a separate process, related to attitude, but not a direct component of attitude (Allport, 1954; Ajzen and Fishbein, 2000).

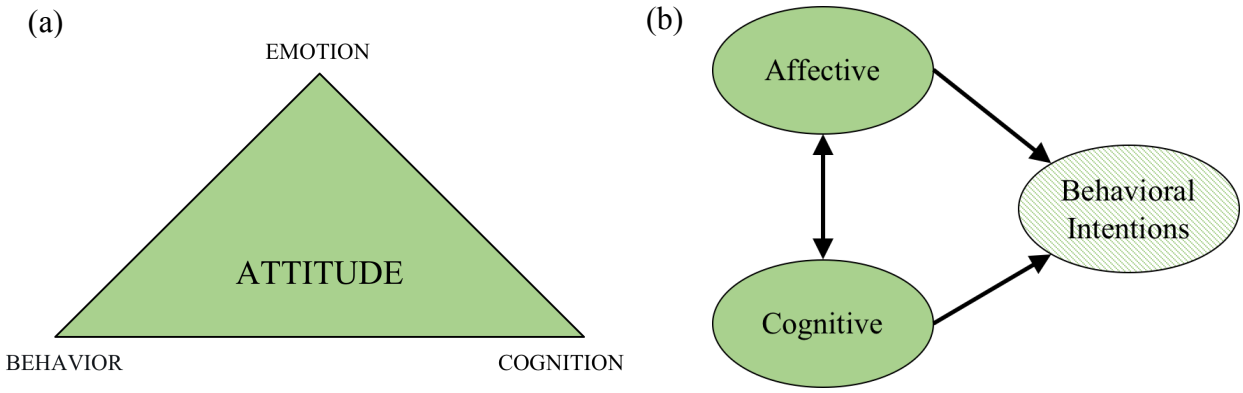


Figure 1.1. (a) The tripartite model of attitude. (b) Attitude model derived from the tripartite model. Attitude components (solid green ovals) are correlated. Each component of attitude can contribute to behavioral intentions (patterned green oval) which is a related, but separate process from attitude.

Why Do We Study Attitude?

Early in the 1900's theorists in the field of social psychology were divided between those who thought attitude was a worthy concept to study and those who thought that time would be better spent studying concepts that could be observed and measured directly (Oskamp and Schultz, 2005). Behaviorists such as Bain (1928) and Skinner (1957) thought the study of attitude to be a 'hinderance' to the advancement of the field. However, the concept of attitude has captivated dozens of theorists and even more researchers because it is a concept that encompasses a wide range of mental processes that can lead to or predict behaviors (Oskamp and Schultz, 2005). Therefore, researchers interested in most aspects of social psychology can benefit from the study of attitude to understand the underpinnings of human behavior through the study of the mental processes that can lead to those behaviors.

Educators can also benefit from the investigation of student and teacher attitude. The study of attitude in educational settings has gained traction in the last few decades as it has been shown to positively relate to metrics of achievement and retention (Koballa, 1988; Kanadli, 2016; Vilia *et al*., 2017). It has proven particularly useful to study attitude in longitudinal studies to examine its effects on achievement and retention which will be detailed in chapters 3, 5 and 6. And although achievement and retention are complex topics of investigation, attitude has been shown to positively impact these important educational metrics and thus is a topic worth exploring.

Additionally, as educators, we care about students' attitude toward chemistry because of the societal impacts these attitudes can have (Ramsden, 1998). When students step into our

chemistry classrooms they are chronically exposed to the attitude object and can form stable attitudes toward it (Ajzen and Sexton, 1999; Reid, 2006). Once they step out of our classrooms and stop their direct exposure to the field of chemistry, very little can be done about influencing their attitude toward chemistry, which can then influence their perceptions and ultimately their behaviors toward the role of chemistry in society (*i.e*, pharmaceuticals, environmental effects, etc.). Thus, having the opportunity to help students develop more permanent positive attitudes toward chemistry when they participate in our classrooms warrants the study of how attitudes are formed and how we can positively influence them.

Attitude Toward Science

In the U.S. and in the U.K. studies of attitude toward science arose simultaneously in the 1960's and 1970's owing to curriculum reform movements and important scientific advances with global impact, such as the first Sputnik satellite (Reid, 2006). Ormerod and Duckworth (1975) provided a review of about 500 studies on students' attitude toward science. In that review of mostly biological and physical science disciplines at the time, various issues that impact students' attitude were discovered, such as gendered differences, effectiveness of various learning strategies, and the importance of a supporting role of the science teacher, among others (Ormerod and Duckworth, 1975). In another review article by Ramsden (1998) it was shown that conclusions about science included that science was perceived as 'not relevant' to most people, science produces environmental damage, there are gendered differences in the attractiveness of science favoring males over females, students tend to lose interest in science in high school and beyond,

and physical sciences were viewed more negatively than other science disciplines. These general negative attitudes toward science, particularly physical science remain and investigations to address some of these attitudes are currently an important focus (*e.g.*, Lewis *et al.*, 2016; Vilia *et al.,* 2017).

Research in science education has included the construct of attitude for the purpose of further understanding the dynamic process of the attitude-achievement relationship (Salta and Tzougraki, 2004; Vilia *et al.*, 2017). Koballa and Crawley (1985) stated that attitude is an important focus of study in science fields because it can allow for predictions of behaviors toward science. It is known that achievement can be influenced in a variety of ways, with attitude being a significant affective predictor (Brandriet, Ward, and Bretz, 2013; Xu, Villafañe, and Lewis, 2013; Villafañe and Lewis, 2016; Rocabado *et al.*, 2019). Thus, implementing pedagogical strategies that can improve attitude, such as active learning and flipped classrooms can also improve student achievement in the course (*e.g.*, Mooring *et al.*, 2016). Another purpose of investigating attitude in secondary and post-secondary science courses is the attitude-retention relationship (Halpern *et al.,* 2007). The issue of retention is of great concern due to the increasing numbers of students who leave science fields (Seymour and Hewitt, 1997; Osborne, Simon, and Collins, 2003; Seymour and Hunter, 2019). Understanding the relationship of attitudes to achievement and retention is a necessary endeavor. Thus, this work aims to continue to explain the attitude-achievement (see chapters 3 and 6) and attitude-retention relationships (see chapter 5) with particular focus on the attitudes of Women of Color within organic chemistry courses.

Measuring Attitude

In order to investigate attitude-achievement and attitude-retention relationships, researchers and practitioners must be able to measure all of those dimensions as well as their connection with each other. In a review article, Osborne, Simon, and Collins (2003) argue that in order to measure attitude toward an attitude object, such as chemistry, we must first operationalize attitude toward chemistry and identify the various elements that may influence this construct, for instance, gender, race/ethnicity, classroom climate, etc. Additionally, we must consider the various ways, both qualitative and quantitative, in which attitude could be measured.

Attitude can be investigated qualitatively through in-depth interviews (Curtis and Curtis, 2017). By employing this method, researchers can expand upon various aspects of attitude and gather participant narratives about their experiences, thoughts, and feelings toward the attitude object (*i.e.*, chemistry). Interviews provide rich, in-depth information, although they are time-consuming and participants tend to be few in numbers. Another qualitative method used to study attitude is open-ended questions (Campbell, 1971). With this approach, the number of participants may increase; however, the responses may be less rich compared to interviews.

On the other hand, attitude can be investigated quantitatively through self-report, multiple choice instruments (Cook and Selltiz, 1964). By utilizing this method, researchers can investigate data for large numbers of participants as well as a generalized notion of the perceptions of an entire group of people (*i.e.*, classroom). Although quantitative methods are not suited to study individual lived experiences, they are widely used in education research to investigate construct relationships,

make group comparisons, and much more. Often chemistry classrooms are taught as large lectures, thus quantitative methods become more practical to investigate attitudes of large numbers of students in a rigorous way. Additionally, chemistry faculty in general may be more familiar and comfortable utilizing quantitative methods in their chemistry classrooms to investigate their students' attitudes toward the discipline. Thus, creating appropriate methods and useful instruments to measure attitudes has been an important focus during the last century. Thurstone's (1928) equally appearing intervals approach was one of the first attempts to quantitatively measure attitude. This method aimed to measure attitudes in discrete levels that were precisely the same distance apart. Shortly after, Likert (1932) suggested a less cumbersome system to construct scales for attitude measurement. This method expanded from a "yes or no" response to an extent or level of agreement or disagreement with a statement about attitude (Oskmap and Schultz, 2005). Almost a century later, Likert's method is currently one of the most widely used in survey research. Likert's approach gave place for other researchers to design different scaling methods. Guttman's (1944) cumulative method was proposed to create respondent's cumulative scores with unique meanings, which is in contrast to Thurstone's and Likert's scales (Oskamp and Schultz, 2005). Another type of scale that is often used to measure attitude is Osgood *et al*.'s (1957) semantic differential. This scale is convenient, since it can be applied to any attitude object and does not rely on opinion items, rather in the connotative meaning of the attitude object (Osgood *et al.,* 1957). Based on the early research done with semantic differential scales, Osgood (1965) discovered that an evaluative dimension is the most recommended way to measure attitude toward an attitude object, such as chemistry.

ASCI, ASCIv2, ASCIv3, and ASCI-UE

Currently there are several published instruments that measure attitude in chemistry. These instruments include Adam's (2008) Colorado Learning Attitudes about Science Survey (CLASS) for the use in chemistry, Fraser's (1977) Test of Science-Related Attitudes (TOSRA), Cheung's (2011) Attitude Toward Chemistry Lessons Scale (ATCLS), and Bauer's (2008) Attitude toward the Subject of Chemistry Inventory (ASCI), as well as its shortened version the ASCIv2 (Xu and Lewis, 2011). These instruments have been used in chemistry classrooms in the last two decades to measure attitude toward chemistry (Heredia and Lewis, 2012; Navarro *et al*., 2016; Villafañe and Lewis, 2016). Of these, the ASCI and subsequent versions of this instrument are the focus of the work presented in this dissertation.

The ASCI was originally developed by Bauer in 2008 and is a 20-item semantic differential instrument with five proposed subscales. Many researchers have used this instrument in their classrooms (Brown *et al*., 2014; Chan and Bauer, 2015; May *et al*., 2018; Ross, Nuñez and Lai, 2018) and laboratories (Hensen and Barbera, 2019; An, Poly and Holme, 2020). In 2011 Xu and Lewis refined the ASCI and presented a two-factor, eight-item version of the instrument (ASCIv2) that measured a cognitive dimension (intellectual accessibility), and an affective dimension (emotional satisfaction) following the theoretical underpinnings of the attitude construct (Ajzen and Fishbein, 2000). This instrument exhibited good psychometric properties evidenced by factor analytic techniques (Xu and Lewis, 2011). Xu (2010) also created several other versions of the ASCIv2 by introducing modifications such as altering the item order (ASCIv3), or changing the attitude object to calculus (ASCIv3.1). The stability of the two-factor model described for these

altered versions of the instrument came in question when the item loadings were inconsistent and factor analysis fit indices indicated poor model fit (Xu, 2010). Thus, further analyses and modifications may be needed for versions 3 and 3.1 of this instrument.

Since its inception, the ASCIv2 has been widely used in chemistry classrooms in the United States (Brandriet *et al*., 2011; Xu and Lewis, 2011; Brandriet, Ward, and Bretz, 2013; Xu, Villafañe, and Lewis, 2013; Cracolice and Busby, 2015; Chan and Bauer, 2014, 2016; Mooring *et al*., 2016; Underwood, Reyes-Gastelum, and Cooper, 2016; Stanich *et al*., 2018; Nenning *et al.,* 2019), in Australia (Xu, Southam, and Lewis, 2012; Vishnumolakala *et al*., 2017; Vishnumolaka *et al*., 2018) in the Phillipines (Damo and Prudente, 2019), in Saudi Arabia (Xu, Alhoosani, Southam, and Lewis, 2015), in Chile (Montes, Ferreira, and Rodriguez, 2018), and in Turkey (Khaveci, 2015; Sen, Yilmaz, and Temel, 2016) and has been translated to various languages to serve students around the globe. On occasion, the ASCIv3 has also been utilized (Xu, 2010; Rocabado *et al.,* 2019). Additionally, by changing the attitude object to mathematics, this instrument was also used to probe the attitude of life science students toward math, yet it was apparent that the two-factor structure that holds when the attitude object is chemistry is not the same when it changes to mathematics (Wachsmuth et al., 2017).

Largely, the ASCIv2 has been utilized longitudinally with the purpose of investigating or evaluating the impact of pedagogical interventions on attitude (*e.g.*, Underwood, Reyes-Gastelum, and Cooper, 2016; Stanich *et al*., 2018; Vishnumolaka *et al*., 2018). Many of these studies tested an intervention group and a non-intervention group and investigated the differences in attitude and achievement of both groups (*e.g.*, Mooring *et al*., 2016). Few studies presented longitudinal

investigations of attitude in a non-intervention classroom (*e.g.*, chapter 5). Additionally, very few studies have investigated the attitude-retention relationship in chemistry classrooms by utilizing this instrument (*e.g.*, chapter 5). Finally, few studies to date have investigated specific intersectional groups as the focus of the study (*e.g.*, Villafañe, Garcia, and Lewis, 2014), but to my knowledge, none other than the work presented herein (particularly chapters 3 and 5), has investigated attitude for Women of Color in chemistry classrooms.

Additionally, chapter 6 informs of the process of development of yet another version of the ASCI in both English and Spanish. Here I demonstrate the process of instrument development following the guidelines prescribed by the *Standards for Education and Psychological Testing* (Arjoon *et al.,* 2013; AERA *et al.,* 2014). This process included cognitive interviews with chemistry students in the U.S. and in Chile as well as a panel of experts in chemistry, CER, and attitude theory. From this process, a new two-factor, nine-item instrument emerged, the ASCI-UE, which contains a revised *emotional satisfaction* scale and a new *utility* scale. The ASCI-UE has been used in chemistry classrooms in the U.S. (Wang *et al*., 2020) and has also been translated to Spanish and used with university chemistry students in Chile (Chapter 6).

Subgroup Comparisons

The significance of investigating subgroups of students in our classrooms is that students of various backgrounds and intersectional identities experience shared events (*i.e.*, a chemistry course) in different ways (chapter 5), some of which may greatly influence their subsequent

decisions and behaviors. Catsambis (1995) stated that female students held more negative attitudes toward science than male students when controlling for variables like background and achievement. Moreover, Catsambis (1995) argued that the gendered effect on attitude was felt more strongly among Black female students. Thus, it is not enough to investigate subgroups based on gender *or* race, but it becomes pertinent to investigate groups that embody the "double bind" (Ong *et al.,* 2011) at the different intersections of gender, race, ethnicity, first generation status, or other relevant background identities (Crenshaw, 1989; Else-Quest, Mineo, and Higgins, 2013; Litzler, Samuelson, and Lorah, 2014; Ireland *et al.*, 2018).

An important focus of research and practice should become the careful measurement of attitude for these subgroups of students to avoid propagating systemic biases and social inequities (García *et al.*, 2018; Gillborn *et al.,* 2018). Taking every available step to safeguard against inappropriate inferences in our measurement is essential to diversity, inclusion, equity, and social justice initiatives in education in general and in our classrooms specifically. Chapter 4 in this work demonstrates measurement invariance testing, which is one quantitative method with various steps to check that group comparisons with instruments measuring latent variables (*i.e.*, attitude) are performed appropriately, checking at every level that the groups being compared answered the same instrument in similar ways (Rocabado *et al*., 2020). This technique is one method, among others, to provide support to conduct subgroup comparisons in a research study.

Few studies in CER have taken the opportunity to disaggregate data with the purpose of subgroup comparisons (*e.g*., Villafañe, Garcia, and Lewis, 2014; Rocabado et al., 2019; chapter 5), although some have investigated URG versus non-URG (*e.g*., Fink *et al*., 2018; Harris *et al*.,

2020). As stated before, the problem with investigating URG and non-URG is that this practice omits the opportunity to learn about specific groups and intersections. Furthermore, investigating particular intersections to consider the experiences of diverse groups (Crenshaw, 1989; Thomsen and Finley, 2019) is crucial to the drive for greater inclusion in our classrooms. By inspecting the experiences of different groups, the field of CER can move toward addressing issues of lower achievement and retention rates among URG, particularly Women of Color. This goal can be accomplished when we are able to understand and meet the needs of each particular subgroup, as well as acknowledge and support their strengths (Kretzman and McKnight, 1993; Yosso, 2005; Donaldson and Daugherty, 2011; Cantú, 2012; Peralta, Caspary, and Boothe, 2013; Myende, 2015; Rodriguez, Cunningham, and Jordan, 2019).

Organic Chemistry as Context for Studies

Students in most science and health-related majors are required to take organic chemistry during their undergraduate years (Barr *et al.,* 2008; Cooper, Grove and Underwood, 2010). This course is perceived as one of the most difficult in the undergraduate curriculum (Rowe, 1983; Barr *et al.*, 2010; Horowitz, Rabin and Brodale, 2013), and thus many students begin organic chemistry with fear (Flynn, 2015). Compounding to their fear, students face significant barriers in understanding the complex content in a fast-paced course (Fautch, 2015). Therefore, much of the focus in CER has been to uncover some of the difficulties in understanding the material that students experience in organic chemistry (*e.g.*, Anderson & Bodner, 2008; Cooper et al., 2010;

Grove & Bretz, 2010; Kraft et al., 2010; Anzovino & Bretz, 2015; Dood *et al.*, 2019; Crandell, Lockhart and Cooper, 2020).

On the other hand, only few studies have shed light on affective aspects of learning in organic chemistry (*e.g.*, Black and Deci, 2000; Liu, Raker and Lewis, 2018). Taber (2015) states that learning is a combination of cognitive and affective processes and researchers would do well to investigate not only student understanding but also students' perceptions and experiences as they engage in learning. Similarly, Gibbons *et al*. (2018) emphasized that "learning is an emotional experience" (p. 838). Thus, it becomes vital to provide insight into the affective domain of learning in a highly emotional environment such as organic chemistry. To address this issue, this work provides several studies that focused on attitude toward chemistry in organic chemistry classrooms with the purpose of investigating achievement emotions (Pekrun, Maier and Elliot, 2009) in this course while also considering the differential impacts for Women of Color.

Overview of Studies

This dissertation is composed of four distinct studies which comprise chapters 3-6. Chapter 3 (Rocabado *et al*., 2019) emerged from the curiosity of investigating whether the positive attitude gains observed in a flipped organic chemistry course demonstrated by Mooring *et al*. (2016) extended to the Black female students in the class. In this study, methods like measurement invariance testing and structural equation modeling were utilized to explore the feasibility of group comparisons as well as the investigation of the attitude-achievement relationship. Through this

first study, a realization of the underuse of measurement techniques in the field of CER became apparent, thus chapter 4 (Rocabado *et al*., 2020) provides a step-by-step tutorial on measurement invariance testing including software code for the reader to perform this technique. In this article I began to explore more deeply and actively challenge a deficit mindset, realizing that 'numbers are not neutral' and researchers along with their research are not objective nor without biases (García *et al.*, 2018; Gillborn *et al*., 2018). Therefore, the constant checking and safeguarding against the propagation of social injustices through my research became an important focus. Chapter 5 explores attitude trends among Hispanic and White female students in an organic chemistry classroom. This study demonstrates a challenge of a deficit mindset and the evolution of the study through the process of reflection and awareness of the tenets of QuantCrit (García *et al*., 2018; Gillborn *et al.,* 2018). Lastly, chapter 6 (previously unpublished) explains the process of adaptation and refinement of the ASCIv2 in English and Spanish to include a dimension of utility, which emerged from cognitive interviews with students in the U.S. and in Chile. Following the guidelines of instrument development (Arjoon *et al*., 2013; AERA *et al*., 2014), several aspects of validity evidence were gathered resulting in a new instrument (ASCI-UE) which is proposed as a candidate to measure attitude with dimensions of *emotional satisfaction* and *utility* in chemistry classrooms. Each of these studies focus on the importance of subgroup comparisons to further diversity and inclusion initiatives in STEM fields.

References

AERA, APA and NCME, (2014), *Standards for educational and psychological testing*, Washington, DC: American Psychological Association.

Adams W. K., Wieman C. E., Perkins K. K. and Barbera J., (2008), Modifying and validating the Colorado learning attitudes about science survey for use in chemistry, J. Chem. Educ., 85(10), 1435-1439. DOI:10.1021/ed085p1435

Ajzen I., (2001), Nature and operation of attitudes, *Annual Rev. Psychol.,* **52**, 27-58. DOI:10.1146/annurev.psych.52.1.27

Ajzen L. and Fishbein M., (2000), *Attitudes and the attitude-behavior relation: Reasoned and automatic processes*. In European review of social psychology (Vol. 11) Stroebe W. and Hewstone M. (Eds.), Chichester, England: Wiley, pp. 1-33.

Ajzen L., and Sexton J., (1999), *Depth of processing, belief congruence, and attitude-behavior correspondence*. In Dual-process theories in social psychology, Chaiken S. and Trope Y. (Eds.), New York: Guilford, pp. 117-140.

Allport G. W., (1935), *Attitudes*. In A handbook of social psychology, Murchison C. (Ed.), Worcester, MA: Clark University Press, pp. 798-844.

Allport G. W., (1954), *The nature of prejudice.* Reading, MA: Addison-Wesley

Allport G. W., (1985), *The historical background of social psychology*. In The handbook of social psychology, (3rd ed., Vol. 1), Lindzey G. and Aronson E. (Eds.), New York: Random House , pp. 1-46.

An J., Poly L.-P. and Holme T. A., (2020), Usability testing and the development of an augmented reality application for laboratory learning, *J. Chem. Educ*., **97**, 97-105. DOI:10.1021/acs.jchemed.9b00453

Anderson T. L. and Bodner G. M., (2008), What can we do about 'Parker'? A case study of a good student who didn't 'get' organic chemistry, *Chem. Educ. Res. Pract.,* **9**, 93-101. DOI:10.1039/B806223B

Anzovino M. E. and Bretz S. L., (2015), Organic Chemistry Students' Ideas About Nucleophiles and Electrophiles: The Role of Charges and Mechanisms, *Chem. Educ. Res. Pract*, **16**, 797−810. DOI:10.1039/C5RP00113G

Arjoon J. A., Xu X. and Lewis J. E., (2013), Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence, *J. Chem. Educ*., **90**, 536–545. DOI:10.1021/ed3002013

Bagozzi R. P. and Burnkrant R. E., (1979), Attitude organization and the attitude-behavior relationship, *J. Pers. Soc. Psychol.,* **37,** 913-929. DOI:10.1037/0022-3514.37.6.913

Bagozzi R. P., Tybout A. M., Craig C. S. and Sternthal B., (1979), The construct validity of the tripartite classification of attitudes, *J. Marketing Res.,* **16**, 88-95. http://www.jstor.com/stable/3150879

Bain R., (1928), An attitude on attitude research. *Am. J. Sociol.,* **33**, 940-957. http://www.jstor.com/stable/2765988

Ballen C. J., Wieman C., Salehi S., Searle J. B. and Zamudio K. R., (2017), Enhancing diversity in undergraduate science: Self-efficacy drives performance gains with active learning, *CBE-Life Sci. Educ.,* **16**(4), 1-6. DOI:10.1187/cbe.16-12-0344

Barr D., Matsui J., Wanat S. F. and Gonzalez, M., (2010), Chemistry Courses as the Turning Point for Premedical Students, *Adv. Health Sci. Educ*., **15**, 45−54. DOI:10.1007/s10459-009-9165-3

Bauer C. F., (2008), Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts, *J. Chem. Educ.*, **85**, 1440−1445. DOI:10.1021/ed085p1440

Bern D. J., (1970), *Beliefs, attitudes, and human affairs,* Belmont, CA: Brooks/Cole.

Black A. E. and Deci E. L., (2000), The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective, *Sci. Educ*., **84**, 740-756. DOI:10.1002/1098-237X(200011)84:6<740::AID-SCE4>3.0.CO;2-3

Bodner G. M., (1986), Constructivism: A theory of knowledge, *J. Chem. Educ.,* **63**(10), 873. DOI:10.1021/ed063p873

Bodner G. M., (2015), Research on problem solving in chemistry. In Chemistry education, García-Martínez J. and Serrano-Torregrosa E., (Eds.), Wiley-VCH Verlag GmbH&Co. KGaA: Weinheim, Germany, pp 181−202.

Bodner G. M., Herron J. D., (2002), *Problem-solving in chemistry*. In Chemical education: Towards research-based practice, Gilbert J. K., de Jong O., Justi R., Treagust D. F. and van Driel J. H. (Eds.), Science & Technology Education Library, Springer: Dordrecht, pp 235−266.

Bourdieu P. and Passeron J., (1977), *Reproduction in education, society and culture,* SAGE, London.

Brandriet A. R., Ward R. M. and Bretz S. L., (2013), Modeling meaningful learning in chemistry using structural equation modeling, *Chem. Educ. Res. Pract.*, **14**, 421-430. DOI:10.1039/C3RP00043E

Brandriet A. R., Xu X., Bretz S. L. and Lewis J. E., (2011), Diagnosing changes in attitude in first-year college chemistry students with a shortened version of Bauer's semantic differential, *Chem. Educ. Res. Pract*., **12**, 271-278. DOI:10.1039/C1RP90032C

Breckler, S. I., (1984), Empirical validation of affect, behavior, and cognition as distinct components of attitude, *J. Pers. Soc. Psychol.,* **47***,* 1191-1205. DOI:10.1037/0022-3514.47.6.1191

Bretz S. L., (2001), Novak's theory of education: Human constructivism and meaningful learning, *J. Chem. Educ*., **78**, 1107−1115. DOI:10.1021/ed078p1107.6

Brown S. J., Sharman, B. N., Wakeling L., Naiker M., Chandra S., Gopalan R.D. and Bilimoria, V. B., (2014), Quantifying attitude to chemistry in students at the University of the South Pacific, *Chem. Educ. Res. Pract*., **15**, 184-191. DOI:10.1039/C3RP00155E

Campbell A., (1971), *White attitudes toward black people.* Ann Arbor: Institute for Social Research, University of Michigan.

Cantú N., (2012), Getting there *Cuando no hay camino* (when there is no path): Paths to discovery *Testimonios* by Chicanas in STEM, *Equity & Excellence Educ.,* **45**(3), 472-487. DOI:10.1080/10665684.2012.698936

Carlone H. B. and Johnson A., (2007), Understanding the science experiences of successful Women of Color: Science identity as an analytic lens, *J. Res. Sci. Teach*., **44**(8), 1187−1218. DOI:10.1002/tea.20237

Catsambis S., (1995), Gender, Race, Ethnicity, and Science Education in the Middle Grades, *J. Res. Sci. Teach*., **32**(2), 243−257. DOI:10.1002/tea.3660320305

Chan J. Y. K. and Bauer C. F., (2014), Identifying at-risk students in general chemistry via cluster analysis of affective characteristics, *J. Chem. Educ*., **91**, 1417-1425. DOI:10.1021/ed500170x

Chan J. Y. K. and Bauer C. F., (2015), Effect of peer-led team learning (PLTL) on student achievement, attitude, and self-concept in college general chemistry in randomized and quasi experimental designs, *J. Res. Sci. Teach*., **52**(3), 319-346. DOI:10.1002/tea.21197

Chan J. K. and Bauer C. F., (2016), Learning and studying strategies used by general chemistry students with different affective characteristics, *Chem. Educ. Res. Pract.*, **17**, 675-684. DOI:10.1039/C5RP00205B

Cheung D., (2011), Evaluating students attitudes toward chemistry lessons to enhance teaching in the secondary school, *Educ. Quim.,* **22**(2), 117-122. ISSN:0187-893-X

Chi M. T. H., (2005), Commonsense conceptions of emergent processes: Why some misconceptions are robust, *J. Learn. Sci.*, **14**, 161−199. DOI:10.1207/s15327809jls1402_1

Clair-Thompson H. S., Overton T. and Bugler M., (2012), Mental capacity and working memory in chemistry: Algorithmic versus open-ended problem solving, *Chem. Educ. Res. Pract.*, **13**, 484−489. DOI:10.1039/C2RP20084H

Cook S. W. and Selltiz C., (1964), A multiple-indicator approach to attitude measurement, *Psychol. Bulletin,* **62**(1), 36–55. DOI:10.1037/h0040289

Cooper M. M., Grove N. and Underwood S. M., (2010), Lost in Lewis structures: An investigation of student difficulties in developing representational competence, *J. Chem. Educ.,* **87**(8), 869-874. DOI:10.1021/ed900004y

Cooper M. M., Williams L. C., Underwood S. M., (2015), Student understanding of intermolecular forces: A multimodal study, *J. Chem. Educ.*, **92**, 1288−1298. DOI:10.1021/acs.jchemed.5b00169

Cracolice M. S. and Busby B. D., (2015), Preparation for College General Chemistry: More than just a matter of content knowledge acquisition, *J. Chem. Educ.,* **92**, 1790-1797. DOI:10.1021/acs.jchemed.5b00146

Crandell O. M., Kouyoumdjian H., Underwood S. M. and Cooper M. M., (2019), Reasoning about reactions in organic chemistry: Starting it in general chemistry, *J. Chem. Educ.,* **96**(2), 213-226. DOI:10.1021/acs.jchemed.8b00784

Crenshaw K., (1989), Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, *University of Chicago Legal Forum*, 139−168. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Curtis B. and Curtis C., (2017), In-Depth Interviewing – The Interactive Base. In *Social Research: A Practical Introduction,* SAGE Publications, Inc.

Damo K. L. and Prudente M. S., (2019), Investigating students' attitude and achievement in organic chemistry using interactive application, *Assoc. Computing Machinery,* 36-41, Tokyo, Japan. DOI:10.1145/3306500.3306562

diSessa A. A., (1988), *Knowledge in Pieces*. In Constructivism in the computer age, Forman G. E. and Pufall P. B. (Eds.), Lawrence Erlbaum Associates: Hillsdale, NJ, pp 49−70.

Donaldson L. P. and Daugherty L., (2011), Introducing asset-based models of social justice into service learning: A social work approach, *J. Community Pract.* **19**(1), 80-99. DOI:10.1080/10705422.2011.550262

Dood A. J., Fields K. B., Cruz-Ramírez de Arellano D. and Raker J. R., (2019), Development and evaluation of a Lewis acid-base tutorial for use in post-secondary organic chemistry courses, *Can. J. Chem.,* **97***,* 711-721. DOI:10.1139/cjc-2018-0479

Eagly A. H. and Chaiken S., (1993), *The psychology of attitudes.* Fort Worth, TX: Harcourt Brace Jovanovich.

Eagly A. H. and Chaiken S., (2007), The advantages of an inclusive definition of attitude, *Soc. Cognition,* **25**(5), 582-602. DOI:10.1521/soco.2007.25.5.582

Eagly A. H., Mladinic A. and Otto S., (1994), Cognitive and affective bases of attitudes toward social groups and social policies, *J. Exp. Soc. Psychol.,* **30**, 113-137. DOI:10.1006/jesp.1994.1006

Else-Quest N. M., Mineo C. C. and Higgins A., (2013), Math and science attitudes and achievement at the intersection of gender and ethnicity, *Psychol. Women Quart.,* **37**(3), 293-309. DOI: 10.1177/0361684313480694

Entwistle N. J., (1991), Approaches to learning and perceptions of the learning environment, *High. Educ*., **22**, 201−204. https://www.jstor.org/stable/3447172

Fautch J. M., (2015), The flipped classroom for teaching organic chemistry in small classes: is it effective? *Chem. Educ. Res. Pract.,* **16**, 179-186. DOI: 10.1039/C4RP00230J

Fazio R. H., (1986), *How do attitudes guide behavior?* In The handbook of motivation and cognition: Foundations of social behavior, Sorrentino R. M. and Higgins E. T. (Eds.), New York: Guilford.

Fink A., Cahill M. J., McDaniel M. A., Hoffman A. and Frey R. F., (2018), Improving general chemistry performance through a growth mindset intervention: Selective effects on underrepresented minorities, *Chem. Educ. Res. Pract.,* **19,** 783-806. DOI:10.1039/C7RP00244K

Fink A., Frey R. F. and Solomon E. D., (2020), Belonging in general chemistry predicts first-year undergraduate's performance and attrition, *Chem. Educ. Res. Pract.,* (Published) DOI:10.1039/D0RP00053A

Fishbein M. and Ajzen I., (1975), *Belief, attitude, intention, and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Flynn A. B., (2015), Structure and evaluation of flipped chemistry courses: Organic & spectroscopy, large and small, first to third years, English and French, *Chem. Educ. Res. Pract,* **16**, 198-211. DOI:10.1039/C4RP00224E

Fraser B. J., (1977), Selection and validation of attitude scales for curriculum evaluation, *Sci. Educ*., **61**(3), 317–329. DOI:10.1002/sce.3730610307.

García N. M., López N. and Vélez V. N., (2018), QuantCrit: Rectifying quantitative methods through critical race theory, *Race Ethn. Educ.,* **21**(2), 149-157. DOI:10.1080/13613324.2017.1377675

Gibbons R. E., Xu X., Villafañe S. A. and Raker J. R., (2018), Testing a reciprocal causation model between anxiety, enjoyment and academic performance in postsecondary organic chemistry, *Educ. Psychol.,* **38**(6), 838-856. DOI:10.1080/01443410.2018.1447649

Gillborn D., Warmington P. and Demack S., (2018), QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics, *Race Ethn. Educ.,* **21**(2), 158-179. DOI:10.1080/13613324.2017.1377417

Graulich N., (2015), The tip of the iceberg in organic chemistry classes: How do students deal with the invisible? *Chem. Educ. Res. Pract*., **16**, 9−21. DOI:10.1039/C4RP00165F

Grove N. P. and Bretz S. L., (2010), Perry's scheme of intellectual and epistemological development as a framework for describing student difficulties in learning organic chemistry, *Chem. Educ. Res. Pract.,* **11**, 207–211. DOI:10.1039/C005469K

Guttman L., (1944), A basis for scaling qualitative data, *Am. Sociol. Rev.,* **9**, 139-150. http://www.jstor.com/stable/2086306

Halpern D. F., Benbow C. P., Geary D. C., Gur R., Hyde, J. S. and Gernsbacher M. A. (2007), The science of sex differences in science and mathematics, *Psychol. Sci. Pub. Int.,* **8**, 1–51. DOI:10.1111/j.1529-1006.2007.00032.x

Harris R. B., Mack M. R., Bryant J., Theobald E. J. and Freeman S., (2020), Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a "hyperpersistent zone", *Science Advances,* **6**(24), (Published) DOI:10.1126/sciadv.aaz5687

Hazari Z., Sadler P. M. and Sonnert, G., (2013), The science identity of college students: Exploring the intersection of gender, race, and ethnicity. *J. Coll. Sci. Teach*., **42**, 82−91. https://www.jstor.org/stable/43631586

Hensen C. and Barbera J., (2019), Assessing affective differences between a virtual general chemistry experiment and a similar hands-on experiment, *J. Chem. Educ*., **96**, 2097-2108. DOI:10.1021/acs.jchemed.9b00561

Heredia K. and Lewis J. E., (2012), A psychometric evaluation of the Colorado learning attitudes about science survey for the use in chemistry, J. Chem. Educ., 89(4), 436-441. DOI:10.1021/ed100590t

Horowitz G., Rabin L. A. and Brodale D. L., (2013). Improving student performance in organic chemistry: Help seeking behaviors and prior chemistry aptitude, *J. Scholarship Teach. Learn.,* **13**(3), 120-133. https://scholarworks.iu.edu/journals/index.php/josotl/article/view/3152

Hosbein K. N. and Barbera J., (2020), Alignment of the theoretically grounded constructs for the measurement of science and chemistry identity, *Chem. Educ. Res. Pract.,* **21,** 371-386. DOI:10.1039/C9RP00193J

Huskinson T. and Haddock G., (2004), Individual differences in attitude structure: Variance in the chronic reliance on affective and cognitive information, *J. Exp. Soc. Psychol.,* **40***,* 82-90. DOI:10.1016/S0022-1031(03)00060-X

Ireland D. T., Freeman K. E., Winston-Proctor C. E., DeLaine K. D., McDonald Lowe S. and Woodson K. M., (2018), (Un)hidden Figures: A Synthesis of Research Examining the Intersectional Experiences of Black Women and Girls in STEM, *Rev. Res. Educ*., **42**, 226−254. DOI: 10.3102/0091732X18759072

Johnstone A. H., (1993), The development of chemistry teaching: A changing response to changing demand, *J. Chem. Educ*., **70**, 701. DOI:10.1021/ed070p701

Kahveci A., (2015), Assessing High School Students' Attitudes Toward Chemistry with a Shortened Semantic Differential, *Chem. Educ. Res. Pract*., **16**, 283−292. DOI: 10.1039/C4RP00186A

Kanadli S., (2016), A meta-analysis on the effect of instructional designs based on the learning styles models on academic achievement, attitude, and retention, *Educ. Sci. Theory Pract.,* **16**(6), 2057-2086. DOI:10.12738/estp.2016.6.0084

Kempf D. S., (1999), Attitude formation from product trial: Distinct roles of cognition and affect for hedonic and functional products, *Psychol. Mark.,* **16**, 35-50. DOI:10.1002/(SICI)1520-6793(199901)16:1<35::AID-MAR3>3.0.CO;2-U

Koballa T. R. and Crawley F. E., (1985), The influence of attitude on science teaching and learning. *School Sci.Math*., **85**, 222–232. DOI:10.1111/j.1949-8594.1985.tb09615.x.

Koballa T. R., (1988), Attitude and related concepts in science education, *Sci. Educ*., **72**, 115–126. DOI:10.1002/sce.3730720202

Kraft A., Strickland A. M. and Bhattacharyya G., (2010), Reasonable reasoning: multi-variate problem-solving in organic chemistry, *Chem. Educ. Res. Pract.,* **11**, 281-292. DOI:10.1039/C0RP90003F

Krech D., Crutchfield R. and Ballachey E., (1962), *Individual in society,* New York: McGraw-Hill.

Kretzman J. and McKnight J., (1993), *Building communities from the inside out: A path toward finding and mobilizing a community's assets.* ABCD Institute, Evanston, IL.

Lewis K. L., Stout J. G., Pollock S. J., Finkelstein N. D. and Ito T. A., (2016), Fitting in or option out: A review of key social-psychological factors influencing a sense of belonging for women in physics, *Phys. Rev. Phys. Educ. Res.,* **12**(2), 1-10. DOI:10.1103/PhysRevPhysEducRes.12.020110

Likert R., (1932), A technique for the measurement of attitudes. *Archives Psychol*., **140**, 1–55.

Litzler E., Samuelson C. C. and Lorah J. A., (2014), Breaking it Down: Engineering Student STEM Confidence at the Intersection of Race/ Ethnicity and Gender, *Res. High. Educ*., **55**, 810−832. DOI:10.1007/s11162-014-9333-z

Liu Y., Raker J. R. and 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, *Chem. Educ. Res. Pract.,* **19,** 251-264. DOI:10.1039/C7RP00153C

May N. W., McNamara S. M., Wang S., Kolesar K. R., Vernon J., Wolfe J. P., Goldberg D. and Pratt K. A., (2018), Polar plunge: Semester-long snow chemistry research in the general chemistry laboratory, *J. Chem. Educ*., **95**, 543-552. DOI:10.1021/acs.jchemed.7b00823

McGuire W. J., (1969), *The nature of attitudes and attitude change*, In The handbook of social psychology (2nd ed., Vol. 3), Lindzey G. and Aronson E. (Eds.), Reading, MA: Addison-Wesley, pp. 136-314.

Montes L. H., Ferreira R. A. and Rodriguez C., (2018), Explaining Secondary School Students' Attitudes Towards Chemistry in Chile, *Chem. Educ. Res. Pract*., **19**(2), 533−542. DOI:10.1039/C8RP00003D

Mooring S. R., Mitchell C. E. and Burrows, N. L., (2016), Evaluation of a flipped, large enrollment organic chemistry course on student attitude and achievement, *J. Chem. Educ.*, **93**, 1972−1883. DOI:10.1021/acs.jchemed.6b00367

Myende P. E., (2015), Tapping in the asset-based approach to improve academic performance in rural schools, *J. Hum. Ecol,* **50**(1), 31-42. DOI:10.1080/09709274.2015.11906857

Navarro M., Förster C., González C. and González-Pose P., (2016), Attitudes toward science: measurement and psychometric properties of the Test of Science-Related Attitudes for its use in Spanish-speaking classrooms*, Int.l J. Sci. Educ*., **38**(9), 1459-1482. DOI:10.1080/09500693.2016.1195521

Nenning H. T., Idarraga K. L., Salzer L. D., Blaske-Rechek A. and Theisen R. M., (2019), Comparison of student attitudes and performance in an online and face-to-face inorganic chemistry course, *Chem. Educ. Res. Pract*., **21,** 168-177. DOI:10.1039/C9RP00112C

Obama B., (2010), Science, technology, engineering and math: Education for global leadership, *Retrieved on* July 8, 2020. https://www.ed.gov/sites/default/files/stem-overview.pdf

Ong M., Wright C., Espinosa L. L. and Orfield G., (2011), Inside the Double Bind: A Synthesis of Empirical Research on Undergraduate and Graduate Women of Color in Science, Technology, Engineering, and Mathematics, *Harvard Educ. Rev.*, **81**(2), 172− 208. DOI:10.17763/haer.81.2.t022245n7x4752v2

Ormerod M. D. and Duckworth D. (1975), *Pupils' attitudes to science: A review of research*, Windsor, Berks: Humanities Press.

Osborne J., Simon S., and Collins S., (2003), Attitudes towards science: A review of the literature and its implications, *Int. J. Sci. Educ*., **25**(9), 1049-1079, DOI:10.1080/0950069032000032199

Osgood C. E., (1965), *Cross cultural comparability of attitude measurement via multi-lingual semantic differentials*. In Recent studies in social psychology, Steiner I. S. and Fishbein M. (Eds.), New York: Holt, Rinehart & Winston, pp. 95-107.

Osgood C. E., Suci G. J. and Tannenbaum P. H., (1957), *The measurement of meaning,* Urbana: University of Illinois Press.

Oskamp S. and Schultz P. W., (2005), *Attitudes and opinions,* (3rd ed), Mahwah, New Jersey: Lawrence Erlbaum Associates.

Pekrun R., Maier M. A., Elliot A. J., (2009), Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance, *J. Educ. Psychol*., **101**(1), 115−135. DOI:10.1037/a0013383

Pekrun R.. (2006), The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice, *Educ. Psychol. Rev.,* **18,** 315–341. DOI:10.1007/s10648-006-9029-9

Peralta C., Caspary M. and Boothe D., (2013), Success factors impacting Latina/o persistence in higher education leading to STEM opportunities, *Cult. Stud. Sci. Educ.,* **8**, 905-918. DOI:10.1007/s11422-013-9520-9

Posner G. J., Strike K. A., Hewson P. W., Gertzog W. A., (1982), Accommodation of a scientific conception: Toward a theory of conceptual change, *Sci. Educ*., **66**, 211−227.

Ramsden J. M., (1998), Mission impossible?: Can anything be done about attitudes to science? *Int. J. Sci. Educ*., **20**, 125–137. DOI:10.1080/0950069980200201

Reid N., (2006), Thoughts on attitude measurement, *Res. Sci.Technol. Educ.*, **24**(1), 3-27, DOI:10.1080/02635140500485332

Rocabado G. A., Kilpatrick N. A., Mooring S. R., and Lewis J. E., (2019), Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course, *J. Chem. Educ.,* **96**(11), 2371-2382. DOI:10.1021/acs.jchemed.9b00516

Rocabado G. A., Komperda R., Lewis J. E. and Barbera J., (2020), Addressing diversity and social inclusion through groups comparisons: A primer on measurement invariance testing, *Chem. Educ. Res. Pract.,* **21,** 969-988. DOI:10.1039/D0RP00025F

Rodríguez S., Cunningham K. and Jordan A., (2019), STEM identity development for Latinas: The role of self- and outside recognition, *J. Hispan. High. Educ.,* **18**(3), 254-272. DOI:10.1177/1538192717739958

Rosenberg J. and Hovland I., (1960), *Cognitive, affective, and behavioral components of attitudes,* In Attitude rganization and change: An analysis of consistency among attitude components, Rosenberg M. J. et al., (Eds), New Haven, CT: Yale University Press, pp. 1-14.

Ross J., Nuñez L. and Lai C. C., (2018), Partial least squares structural equation modeling of chemistry attitude in introductory college chemistry, *Chem. Educ. Res. Pract*., **19**, 1270-1286. DOI:10.1039/C7RP00238F

Rowe M. B., (1983), Getting chemistry off the killer course list, *J. Chem. Educ.*, **60**, 954−956. DOI:10.1021/ed060p954

Salta K. and Tzougraki C., (2004), Attitudes toward chemistry among 11th grade students in high school in Greece, *Sci. Ed.,* **88,** 535-547. DOI 10.1002/sce.10134

Seadler A., (2012), Obama introduces plan to increase U. S. STEM undergraduates, *Earth*, **57**(6), 27. Online ISSN:1865-0678

Sen S., Yilmaz A. and Temel S., (2016), Adaptation of the Attitude toward the Subject of Chemistry Inventory (ASCI) into Turkish. *J. Educ. Training Stud,* **4**(8), 27-33. ISSN-2324-805X

Seymour E. and Hewitt N., (1997), *Talking About Leaving: Why Undergraduates Leave the Sciences*, Boulder, CO: Westview Press.

Seymour E. and Hunter A-B., (2019), *Talking About Leaving Revisited: Persistence, relocation, and loss in undergraduate STEM education*, Springer Nature Switzerland.

Skinner B. F., (1957), *Verbal behavior,* New York: Appleton-Century-Crofts.

Spagnoli D., Wong L., Maisey S. and Clemons T. D., (2017), Prepare, do, review: A model used to reduce the negative feelings towards laboratory classes in an introductory chemistry undergraduate unit, *Chem. Educ. Res. Pract.,* **18**, 26-44. DOI:10.1039/C6RP00157B

Stanich C. A., Pelch M. A., Theobald E. J. and Freeman S., (2018), A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women, *Chem. Educ. Res. Pract*., **19**, 846-866. DOI:10.1039/C8RP00044A

Sullivan A., (2001), Cultural capital and educational attainment, *Sociology,* **35**(4), 893-912. DOI:10.1017/S0038038501008938

Taber K. S., (2015), *Meeting educational objective in the affective and cognitive domains: Personal and social constructivist perspectives on enjoyment, motivation and learning chemistry*. In Affective Dimensions in Chemistry Education*,* Kahveci M., Orgill M. (Eds), Springer Heidelberg, pp. 3-27.

Thomsen C. and Finley J., (2019), On intersectionality: A review essay. *Hypatia,* **34**(1), 155-160. DOI:10.1111/hypa.12450

Thurstone L. L., (1928), Attitudes can be measured, *Am. J. Sociol.,* **33***,* 529-554.

Underwood S. M., Reyes-Gastelum D. and Cooper M. M., (2016), When do students recognize relationships between molecular structure and properties? A longitudinal comparison of the impact of traditional and transformed curricula, *Chem. Educ. Res. Pract.*, **17**, 365-380. DOI:10.1039/C5RP00217F

Vilia P. N., Candeias A. A., Neto A. S., Franco M. D. G. S., and Melo M., (2017), Academic achievement in physics-chemistry: The predictive effect of attitudes and reasoning abilities, *Frontiers in Psychology,* **8,** 1-9. DOI:10.3389/fpsyg.2017.01064

Villafañe S. M. and Lewis J. E., (2016), Exploring a measure of science attitude for different groups of students enrolled in introductory college chemistry, *Chem. Educ. Res. Pract*., **17,** 731−742. DOI:10.1039/C5RP00185D

Villafañe S. M., Garcia C. A. and Lewis J. E., (2014), Exploring diverse students' trends in chemistry self-efficacy throughout a semester of college-level preparatory chemistry, *Chem. Educ. Res. Pract.,* **15***,* 144-127. DOI:10.1039/C3RP00141E

Vishnumolakala V. R., Qureshi S. S., Treagust D. F., Mocerin, M., Southam D. S. and Ojeil J., (2018), Longitudinal impact of process-oriented guided inquiry learning on the attitudes, self-efficacy and experiences of pre-medical chemistry students. *QScience Connect,* **1**, 1-12. DOI:10.5339/connect.2018.1

Vishnumolakala V. R., Southam D. C., Treagust D. F., Mocerino M. and Qureshi S. (2017), Students' attitudes, self-efficacy and experiences in a modified process-oriented guided

inquiry learning undergraduate chemistry classroom, *Chem. Educ. Res. Pract*., **18**, 340-352. DOI:10.1039/C6RP00233A

Wachsmuth L. P., Runyon C. R., Drake J. M. and Dolan E. L., (2017), Do biology students really hate math? Empirical insights into undergraduate life science majors' emotions about mathematics, *CBE-Life Sci.,* **49**, 1-10. DOI:10.1187/cbe.16-08-0248

Wang Y., Rocabado G. A., Lewis J. E. and Lewis S. E., (2020), Utility-value and growth mindset interventions in STEM at college level, *J. Chem. Educ.,* (Submitted).

Wang, Y. and Lewis, S. E., (2020), Analytical chemistry students' explanatory statements in the context of their corresponding lecture, *Chem. Educ. Res. Pract.,* (Published) DOI:10.1039/D0RP00063A

Xu X. and Lewis J., (2011), Refinement of a Chemistry Attitude Measure for College Students, *J. Chem. Educ.,* **88**, 561-568. DOI:10.1021/ed900071q

Xu X., (2010), *Refinement of a Chemistry Attitude Measure for College Students*, Dissertation, University of South Florida, Tampa, FL.

Xu X., Alhoosani K., Southam D. and Lewis J. E., (2015), *Gathering Psychometric Evidence for ASCIv2 to Support Cross-Cultural Attitudinal Studies for College Chemistry Programs*. In Affective Dimensions in Chemistry, Springer-Verlag: Berlin, pp. 177−194.

Xu X., Southam D. and Lewis J. E., (2012), Attitude Towards the Subject of Chemistry in Australia: An ALIUS and POGIL Collaboration to Promote Cross-National Comparisons, *Aus. J. Educ. Chem.*, **72**, 32−36. ISSN-14459698

Xu X., Villafañe, S. M. and Lewis J. E., (2013), College students' attitudes toward chemistry, conceptual knowledge and achievement: Structural equation model analysis. *Chem. Educ. Res. Pract*., **14**, 188−200. DOI:10.1039/C3RP20170H

Yosso T., (2005), Whose culture has capital? A critical race theory discussion of community cultural wealth, *Race Ethn. Educ.,* **8**(1), 69-91. DOI:10.1080/1361332052000341006

CHAPTER 2

INTRUMENTS AND METHODS

In this chapter, a preview of the most common quantitative methods used throughout this dissertation will be given along with a justification for the choice of method in each case. Chapters 3, 5, and 6 each have their own detailed methods sections; therefore, no such detail will be given here, rather I provide a more thorough clarification of assumptions, data cleaning procedures, etc. Additionally, chapter 4 is a primer on measurement invariance testing, which is a method I have used throughout this work consistently and will not be described here since chapter 4 is entirely dedicated to this method.

A positivist view governed the studies presented in this dissertation. Positivism entails a belief that the observer examines phenomena of interest and does not influence its outcomes. This view embraces the notion that truth is independent from the observer, it is self-governing, and objective (Aliyu *et al.*, 2014). Although I have operated under this idea throughout my research studies, I also believe that research and researchers are not objective, and 'numbers are not neutral' (García, López and Vélez, 2018; Gillborn, Warmington and Demack, 2018). Therefore, I have carefully employed methods that allowed ample opportunity to scrutinize the data, the results, and the inferences made throughout these studies, particularly when it came to drawing conclusions for marginalized and disadvantaged students that could continue to disenfranchise them and favor

the systemic issues that oppress them. To this end I have chosen to use methods that provide evidence of support for subgroup comparisons with ample opportunities for scrutiny against implicit biases.

Instruments and Participants

In this work I have reported several studies that have relied on the use of instruments to examine students' attitudes toward chemistry. The Attitude toward the Subject of Chemistry Inventory version 2 (ASCIv2; Xu and Lewis, 2011) and variations of this instrument were used in each of the research studies. This instrument first developed by Bauer (2008) and refined by Xu and Lewis (2011) utilizes a semantic differential scale (Osgood, 1965) that provides each student with items containing two opposing adjectives to describe their feelings toward the discipline of chemistry. Each time, the instrument was administered via online platforms such as Qualtrics or Canvas to students enrolled in general chemistry or organic chemistry courses. The studies in this dissertation focused on the attitudes of students enrolled in organic chemistry courses; however, data obtained from general chemistry students was used on occasion for pilot studies.

In chapter 3, the ASCIv3 (Xu, 2010) was completed by 395 Organic Chemistry I (OCI) students at a southeastern public research institution in the fall of 2015 at the beginning and at the end of the semester. Of those 395 students, 270 were Black female students, and 125 were all other students in the course. The ASCIv3 is an eight-item, two-factor instrument similar to the ASCIv2 with a single variation in item order (Xu, 2010). In the ASCIv3 items 2 (Complicated-Simple) and

8 (Chaotic-Organized) switch places. The two factors measured by the ASCIv2 and ASCIv3 are *Intellectual Accessibility* (IA) and *Emotional Satisfaction* (ES). The data from this study was originally collected and analyzed in fall 2015 for a study conducted by Mooring and colleagues (2016). This study evaluated the impact of a flipped organic chemistry classroom and measured students' attitudes as a gauge of success in the implementation of a flipped classroom pedagogy. Mooring et al. (2016) observed that students in the flipped classroom experienced significant gains in attitude as well as achievement (measured by exams). I used these data to inspect whether these attitude and achievement gains extended to the Black female students in this course (Rocabado *et al*., 2019).

Chapter 5 describes the attitudes of 171 White female and 84 Hispanic female students in a traditional lecture OCI classroom at a southeastern public research university in the Fall 2018. These students were part of a cohort of 650 students. The students completed the ASCIv2 several times during the semester two days before each of the course exams including the final exam.

Chapter 6 describes the development of another variation of the ASCIv2, the ASCI-UE, a nine-item, two-factor instrument measuring *Utility* and *Emotional Satisfaction.* This instrument was developed in English and Spanish simultaneously. I conducted cognitive interviews with eleven students enrolled in general and organic chemistry courses. Additionally, chapter 6 describes the administration of this instrument in an Organic Chemistry II (OCII) course in a southeastern public research institution in Fall 2019. A total of 291 students completed the survey at the beginning of the semester. Several times during the semester students were asked to complete

the survey two days before each of their exams including the final exam. A comparison of attitude between high- and low-achieving students was also conducted in this study.

Methods and Analyses

Data Cleaning Process

When beginning a study, data were gathered in a Microsoft Excel file. Throughout the semester, each time students completed the survey and took an exam the data were matched by the students' university identification number in a single master file. At the end of the semester demographic data for each student was obtained from the university records following IRB approved protocols and was matched to each student in the master file. Once all data were gathered, I processed all the missing data by including a number that would not be possible to obtain from any of the categories such as -999. Analysis of missing data was conducted on SPSS leading to the conclusion that all missing data was missing at random. I also scrutinized some patterns in the data that could be problematic, such as students choosing only the extremes or the middle options. No cases were excluded from this evaluation of the data. Finally, all categories were given numerical values, such as 1 for female students, and 2 for male students. The categories for gender were male or female in each instance. The categories for race/ethnicity for chapter 3 were White/Caucasian, Black/African American, Asian, Native American, Pacific Islander, and Other. The categories for race/ethnicity for chapters 5 and 6 were White, Black/African American, Hispanic, Asian, Native American, Pacific Islander, Foreign, and Unknown. Other categories that could be found in the

data were course modality (*i.e.*, flipped classroom, traditional), majors (*i.e.*, STEM, health) and other relevant demographics. Once all data were converted to numerical values, I deidentified the data by assigning a number to each student and subsequently removing all identifiable information such as names, student identification numbers, etc.

Descriptive Statistics

Once the data were deidentified descriptive statistics were computed first for the entire sample, and then if deemed appropriate for subgroups within the sample. Descriptive statistics were computed in SPSS. This type of statistics lets the researcher obtain a broad depiction of the data by computing simple features such as mean and standard deviation. These simple features describe the basic tendency and variability of the data, respectively. Descriptive statistics were used throughout all of the studies to provide basic knowledge of the data and later to complete longitudinal or group comparisons. Most descriptive statistics are provided in each of the subsequent chapter (3-6) and also in Appendix C.

In addition to the mean and standard deviation values, I also explored measures of normality of the distribution for each variable. This assumption is an important feature of the data because further analyses, both univariate and multivariate, assume normal distribution curves for continuous data (or ordinal data that is treated as continuous). Measures of normality include skewness and kurtosis values. Throughout all of the studies in this dissertation, values outside of the +/- 1.00 range, were deemed non-normal (Bulmer, 1979). Most items displayed skewness and

kurtosis values that were considered within a normal range, however, in each study there were some cases that skewness and kurtosis were outside of the normal values. For further analyses of the data, I utilized a robust estimator (maximum likelihood robust) which took into account the non-normally distributed data (Cheng-Hsien, 2016). Additionally, Levene's test of homogeneity was tested, and each item displayed a non-significant (>0.05) statistic, meaning that the assumption of homogeneity was achieved.

Measurement Models

Throughout this work there was a focus in longitudinal comparisons and subgroup comparisons of attitude throughout a course. These comparisons are meaningful when evidence that the internal structure of the instrument holds longitudinally and for the subgroups. Therefore gathering evidence of internal structure validity and reliability is paramount for conducting comparisons (Arjoon *et al*., 2013; AERA *et al*., 2014).

One way to demonstrate that the internal structure of the instrument holds for the data collected is by first conducting statistical analyses such as confirmatory factor analysis (CFA; Brown, 2006). Throughout this work, conducting CFA was a standard procedure to ensure meaningful interpretation of observed factor scores. CFA was conducted using Mplus software (Muthén and Muthén, 1998-2007) with a maximum likelihood robust (MLR) estimator to handle non-normal data (Cheng-Hsien, 2016). Additionally, Mplus handles missing data by using full-information maximum likelihood (FIML) as opposed to pairwise or listwise deletion, which are

more common in software packages such as SPSS. When conducting CFA, all analyses terminated normally and convergence was achieved unless otherwise listed in each specific study. If convergence was not achieved, or the internal structure of the instrument did not hold, any result would be deemed unfit for interpretation.

Certain standards were employed when determining whether there was a good data-model fit. Hu and Bentler (1999) provide proposed cutoffs or guidelines for model fit indices common when conducting CFA with continuous data. It is in the best interest of the researcher to review several kinds of model fit indices that provide insight into different aspects of model fit. There are three common categories of fit indices, namely, absolute fit, comparative fit, and parsimony correction (Brown, 2006). The absolute fit indices, such as the Chi-square (χ^2) test statistic and the standardized root-mean square residual (SRMR), investigate how closely the data fit the model when compared to a null hypothesis that the data-model fit is perfect. The χ^2 statistic should be close to zero with no evidence of significant difference; however, this statistic is highly influenced by sample size (Brown, 2006). The SRMR has a suggested cutoff of <0.08 as acceptable based on simulation studies by Hu and Bentler (1999). Comparative fit indices examine the data-model fit in comparison to a baseline model where there are no relations between items through an underlying factor (Brown, 2006). Two common examples of comparative fit are: the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI), which have a suggested cutoff of >0.90 as acceptable, but best if >0.95 (Hu and Bentler 1999). The parsimony correction indices are comparable to the absolute fit indices with the addition of a penalty for poor model parsimony (Brown, 2006). The root mean square of approximation (RMSEA) index is an example of parsimony correction index and is evaluated with acceptable cutoff criteria of <0.06 (Hu and

Bentler 1999). However, this fit index behaves idiosyncratically when instruments are short leading to small degrees of freedom (Kenny, Kaniskan and McCoach, 2015). Therefore, in many instances throughout this work, the RMSEA was not used for model fit evaluation since the instruments used only have eight or nine items leading to small degrees of freedom.

Longitudinal and subgroup comparisons were only considered appropriate if the internal structure held for the subgroups or over time, simultaneously. Chapter 4 describes in detail measurement invariance testing, the method used to obtain this evidence throughout this work. The model fit criteria describe previously is used when evaluating measurement invariance testing models, along with change in fit criteria described by Chen (2007) and detailed in Chapter 4. Gathering this evidence before conducting comparisons provided safeguards against inferences that would be inappropriate to make due to respondents' inconsistent interpretation of the items or constructs being measured. Therefore, this method was fundamental in the work presented in this dissertation. I reported the results of CFA and measurement invariance testing within each of the subsequent chapters (3-5) along with more detailed information in Appendix C.

In addition to gathering internal structure validity by conducting CFA and measurement invariance testing, gathering reliability evidence is also common and a highly suggested practice. Normally researchers have used Cronbach's alpha as a measure of reliability in studies of this nature (Cronbach, 1957; Cortina, 1993). However, this statistic assumes a *tau*-equivalent model in which all parameters of the model are freely estimated except for the factor loadings which are constrained to be the same for all items. This scenario is rare, therefore Cronbach's alpha is generally not an appropriate coefficient of reliability. Komperda, Pentecost, and Barbera (2018)

describe several alternatives to Cronbach's alpha that are more appropriate for models that contain all freely estimated parameters. One of their suggestions was the Mcdonald's Omega coefficient which takes into account an appropriate data-model fit and uses the factor loadings and error variance parameters to calculate reliability of each factor in a model. Omega was used as a measure of reliability throughout this work, with values >0.70 and closer to 1.00 as good measures of reliability.

Longitudinal and Subgroup Comparisons

After conducting measurement invariance testing and only if the results of this analysis were optimal, longitudinal or subgroup comparisons could take place. Some of the statistical analyses used for comparisons were *t*-tests and multivariate analysis of variance (MANOVA). Paired-samples *t*-tests were conducted for longitudinal comparisons because the sample was the same and the comparison was between two time points (pre and post). This analysis compares the mean scores from time 1 (pre) to the mean scores from time 2 (post) and examines whether the mean scores are significantly different, taking into account standard deviation and sample size. Similarly, independent samples *t*-test was used to compare the means of two groups of students at a specific time point (*i.e.*, pre or post). The sample size is important to consider when conducting *t*-tests because the results are meaningful only when the sample size is big enough to have statistical power.

In MANOVA researchers can use one or more categorical dependent variables and two or more continuous dependent variables (Harlow, 2014). In this work, simple MANOVA tests were ran with two groups (independent variable), and two dependent variables which were the factors of the instrument used. When there are three or more independent variables, a post hoc Tukey test can be applied to see which of the group comparisons displayed statistical significance. In this work, this step was not necessary since there were only two groups being compared to each other at a time.

In addition to the significance tests described, I also examined the effect size of the difference by utilizing Cohens' *d* (Cohen, 1988), or similar effect size indicators. Effect size is another way to test the null hypothesis by not only indicating whether to reject or fail to reject it, but also the degree to which the results deviate from the null hypothesis (Cohen, 1988; Lipsey and Wilson, 2001). Additionally, effect size is a vector measure because it provides a magnitude or degree of deviation from the null hypothesis and also a direction (positive or negative) for the difference when comparing groups or over time (Lipsey and Wilson, 2001).

Relationships with Other Variables

Gathering validity evidence with regards to relationship to other variables is an aspect of state-of-the-art practices in education research delineated by *The Standards* (AERA et al., 2014). Some ways to investigate relationships between variables are correlation analyses, or structural equation modeling (SEM; Kline, 2015). SEM is a multivariate analysis technique in which

researchers can model variable relationships based on theory or empirical evidence (Xu et al., 2013). This technique was used in chapters 3 and 6 to describe the reciprocal relationship between attitude and achievement based on Pekrun's (2006) control-value theory (CVT) of achievement emotions. CVT describes achievement emotions as affective aspects that have direct links to achievement outcomes and can occur before, after, or during achievement activities such as an exam (Pekrun, 2006). In essence, Pekrun (2006) describes a reciprocal causation theoretical model to investigate the influence of achievement on affect and the influence of affect on achievement over a period of time (Marsh *et al*., 2005; Pekrun, Maier and Elliot, 2009; Pekrun *et al.,* 2014). In the field of chemistry education research, these reciprocal models have been investigated in chemistry classrooms (*i.e.*, Gibbons et al., 2018; Gibbons and Raker, 2018). Therefore in this dissertation these models are also utilized to investigate the reciprocal relationship between attitude and achievement (exam scores) over the course of a semester in organic chemistry classrooms.

Cognitive Interviews

In chapter 6, cognitive interviews (Willis, 1999) were conducted with eleven students enrolled in General and Organic Chemistry courses in a southeastern public research university. These interviews had the focus of investigating respondent's interpretation of the items in the ASCIv2 with the purpose of further refining this instrument to reflect the respondents' perceptions. I followed a semi-structured interview approach (Guba and Lincoln, 1983; Wilkinson, Joffe and Yardley, 2004; Curtis and Curtis, 2017) following IRB approved guidelines (see Appendix D).

The interviews were recorded, transcribed and analyzed by the members of the research team in the U.S. Themes emerged from the data that allowed the research team along with collaborators in Chile to refine the *Emotional Satisfaction* scale and develop a new *Utility* scale to create a new instrument reflecting these two factors (ASCI-UE). The results of the qualitative data are reported in chapter 6 as well as in Appendix C.

Data Storage

All data, qualitative and quantitative, was obtained following IRB approved protocols. After data cleaning, all student identifying information was replaced by numerical identifiers to which only the research team has access. The data files are stored in a password-protected work computer which is only in the hands of the researcher. For the qualitative data collection, informed consent forms were signed by each of the students. Those forms are under lock and key with access only by the research team. I have taken every care to follow IRB approved protocols to protect the students' identity in every one of the studies presented in this dissertation.

References

AERA, APA and NCME, (2014), *Standards for educational and psychological testing*, Washington, DC: American Psychological Association.

Aliyu A. A., Bello M. U., Kasim R. and Martin D., (2014), Positivist and non-positivist paradigm in social science research: Conflicting paradigms or perfect partners?, *J. Manage. Sustain.,* **4**(3), 79-95. DOI: 10.5539/jms.v4n3p79

Arjoon J. A., Xu X. and Lewis J. E., (2013), Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence, *J. Chem. Educ.*, **90**, 536–545. DOI:10.1021/ed3002013

Bauer C. F., (2008), Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts, *J. Chem. Educ.*, **85**, 1440−1445. DOI:10.1021/ed085p1440

Brown T. A., (2006), *Confirmatory Factor Analysis for Applied Research*, The Guilford Press, New York, NY.

Bulmer M. G., (1979), *Principles of Statistics*. New York: Dover

Chen F. F., (2007), Sensitivity of goodness of fit indexes to lack of measurement invariance, *Struct. Equ. Modeling*, **14**(3), 464-504. DOI: 10.1080/10705510701301834

Cheng-Hsien L., (2016), Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares, *Behav. Res.,* **48,** 936-949. DOI:10.3758/s13428-015-0619-7

Cohen J., (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ.

Cortina J. M., (1993), What is coefficient alpha? An examination of theory and applications, *J. App. Psychol.,* **78,** 98-104. DOI: 10.1037/0021-9010.78.1.98

Cronbach L. J., (1951), Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**, 297-334. DOI: 10.1007/BF02310555

Curtis B. and Curtis C., (2017), I-Depth Interviewing – The Interactive Base. In *Social Research: A Practical Introduction,* SAGE Publications, Inc.

García N. M., López N. and Vélez V. N., (2018), QuantCrit: Rectifying quantitative methods through critical race theory, *Race Ethn. Educ.,* **21**(2), 149-157. DOI:10.1080/13613324.2017.1377675

Gibbons R. E. and Raker J. R., (2018), Self-beliefs in organic chemistry: Evaluation of a reciprocal causation, cross-lagged model*, J. Res. Sci. Teach.,* **56**(5), 598-615. DOI:10.1002/tea.21515

Gibbons R. E., Xu X., Villafañe S. A. and Raker J. R., (2018), Testing a reciprocal causation model between anxiety, enjoyment and academic performance in postsecondary organic chemistry, *Educ. Psychol.* **38**(6), 838-856. DOI:10.1080/01443410.2018.1447649

Gillborn D., Warmington P. and Demack S., (2018), QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics, *Race Ethn. Educ.,* **21**(2), 158-179. DOI:10.1080/13613324.2017.1377417

Guba E. G. and Lincoln Y. S., (1983), Epistemological and methodological bases of naturalistic inquiry, *Educ. Comm. Tech. J.,* **4**(30), 311-333. ISSN 0148-5806

Harlow, L. L., (2014), *The essence of multivariate thinking: Basic themes and methods,* (2nd Eds), New York, NY: Routledge.

Hu L. T. and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Struct. Equ. Modeling,* **6**(1), 283-292. DOI: 10.1080/10705519909540118

Kenny D. A., Kaniskan B. and McCoach D. B., (2015), The performance of RMSEA in models with small degrees of freedom, *Sociol. Methods Res.*, **44**(3), 486-507. DOI:10.1177/0049124114543236

Kline R. B., (2015), *Principles and Practice of Structural Equation Modeling*, 3rd ed., Guilford Press: New York.

Komperda R., Pentecost T. C. and Barbera J., (2018), Moving beyond alpha: A primer on alternative sources of single-administrations reliability evidence for quantitative chemistry education research, *J. Chem. Educ.*, **95**, 1477-1491. DOI: 10.1021/acs.jchemed.8b00220

Lipsey M. W., and Wilson D. B., (2001), *Practical Meta-Analysis,* Applied Social Research Method Series (Vol. 49), Thousand Oaks: SAGE Publications.

Marsh H. W., Trautwein U., Lüdtke O., Köller O. and Baumert J., (2005), Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering, *Child Dev.,* **76**(2), 397-416. DOI: 10.1111/j.1467-8624.2005.00853.x

Mooring S. R., Mitchell C. E. and Burrows, N. L., (2016), Evaluation of a flipped, large enrollment organic chemistry course on student attitude and achievement, *J. Chem. Educ.*, **93**, 1972−1883. DOI:10.1021/acs.jchemed.6b00367

Muthén L. K. and Muthén B. O., (1998-2007), *Mplus User's Guide*, 5th ed., Muthén & Muthén: Los Angeles, CA.

Osgood C. E., (1965), *Cross cultural comparability of attitude measurement via multi-lingual semantic differentials*. In Recent studies in social psychology, Steiner I. S. and Fishbein M. (Eds.), New York: Holt, Rinehart & Winston, pp. 95-107.

Pekrun R., Maier M. A., Elliot A. J., (2009), Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance, *J. Educ. Psychol*., **101**(1), 115−135. DOI:10.1037/a0013383

Pekrun R., Hall N. C., Goetz T. and Perry R. P., (2014), Boredom and academic achievement: Testing a model of reciprocal causation, *J. Educ. Psychol.,* **106**, 696-710. DOI:10.1037/a0036006

Pekrun R.. (2006), The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice, *Educ. Psychol. Rev.,* **18,** 315–341. DOI:10.1007/s10648-006-9029-9

Rocabado G. A., Kilpatrick N. A., Mooring S. R., and Lewis J. E., (2019), Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course, *J. Chem. Educ.,* **96**(11), 2371-2382. DOI:10.1021/acs.jchemed.9b00516

Sass D., (2011), Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework, *J. Psychoeduc. Assess.*. **29**(4), 347-363. DOI:10.1177/0734282911406661

Wilkinson S., Joffe H. and Yardley L., (2004), Qualitative data collection: interviews and focus groups. SAGE Publications.

Willis G. B., (1999), Cognitive Interviewing: A "How To" Guide. *Meeting of the American Statistical Association,* Research Triangle Institute.

Xu X. (2010), *Refinement of a Chemistry Attitude Measure for College Students*, Dissertation, University of South Florida, Tampa, FL.

Xu X. and Lewis J., (2011), Refinement of a Chemistry Attitude Measure for College Students, *J. Chem. Educ.,* **88**, 561-568. DOI:10.1021/ed900071q

Xu X., Villafañe S. M. and Lewis J. E., (2013), College students' attitudes toward chemistry, conceptual knowledge and achievement: Structural equation model analysis, *Chem. Educ. Res. Pract.,* **14,** 188-200. DOI: 10.1039/C3RP20170H

CHAPTER 3:

CAN WE COMPARE ATTITUDE SCORES AMONG DIVERSE POPULATIONS? AN EXPLORATION OF MEASUREMENT INVARIANCE TESTING TO SUPPORT VALID COMPARISONS BETWEEN BLACK FEMALE STUDENTS AND THEIR PEERS IN AN ORGANIC CHEMISTRY COURSE

Note to Reader

This chapter is a manuscript published in the *Journal of Chemistry Education*, reprinted with permission from:

Rocabado, G. A.; Kilpatrick, N. A.; Mooring, S. R.; Lewis, J. E. Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course. *J. Chem. Educ.* **2019,** *96*(11), 2371-2382. DOI:10.1021/acs.jchemed.9b00516.

Further information regarding permissions can be found on Appendix B.

Introduction

Organic chemistry is a prerequisite course for many fields, not only chemistry careers. Thousands of students attempt mastery of this subject on their way to engineering, science and health professions (Cooper, Grove and Underwood, 2010). With such heavy implications for future professions, organic chemistry is one of the most feared and failed courses in the undergraduate curriculum (Grove, Hershberger and Bretz, 2008; Flynn, 2015), thus acting as a gatekeeper for the target professions (Rowe, 1983; Barr *et al*., 2010). It is important to note that students from all demographic backgrounds who start chemistry, but end up switching to other majors, often do so in the first two years (Zoller, 1990; Grove and Bretz, 2010). The most substantial attrition rates are reported for these gatekeeping courses (Seymour and Hewitt, 1997; Gasiewski *et al*., 2012). Many researchers have focused their efforts in understanding the hurdles that prevent students from succeeding in organic chemistry (Cooper, Grove and Underwood, 2010; Grove and Bretz, 2010; Kraft, Strickland and Bhattacharyya, 2010; Anzovino and Bretz, 2015) and prevailing in their chosen career tracks (Anderson and Bodner, 2008). Seymour and Hewitt (1997) investigated the role of negative feelings toward STEM disciplines and concluded that these play a role in students' decisions to leave. Alternatively, studies have shown that positive emotions such as self-efficacy and perceived autonomy-support (Simon *et al*., 2015), positive attitudes toward science careers (Wyer, 2003), and science identity (Carlone and Johnson, 2007) have influenced students to persist in STEM. Although some researchers have seen significant improvements in success and attrition rates (Grove, Hershberger and Bretz, 2008; Mooring *et al*., 2016) the problem still persists.

Ultimately, as researchers and practitioners we desire to increase retention and success of students in our chemistry classrooms. With this in mind, attitude toward chemistry has been reported to be related to measures of achievement in chemistry courses (Brandriet, Ward and Bretz, 2013; Xu, Villafañe and Lewis, 2013; Villafañe and Lewis, 2016). Villafañe and Lewis (2016) utilize a shortened version of the Test of Science Related Attitudes (TOSRA) instrument and model correlations between students' attitudes and achievement on an American Chemical Society (ACS) final exam in a general chemistry course. Their results indicate a small but significant relationship between two of the TOSRA factors and achievement with a small effect size (f^2 = .02 to .06) when the predictors include race/ethnicity and prior math knowledge (Villafañe and Lewis, 2016). Xu and colleagues (2013) reported a small but significant effect of attitude toward chemistry on achievement measures in general chemistry with medium effect size (f^2 = .19). Brandriet and colleagues (2013) showed a correlational relationship of the two constructs (*emotional satisfaction*, *intellectual accessibility*) in the Attitude toward the Subject of Chemistry Inventory version 2 (ASCIv2) and the ACS exam taken as the final exam in a general chemistry course with large and significant correlation coefficients (between .411 and .522). These studies help establish a relationship between attitude toward science or chemistry and achievement measures in general chemistry.

Several studies have investigated the impact on student attitudes when implementing student-centered active learning pedagogies in gatekeeping chemistry courses. In a study done by Richards-Babb and colleagues, students reported significant attitude improvements when offered online homework as formative assessment with small extra-credit incentives in organic chemistry (Richards-Babb et al., 2015). Case study and context-based learning approaches have also been

shown to produce positive results for student attitudes as students are presented with real-life contexts for the material in college chemistry (Overton, Byers and Seery, 2009; Ültay and Çalik, 2012; Mahaffy *et al*., 2017). Tien and colleagues (2002) reported significant improvements in attitude, retention and achievement when implementing a peer-led team learning instructional approach in organic chemistry. Mooring and colleagues (2016) evaluated attitude gains in organic chemistry classrooms, comparing these gains between traditional lecture and active learning or flipped classrooms. Our interest for this study lies in investigating the attitude-achievement relationship in organic chemistry, particularly because of organic chemistry's reputation of high failure and attrition rates (Grove, Hershberger and Bretz, 2008; Flynn, 2015).

In addition, we note that in the last few decades researchers have explored many dimensions that play significant roles in student engagement and achievement in the classroom, one being the student's demographic background (Baumgartner and Johnson-Bailey, 2008; Charleston *et al*, 2014; Owo and Ikwut, 2015). There seems to be an increase in efforts to make classrooms more inclusive of diverse populations; however, the underrepresentation of gender and ethnic minority groups in STEM is still prevalent (Hurtado *et a*l., 2011). Reports have been issued stating that students' demographic backgrounds, such as gender or race and ethnicity, correlate with how students view the importance of their educational investments (Fordham and Ogbu, 1986; Baumgartner and Johnson-Bailey, 2008; Banerjee *et al*., 2018). Additionally, attitude towards a subject, as well as self-perception and beliefs about inherent abilities in a discipline, has also been connected to a student's gender and racial identity (Catsambis, 1995; Else-Quest, Mineo and Higgins, 2013; Leslie *et al.,* 2015). Thus it becomes of utmost importance for educational researchers to explore the experiences of different demographic groups of students. Baumgartner

and Johnson-Bailey (2008) called for an increased awareness of cultural, gender, racial and ethnic stereotypes because these stereotypes may be detrimental to adult students' learning. Pinder and Blackwell (2013) take these ideas further and identify a possible explanation for the persisting underrepresentation of women of color in STEM. They explained that Black female students' socially constructed meaning of their place in the sciences arises from the interactions with teachers and male peers and might be a source of exclusion from these fields. Archer, Dewitt and Osborne (2015) acknowledged a strong parental influence in Black students' views of science and decisions of whether to pursue STEM careers.

Jackson and Winfield (2014) have issued a call to action to "realign the crooked room" and move towards making STEM classrooms and work places more welcoming to women of color (p. 9). Therefore, we turn our attention to a particular group of students within the classroom: Black[1] female students. This group of students has historically been characterized as displaying negative attitudes toward science in middle school and having low self-perception and belief in their inherent ability to succeed in such disciplines in college and even in academic positions (Catsambis, 1995). The aim of the current research is to explore whether Black female students have negative attitudes toward chemistry as compared to the rest of their peers in an organic chemistry course. Black women in chemistry characterize the "double bind" described by Ong and colleagues (2011), meaning they are individuals representing two minority groups in science simultaneously. The idea of intersectionality (Crenshaw, 1989; Litzler, Samuelson and Lorah,

[1] "Black" will be designated for the students of African descent who classify themselves as belonging to the Black and/or African American race. "White" will be designated for all students of European descent who classify themselves as belonging to the White and/or Caucasian race. Other race classifications at the institution include "Asian", "Native American", and "Pacific Islander". "Other" entails all students whose race is not classified with any of the five major option and/or multiracial students. Throughout this study the main groups of students are Black female and all other students, which designates students from all genders and races who are not Black females.

2014; Ireland *et a*l., 2018) to explain and study the "complex and multidimensional experiences within STEM education" (Ireland *et al*., 2018 p.228) of Black female students has gradually been gaining attention within the STEM education community. Without an appreciation of intersectionality, researchers run the risk of focusing on either racial concerns *or* gender concerns, failing to acknowledge the unique experiences that minority women face in STEM fields (Ong *et al*., 2011). Therefore, bringing to light specific outcomes for Black female students is valuable as we work toward greater inclusivity for STEM fields.

In a study done by Mooring and colleagues (2016) a Multivariate Analysis of Variance (MANOVA) test found attitudinal gains for students experiencing a flipped, first semester organic chemistry (OC1) course; however, whether these gains extended to the Black female students within the class was not explored. Given the history of negative attitudes toward science for Black female students (Catsambis, 1995), it seems worthwhile to inquire whether Black female students within this particular class also experienced similar attitude gains, and whether their attitudes can be linked to their performance in the course. The present study will examine this issue by comparing the attitudes of groups within the original sample.

Currently, instruments that measure cognitive or affective learning traits are widely used in education research across the globe (Marsh *et al*., 2006). When an instrument is translated to a different language and/or when people from other countries and cultures utilize the instrument, often some items are inconsistent with these new contexts (Khaveci, 2015; Xu et al., 2015; Montes, Ferreira and Rodríguez, 2018). However, when instruments are administered to students in a classroom, there has often been an underlying assumption that all students in that classroom,

regardless of their gender, racial or ethnic identities, will view the items in similar contexts. With respect to the earlier study by Mooring and colleagues (2016), the underlying assumption for comparing groups with MANOVA is that the internal structure of the instrument used to measure attitude holds for both groups in a similar way. Therefore, the attitudinal data collected must first be subjected to measurement invariance testing (Gregorich, 2006; Wicherts and Dolan, 2010; Xu, Kim and Lewis, 2016). This sophisticated statistical analysis approach has its roots in the confirmation of the internal structure of the instrument which characterizes the latent constructs (factors) being measured (Sass, 2011). Standards for educational measurement call for evidence of internal structure validity before drawing inferences from measured scores on an instrument (Arjoon, Xu and Lewis, 2013; AERA *et al.,* 2014).

This study also investigates whether a common instrument that has been used in chemistry education research functions as intended for Black female students. While the techniques that are demonstrated in this report can be applied to any student population in any discipline for a variety of different constructs, the present study focuses on Black female students' attitudes toward chemistry in a first semester organic chemistry course. Additionally, we examine the relationship between attitude and achievement and explore the feasibility of a reciprocal causation model (Pekrun, Maier and Elliot, 2009; Gibbons and Raker, 2018; Gibbons *et al*., 2018) between attitude toward chemistry and subsequent exams at the beginning and end of the semester. From the Control-Value Theory (CVT) perspective, achievement emotions, which include affective, cognitive, motivational, expressive, and peripheral physiological processes, are directly linked to achievement activities and outcomes (Pekrun, 2006). These emotions can occur before, after, or during achievement activities as either activity emotions or outcome emotions (Pekrun, 2006). A

reciprocal causation model is appropriate for explaining the role of achievement emotions both as predictors of achievement and as influenced by achievement, consistent with CVT (Pekrun, 2006). This model follows recommendations to investigate the interconnectedness of two or more constructs over time (Marsh *et al*., 2005; Pekrun, Maier and Elliot, 2009; Pekrun *et al.,* 2014), which in this case will be attitude and achievement over a period of one semester in organic chemistry. We address this relationship with this model in the flipped classroom only, as our interest is to see how the attitude change relates to achievement, and achievement relates to attitude change. Significant attitude changes are not observed in the traditional classroom (Mooring *et al*., 2016).

Research Questions

The focus of this study is to investigate whether the attitude gains in a flipped classroom (Mooring *et al*., 2016) extend to the Black female students in the sample. In order to undertake this investigation, we are first interested in studying whether the instrument with which attitude was measured, functions similarly for the Black female students as it does for their peers. Additionally, we investigate whether attitude is related to achievement in this organic chemistry course. With these goals in mind, we have three research questions in this study.

1. To what extent do Black female students experience similar attitude gains as all of their peers in the flipped classroom as reported by Mooring and colleagues in 2016?

2. To what extent are attitude score comparisons appropriate among diverse groups within a sample?
3. How are attitude measures related to achievement measures in Organic Chemistry I in a flipped setting?

Methods

Data were gathered on the Attitude toward the Subject of Chemistry Inventory version 3 (ASCIv3) in Fall 2015 in OC1 classes at a large southeastern public research university (Mooring *et al*., 2016). This instrument is another version of the shortened adaptation (ASCIv2; Xu and Lewis, 2011) from the ASCI developed by Bauer in 2008. The two shortened versions of the original ASCI differ only in the item order (see Appendix C Figures S3.1a and b). The ASCIv2 has a well-established factor structure (AERA *et al.*, 2014) and has been utilized in many English-speaking classrooms (Xu and Lewis, 2011; Xu, Southam and Lewis, 2012; Xu, Villafañe and Lewis, 2013; Xu *et al*., 2015; Mooring *et al*., 2016). It has also been translated to other languages (*e.g*., Turkish and Spanish) and administered to these non-English-speaking classrooms (Khaveci, 2015; Montes, Ferreira and Rodríguez, 2018). However, using measurement invariance to test for population bias has never been done. The sample collected in OC1 classrooms displayed a highly diverse population of students, with Black female students being the largest group. These demographics made this sample suitable to investigate whether ASCIv3 scores can be used to compare Black female students with others in the same class. We also investigate whether item order in this instrument disrupts the well-established two-factor structure shown in ASCIv2.

The ASCIv3 is an 8-item semantic differential instrument that consists of two factors: *intellectual accessibility* (IA) and *emotional satisfaction* (ES) which contain 4 items each. In ASCIv2 Item 2 is Complicated-Simple and Item 8 is Chaotic-Organized. In ASCIv3, Items 2 and 8 switch places. This switch was originally made in the interest of investigating potentially inflated measurement error when three items in a row belong to the same factor, such as items 1, 2, and 3 belonging to the IA factor (Xu, 2010). The item order shown in this report is for ASCIv3; however, the interest is in investigating whether the instrument's factor structure holds true even when the item order switches. Therefore, the ASCIv2 factor structure is used to evaluate this sample. Note that prior work with this data utilized the ASCIv2 factor structure (Mooring *et al*., 2016). The instrument was administered within the first two weeks of the semester in the fall of 2015 before the first exam in OC1, and again at the end of the term after the third exam but before the ACS final exam.

Demographic and Missing Data Analysis

A total of 395 students in OC1 were given the ASCIv3. The categories for gender at the institution are male and female, and the categories for race/ethnicity are Black or African American, White or Caucasian, Asian, Pacific Islander, Native American, and Other. Students self-reported gender and race/ethnicity after completing the ASCIv3 on Qualtrics. The percent of missing values in the sample (without taking into account cases with missing data on all values) is 0.3%. A thorough investigation of missing data is found in Tables S3.1 and S3.2 of the Supporting Information (SI) also found in Appendix C. All of the missing values were handled using full-

information maximum likelihood (FIML) estimation procedures (Klein and Moosbrugger, 2000; Muthén and Asparouhov, 2003).

As a follow-up to the study done by Mooring and colleagues (2016) MANOVA tests were conducted utilizing only the cases that had both pre-test and post-test data (n = 297). This initial analysis agreed with the published findings. Mooring and colleagues (2016) report significant differences in IA gain (F (3,428) = 7.764, $p < 0.001$) as well as ES gain (F (3, 428) = 3.813, p = 0.010) in favor of the students in the flipped classroom when compared to students in a traditional classroom during the fall semester in 2015. Our desire was to investigate whether these gains were extended to Black female students in the flipped-classroom. Descriptive statistics as well as results from the MANOVA analysis are found in Tables S3.3 through S3.5 in Appendix C. Differences in factor gain scores between Black female students and all other students were quantified using Cohen's d value, which indicates the effect size of the difference between mean scores (Cohen, 1988). According to Cohen's (1988) standards, the effect sizes range from small (> 0.2), to medium (> 0.5) to large (> 0.8). If the effect size is smaller than 0.2, the difference between mean scores might be negligible (Cohen, 1988).

Descriptive Statistics

Item mean scores and observed factor mean scores[2] were computed with their respective skewness and kurtosis values for the purpose of examining the normality assumption (see Tables

[2] We are aware that there are different labels for scores obtained by averaging observed data. For example, one label is *grouping variable mean score* as described by Thompson and Green in 2013. In this study, we simply label these scores as observed factor scores

3.1 and 3.2). Observed factor mean scores are calculated as the average score of item means. The homogeneity assumption was tested using Levene's test. Since the present study must check whether differences observed between demographic groups within the sample could be an artifact of how the instrument functions for different populations, we begin by utilizing the entire data set collected (n = 395) and conduct confirmatory factor analysis (CFA), and measurement invariance testing. Additionally, we check that the instrument holds for the student populations in the traditional and flipped classrooms utilizing the same analyses. Finally, we approach the question about the attitude-achievement relationship with structural equation modeling (SEM).

Confirmatory Factor Analysis Criteria

Confirmatory factor analysis was conducted using Mplus (Version 8; Muthén and Muthén, 1998-2007) for Black female students and all other students in a parallel fashion to check whether the internal structure was the same for both groups in each course. The two-factor model established with ASCIv2 (Xu and Lewis, 2011) was utilized. The data were treated as continuous and a maximum likelihood robust (MLR) estimator was used. This estimator takes into account non-normally distributed data. The model parameters were estimated by fixing the first factor loading on each factor to 1.00 and allowing all of the other loadings, variances and covariances to be freely estimated. Model fit statistics were used to determine whether the data fits the model well. To evaluate model fit we examined the chi-square (χ^2) value. The χ^2 is highly influenced by sample size; thus it becomes critical that we inspect additional fit indices, such as, the comparative fit index (CFI), the standardized root-mean square residual (SRMR) and the root mean square of

approximation (RMSEA). The accepted cutoff criteria for these fit indices are as follows: for CFI > .90 is acceptable, but best if > .95; for SRMR < .08; and for RMSEA < .06 (Hu and Bentler, 1999). The RMSEA has been shown to produce inconsistent results with a short instrument due to fewer degrees of freedom (Kenny, Kaniskan and McCoach, 2015); therefore, RMSEA values will be provided but not be considered in comparisons of the measurement invariance models.

Reliability

Reliability of scores was also calculated for each factor in each group. Cronbach's alpha is often a reported measure for reliability (Cronbach, 1951; Cortina, 1999). Coefficient alpha is a measure of how closely related the items within a factor are. However, this coefficient works under the assumption that the model in the study is a *tau*-equivalent or essentially *tau*-equivalent model (Komperda, Pentecost and Barbera, 2018). Basically, "tau-equivalent" means that the measurement model for the set of items that comprise the instrument assumes equal factor loadings for each item in the factor. The measurement model used in this study is not *tau*-equivalent. Instead, we evaluate a congeneric measurement model, in which factor loadings, error variances and all other parameters are freely estimated. Therefore, following Komperda's (2018) suggestion, a more appropriate measure of reliability is coefficient omega. This coefficient is directly calculated using the parameter estimates obtained from confirmatory factor analysis and it is interpreted much like Cronbach's alpha, with higher values (closer to one) indicating high reliability. One caution is that the omega coefficient is only interpretable when there is evidence that the model fits the data well. The equation used to calculate the omega coefficient of reliability

is as shown in Equation 3.1, where λ represents the standardized factor loadings and θ represents the error variances.

$$\omega = \frac{(\Sigma\lambda)^2}{(\Sigma\lambda)^2 + \Sigma\theta} \qquad \text{[Eq. 3.1]}$$

Measurement Invariance Model Fit Criteria and Group Comparisons

Measurement invariance testing was performed for the configural, metric and scalar models comparing Black female students and all other students within the traditional and flipped courses for the 2-factor model prescribed for the pre- and post-tests. Additionally, an overall comparison of pre- and post-test survey administrations was also performed utilizing measurement invariance testing. The logic of measurement invariance testing for two groups is straightforward. The configural invariance model is the least constrained. In this model, only the pattern of fixed and freely estimated factor loadings must be the same for both groups. If fit indices are within the range of acceptable values, the configural model is considered invariant. The next step is to impose a more rigorous constraint: metric invariance is tested by fixing the factor loadings to be the same for both groups. If the fit indices are not significantly different from those for the configural model, the metric model is considered invariant. Finally, even more stringent constraints are imposed, with scalar invariance tested by extending the constraints to equal thresholds (intercepts) for each item. The fit indices produced by the scalar model are compared to those for the metric model (Sass, 2011). Based on Chen (2007) we evaluated $\Delta\chi^2$; however, as noted previously, this value is highly influenced by sample size. Therefore, in addition, we evaluated our results based on the

following fit index cutoffs: ΔCFI (< .01), ΔSRMR (< .03), and ΔRMSEA (< .015) for metric invariance, and ΔCFI (< .01), ΔSRMR(< .01), and ΔRMSEA (< .015) for scalar invariance (Chen, 2007). Results of measurement invariance testing are found in Tables 3.5 through 3.7. Once measurement invariance is established, a comparison of attitude scores can yield meaningful results. Utilizing the model, factor scores were compared between groups. The comparison was made within the scalar model with one of the groups being the control group (factor score at zero) and the other group with freely estimated factor scores. The valence of the factor score indicates whether the score is higher or lower than the control group. For this study, Black female students serve as the focal group and all other students are set as the control group.

Having observed the IA and ES factor score comparisons between both groups within the traditional and flipped classrooms as well as the longitudinal comparisons, we want to further investigate the attitude-achievement relationship. Since the attitude gains were observed in the flipped classroom for the Black female students as well as for all other students, we investigate the relationship in this setting.

Relationship to Other Variables

Structural equation modeling (SEM) is a suitable method to explore the relationship between attitude toward chemistry and achievement measures such as exam scores (Kline, 2015). Five models were tested (A-E) for the students in the flipped classroom using a reciprocal causation framework drawn from control-value theory (Pekrun, 2006; Villafañe, Xu and Raker,

2016; Gibbons and Raker, 2018; Gibbons *et al*., 2018). The best fitting model (A), both statistically and theoretically, was chosen to explain the relationships among the variables. Figure 3.1 displays Model A, while Figure S3.3 and Table S3.6 in the SI show models A-E and their respective fit statistics.

Results

All statistical assumptions for the analyses conducted were met or addressed. The normality assumption was violated for Item 6 (Challenging-Not challenging) due to skewness and kurtosis values outside of the acceptable range of -1 to +1 (Bulmer, 1979; see Tables 1 and 2), thus the MLR estimator was used. Regarding the homogeneity of variance assumption, each item displayed a non-significant Levene statistic (> 0.05). Additionally, all analyses terminated normally and convergence was achieved in parameter estimation.

Following up on the study done by Mooring and colleagues (2016), we wanted first to determine whether Black female students had displayed similar IA and ES gains in the flipped classroom as the rest of their peers. Utilizing the same sample (as in Mooring *et al*., 2016) of 297 students who responded to both the pre-test and the post-test ASCIv3, we noted that the observed factor mean gain scores were similar for Black female students and all of their peers in the two classroom settings. For Black female students just as for all other students, positive gains are associated with the flipped classroom. For the traditional classroom, we observe a small but negative "gain" score for Black female students. Following the descriptive analysis, we performed two MANOVA analyses in SPSS (v24). We compared Black female students (n = 43) and all of

their peers (n = 94) in the flipped classroom and observed no evidence of significant difference in IA gains (F (1, 135) = 0.131, p = 0.718; Cohen's d = .07) and ES gains (F (1, 135) = 0.904, p = 0.343; Cohen's d = .17). Similarly, we compared Black female students (n =57) and all of their peers (n = 103) in the traditional classroom and observed no evidence of significant difference in IA gains (F (1, 158) = 2.220, p = 0.138; Cohen's d = .25) and ES gains (F (1, 158) = 0.381, p = 0.538; Cohen's d = .11). Descriptive statistics including observed factor mean gain scores and MANOVA results are found in SI tables S3.3 through S3.5.These positive results indicate that it is worthwhile to work with the full sample (n = 395) and demonstrate that a score comparison is justifiable from a measurement perspective.

Descriptive Statistics

Descriptive statistics for the entire sample (n = 395) were calculated using SPSS (Version 24) for Black female students and all other students with item and observed factor score means, standard deviation, skewness and kurtosis are presented in Tables 3.1 and 3.2. Items 1, 4, 5 and 7 were reverse coded for ease of interpretation. The negative differential adjectives were coded to be on the lower side of the scale and the positive differential adjectives were coded to appear on the higher side of the scale, with 4 indicating a neutral feeling toward chemistry. Therefore higher values describe higher *intellectual accessibility* or *emotional satisfaction* in the context of chemistry. Mean item scores revealed that students viewed chemistry as relatively hard, challenging, and complicated.

Table 3.1. Descriptive Statistics for Black Female Students in Organic Chemistry I for ASCIv3 (n=125)

	Pre-test				Post-Test			
Items/Factors	Mean[b]	S.D.[c]	Skew.	Kurt.	Mean	S.D.[c]	Skew.	Kurt.
1. Hard-Easy[a]	2.61	1.324	0.593	-0.150	2.82	1.610	0.724	-0.114
3. Confusing-Clear	3.48	1.388	0.027	-0.093	3.75	1.677	-0.004	-0.683
6. Challenging-Not Challenging	2.26	1.481	1.253[e]	1.077[e]	2.29	1.504	1.155[†]	0.562
8. Complicated-Simple	2.61	1.370	0.689	0.439	2.88	1.654	0.696	-0.382
Intellectual Accessibility[d]	**2.75**	**1.111**	**0.602**	**0.593**	**2.94**	**1.316**	**0.336**	**-0.614**
2. Chaotic Organized	4.37	1.577	-0.051	-0.587	4.55	1.647	-0.246	-0.509
4. Uncomfortable-Comfortable[a]	3.45	1.426	-0.032	-0.230	3.80	1.559	0.072	-0.587
5. Frustrating-Satisfying[a]	3.44	1.499	0.017	-0.453	3.38	1.718	0.196	-0.956
7. Unpleasant-Pleasant[a]	3.54	1.335	-0.109	0.088	3.79	1.708	0.010	-0.632
Emotional Satisfaction[d]	**3.70**	**1.085**	**-0.261**	**0.710**	**3.88**	**1.375**	**0.091**	**-0.329**

[a]Items 1, 4, 5 and 7 were reverse coded for ease of interpretation. [b]Each score ranges from 1 to 7, with 4 being the midpoint. High scores mean students feel that chemistry is intellectually accessible or emotionally satisfying. [c]S. D. = Standard deviation. [d]Factor label, boldface for emphasis on composite scores, meaning average scores of observed item means. [e]Value outside of acceptable range.

Descriptively, we observe that Black female students' observed factor score means are consistently lower than the rest of their peers in both classrooms. This observation leads to a question regarding whether the achievement measures also reflect lower achievement scores for the Black female students in this sample. Table 3 displays average exam scores by gender and ethnicity for Exam 1 and the Final Exam in OC1. The lowest average scores on Exam 1 are for Black students. The same pattern is observed for the ACS final. Black female students display some of the lowest average scores in this course, despite representing 32% of the student population in this sample. These results give rise to research question 3.

Table 3.2. Descriptive Statistics for All Other Students in Organic Chemistry I for ASCIv3 (n=270)

	Pre-test				Post-Test			
Items/Factors	Mean[b]	S.D.[c]	Skew.	Kurt.	Mean	S.D.[c]	Skew.	Kurt.
1. Hard-Easy[a]	2.75	1.306	0.620	0.444	3.13	1.463	0.588	-0.185
3. Confusing-Clear	3.70	1.377	0.002	-0.366	4.13	1.576	-0.135	-0.528
6. Challenging-Not Challenging	2.41	1.443	1.119[e]	0.796	2.55	1.496	0.959	0.256
8. Complicated-Simple	2.85	1.263	0.349	-0.001	3.10	1.389	0.422	-0.104
Intellectual Accessibility[d]	**2.93**	**1.025**	**0.200**	**-0.190**	**3.23**	**1.104**	**0.186**	**-0.018**
2. Chaotic Organized	4.61	1.461	-0.182	-0.287	4.62	1.632	-0.510	-0.337
4. Uncomfortable-Comfortable[a]	3.71	1.318	0.059	0.078	4.03	1.515	-0.006	-0.509
5. Frustrating-Satisfying[a]	3.89	1.507	0.160	-0.507	3.98	1.773	0.076	-0.959
7. Unpleasant-Pleasant[a]	3.89	1.367	-0.072	0.269	4.03	1.577	-0.151	-0.331
Emotional Satisfaction[d]	**4.02**	**1.097**	**-0.060**	**-0.300**	**4.17**	**1.300**	**-0.092**	**0.015**

[a]Items 1, 4, 5 and 7 were reverse coded for ease of interpretation. [b]Each score ranges from 1 to 7, with 4 being the midpoint. High scores mean students feel that chemistry is intellectually accessible or emotionally satisfying. [c]S. D. = Standard deviation. [d]Factor label, boldface for emphasis on composite scores, meaning average scores of observed item means. [e]Value outside of acceptable range.

Table 3.3. Exam 1 and ACS Final Exam Mean Scores for Each Demographic Group Represented

	Exam 1 percentage					ACS[c] Final raw score				
Demographics	*N*	Mean[a]	S.D.[b]	Min.	Max.	*N*	Mean[a]	S.D.[b]	Min.	Max.
Black Male	39	75.68	17.300	36	101	31	31.87	8.906	11	59
Asian Male	40	81.91	16.511	30	101	36	35.58	9.869	19	58
White Male	32	77.77	22.234	0	102	29	37.17	12.361	18	63
Other Male	17	86.41	11.133	59	99	16	36.25	10.043	15	50
Black Female	121	75.81	16.771	21	101	118	32.51	8.865	10	53
Asian Female	63	79.05	19.032	0	100	63	34.02	8.511	18	53
White Female	47	80.54	17.188	28	103	43	34.49	11.465	0	55
Other Female	26	79.62	20.646	0	100	25	35.36	9.552	19	52

[a]Exam 1 scores are based on percentage scores that students were awarded in class. Students may earn extra credit points on the exam, therefore scores of greater than 100 are possible. The ACS scores are the "raw" scores which are the number of correct responses students got out of 70 possible points. [b]S.D. = Standard deviation. [c]ACS = American Chemical Society.

In order to fully analyze these descriptive findings, we must first check whether the differences observed are an artifact of the instrument. We investigate whether the internal structure of the instrument holds for both groups by first employing confirmatory factor analysis (CFA) techniques.

Confirmatory Factor Analysis and Reliability

CFA was conducted for both groups in a parallel fashion to ensure that the 2-factor structure prescribed in the literature for the ASCIv2 (Xu and Lewis, 2011) could be applied to the ASCIv3 data and would be the same for both groups. Initially the model fit was not acceptable for either group in OC1 at the beginning of the semester (Black female: χ^2 (n = 125, df = 19, p = 0.0001) = 52.535; CFI = .875; SRMR = .071; RMSEA= .126; all other: χ^2 (n = 270, df = 19, $p <$ 0.0001) = 87.505; CFI = .848; SRMR = .094; RMSEA = .123). Therefore, model modification indices were examined for potential correlated errors. Each modification suggestion was evaluated in both the statistical and theoretical sense as suggested by Wang & Wang (2012). The final model with a summary of the modifications and rationale for each can be found in Appendix C (Figure S3.2). The final CFA contains 2 correlated errors that were justified theoretically and empirically (Xu, 2010; Xu *et al*., 2015; Montes, Ferreira and Rodríguez, 2018). Table 3.4 displays model fit statistics for Black female students and all other students for the pre- and post-test as well as the accompanying reliability for each factor. In all cases, the results indicated that the data fit the final model well (Hu and Bentler, 1999). Standardized loadings, error terms and correlations are displayed in Figure S3.2. All parameters were statistically significant at the .05 level.

Reliability was calculated using the omega coefficient (Table 3.4). These results indicate acceptable internal consistency of the two factors with the exception of the IA factor in the post-test for all other students, which displays slightly lower than acceptable reliability scores (< .7). Comparisons between the two groups for IA in the post-test should therefore be cautious. In addition to the omega coefficient, we also calculated reliability coefficients for a multidimensional model with correlated factors as described by Cho (2016). In each case, the reliability results from these calculations were the same as omega (see Table S3.7). We have also calculated Cronbach's alpha coefficients (Table S3.7); however, as previously noted Cronbach's alpha assumes a *tau*-equivalent model, which this is not.

Table 3.4. Confirmatory Factor Analysis and Internal Consistency Reliability for Black Female Students and All Other Students for Pre- and Post-test

Time	Group	*N*	χ^2	*df*	*p*	CFI	SRMR	RMSEA	Omega IA[a]	Omega ES[b]
Pre	**Black Female**	125	33.099	17	0.0110	0.940	0.056	0.092	0.809	0.749
	All Other	270	42.747	17	0.0005	0.943	0.060	0.080	0.725	0.790
Post	**Black Female**	125	39.038	17	0.0018	0.943	0.047	0.107	0.797	0.794
	All Other	270	37.587	17	0.0028	0.965	0.044	0.072	0.688	0.782

[a] IA = Intellectual Accessibility. [b] ES = Emotional Satisfaction.

Measurement Invariance Models

The previous analyses provide sufficient grounds to conduct measurement invariance testing for Black female students as compared to all other students, which requires a well-defined factor structure with goodness of fit indicators that suggest good model fit for both pre- and post-

tests. All the previous results show these requirements are met for this sample; therefore, configural, metric, and scalar invariance testing can be undertaken. The differences in fit statistics for the configural, metric, and scalar models are shown in Table 3.5 and 3.6 for OC1 pre- and post-test respectively. The results indicate that the changes in fit statistics are within the cutoffs, and invariance between groups is well established for this model (Chen, 2007).

Attention to whether the internal structure of the instrument holds over time is important, since the interest is in observing attitude gains over the course of the semester. Measurement invariance testing between the flipped course and the traditional course at the beginning and end of the semester confirmed that that the structure holds for both groups (Tables S3.8 and S3.9 in Appendix C). Furthermore, longitudinal measurement invariance testing was performed for all students in OC1 to check whether the factor structure holds over time for all students. Table 3.7 indicates that the configural and metric and scalar models all hold over time, so comparisons between pre-test and post-test scores can be made.

Table 3.5. Measurement Invariance Between Black Female Students and All Other Students in Organic Chemistry I (Pre-test)

Model	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural[a]	76.242	34	<0.0001	0.942	0.059	-	-	-	-	-
Metric[b] vs. Configural	81.557	40	0.0001	0.943	0.067	5.315	6	0.5041	0.001	0.008
Scalar[c] vs. Metric	85.027	46	0.0004	0.947	0.066	3.470	6	0.7480	0.004	0.001

[a]The configural model is a comparison model for both groups without constraints. [b]The metric model adds the constraint of equal factor loadings for both groups. [c]The scalar model adds the constraint of equal intercepts for both groups. Each constraint is added one at a time. Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Black female students (n = 125) and all other students including Black male, White, Asian and Other both male and female students (n = 270).

Table 3.6. Measurement Invariance Between Black Female Students and All Other Students in Organic Chemistry I (Post-test)

Model	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural[a]	76.520	34	<0.0001	0.957	0.045	-	-	-	-	-
Metric[b] vs. Configural	82.420	40	0.0001	0.957	0.057	5.900	6	0.4345	0.000	0.006
Scalar[c] vs. Metric	91.264	46	0.0001	0.954	0.060	8.844	6	0.1825	0.003	0.003

[a]The configural model is a comparison model for both groups without constraints. [b]The metric model adds the constraint of equal factor loadings for both groups. [c]The scalar model adds the constraint of equal intercepts for both groups. Each constraint is added one at a time. Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Black female students (n = 125) and all other students including Black male, White, Asian and Other both male and female students (n = 270).

Table 3.7. Measurement Invariance Pre-Post for All Organic Chemistry I Students

Model	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural[a]	211.12	94	<0.0001	0.939	0.055	0.056	-	-	-	-	-	-
Metric[b] vs. Configural	222.751	100	<0.0001	0.937	0.059	0.056	11.631	6	0.0710	0.002	0.004	0.000
Scalar[c] vs. Metric	237.915	106	<0.0001	0.932	0.059	0.056	15.164	6	0.0190	0.005	0.000	0.000

[a]The configural model is a comparison model for both groups without constraints. [b]The metric model adds the constraint of equal factor loadings for both groups. [c]The scalar model adds the constraint of equal intercepts for both groups. Each constraint is added one at a time. Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison is between the pre-test and post-test administrations (n=395).

Latent Factor Score Comparisons

With measurement invariance established between groups in the pre- and post-tests we can investigate whether the IA and ES latent factor scores for each group differ. The measurement model at the most constrained setting (scalar model) can be used to compare latent factor scores (DiStefano, Zhu and Mîndrila, 2009; Thompson and Gren, 2013) between a control group and a

focal group. The comparisons of interest appear in Tables 3.8 through 3.11. Latent factor scores utilizing the measurement model are standardized to a mean of zero (DiStefano, Zhu and Mîndrila, 2009; Thompson and Gren, 2013). In other words, the control group's latent factor score is set to zero, whereas the latent factor score for the focal group can deviate from zero, thereby allowing the comparison. The maximum deviation for these standard solutions is -1 to +1. A deviation in the upper or lower quarter of this range is therefore relatively large, whereas deviations closer to zero are quite small. In other words, the deviation of .671 observed for the Post-IA scores between traditional and flipped classrooms represents a large difference favoring the flipped classroom, but the differences between these two classrooms for Pre-IA (.110) and Pre-ES (.073) are not notable (see Table 3.8).

Table 3.8. Latent Factor Score Comparison Between Traditional and Flipped Classroom and the Beginning and End of the Semester

	PRE			POST		
Factor	Traditional[a]	Flipped[b]	*p*	Traditional[a]	Flipped[b]	*p*
Intellectual Accessibility	0.000	0.110	0.364	0.000	0.671	< .0001
Emotional Satisfaction	0.000	0.073	0.181	0.000	0.286	0.002

[a]Reference group with latent mean score of zero. [b]Latent factor score calculated as a deviation from the reference group.

These results support the MANOVA reported in Mooring *et al.,* (2016) in which no evidence of significant difference was observed between the two classrooms at the beginning of the semester, yet a significant difference was observed at the end of the semester. In the present study, the entire sample (n = 395) is analyzed and the measurement model is taken into account, demonstrating that this result is robust. The present study takes the analysis further by positioning

Black female students as the focal group, first within the entire sample (Table 3.9). The comparison of latent factor scores between Black female students and all of their peers, regardless of the classroom in which they were enrolled, confirms that Black female students have lower IA and ES scores (Table 3.9). Though these differences are relatively small, they are larger rather than smaller by the end of the term. This trend was foreshadowed by the simple descriptive statistics (Tables 3.1 and 3.2).

Table 3.9. Latent Factor Score Comparison Between Black Female Students and All Other Students in Both Classrooms at the Beginning and End of the Semester

Factor	PRE			POST		
	All Other[a]	Black Female[b]	*p*	All Other[a]	Black Female[b]	*p*
Intellectual Accessibility	0.000	-0.192	0.149	0.000	-0.359	0.027
Emotional Satisfaction	0.000	-0.147	0.021	0.000	-0.217	0.023

[a]Reference group with latent mean score of zero. [b]Latent factor score calculated as a deviation from the reference group.

Examining the two classrooms separately demonstrates the relative advantage of the flipped classroom for Black female students with respect to attitude. The result in Table 3.10 for the traditional classroom echoes the "negative gains" for attitude in the simple descriptive statistics for the Black female students in this setting (Table S3.3). At the beginning of the term, Black female students have only slightly lower latent factor scores than their peers, but by the end of the term the differences are medium to large. In the flipped classroom (Table 3.11), although the Black female students still do have lower latent factor scores than their peers, the differences are never large.

Table 3.10. Latent Factor Score Comparison Between Black Female Students and All Other Students at the Beginning and End of the Semester in the Traditional Classroom

Factor	PRE			POST		
	All Other[a]	Black Female[b]	*p*	All Other[a]	Black Female[b]	*p*
Intellectual Accessibility	0.000	-0.245	0.180	0.000	-0.522	0.021
Emotional Satisfaction	0.000	-0.136	0.058	0.000	-0.321	0.007

[a]Reference group with latent mean score of zero. [b]Latent factor score calculated as a deviation from the reference group.

Table 3.11. Latent Factor Score Comparison Between Black Female Students and All Other Students at the Beginning and End of the Semester in the Flipped Classroom

Factor	PRE			POST		
	All Other[a]	Black Female[b]	*p*	All Other[a]	Black Female[b]	*p*
Intellectual Accessibility	0.000	-0.184	0.409	0.000	-0.172	0.531
Emotional Satisfaction	0.000	-0.145	0.285	0.000	-0.061	0.668

[a]Reference group with latent mean score of zero. [b]Latent factor score calculated as a deviation from the reference group.

While the research design does not provide evidence that the flipped classroom setting closed an attitude gap between Black female students and their peers, the data suggest that this flipped classroom provided a positive environment for Black female students with respect to attitude. Black female students in this setting report that chemistry is more *intellectually accessible* and *emotionally satisfying* at the end of the term than at the beginning, with end of term attitude scores close to those of their peers. This is a promising finding, but questions remain regarding the relationship between attitude and achievement in this setting. The established measurement model can be used in conjunction with structural equation modeling (SEM) techniques to address these questions

Reciprocal Causation Model for Attitude and Achievement Relationship

After gathering validity evidence for the internal structure of the ASCIv3 for the student groups in the flipped classroom, we moved to address the relationship between attitude and achievement, which in turn is yet another aspect of validity evidence (Arjoon, Xu and Lewis, 2013; AERA *et al*., 2014). It has been reported that attitude toward chemistry has a relationship to achievement measures such as SAT math scores and also class exam scores and American Chemical Society (ACS) exam scores (Xu, Villafañe and Lewis, 2013). For the flipped classroom students, models A-E were tested and analyzed. Fit results for model A (Figure 3.1) meet acceptable standards and are as follows: χ^2 (n = 194, df = 123, p < .0001) = 191.758; RMSEA = .064; CFI = .927; and SRMR = .069. In this model the relationship between attitude and achievement follows a reciprocal causation logic (Marsh *et al*., 2005; Pekrun *et al*., 2014; Gibbons *et al*., 2018) between the attitude constructs and the achievement measures throughout the semester (see Figure 3.1). All paths in Model A were significant at the .05 level except for non-significant paths from IA pre-test and post-test to Exam 1 and the ACS final exam respectively. Models B-E are shown in Appendix C (Figure S3.3), and represent a set of more parsimonious models. Model B removed the path between Exam 1 and IA Post. Model C removed the path between Exam 1 and ES Post. Model D removed the path between ES Post and ACS Final. Model E removed the path between ES Pre and Exam 1. The results for Models B through E displayed convergence and normal estimation; however, the fit indices showed worse fit than model A (Table S3.6).

In Model A, students' ES as measured by the ASCIv3 has a small, yet direct and positive relationship with performance on the subsequent exam, both at the beginning and the end of the

semester. In turn, performance on the first exam has a small to medium, direct, and positive relationship on subsequent attitude toward chemistry as measured by both constructs (IA and ES). It follows that students who have lower ES scores at the beginning of the term have a greater possibility of doing poorly on their exams, and that this trend persists over the course of the semester. It is notable that attitude remains a significant predictor of final exam scores in the model even with first exam scores taken into account. We take this opportunity to notice that Black female students in this classroom, on average, display lower ES scores than their peers and also some of the lowest scores on both tests. Even though the coefficients associated with attitude do not rise above small to medium in effect, the consistent relationship between attitude and achievement is worth considering both in research and teaching.

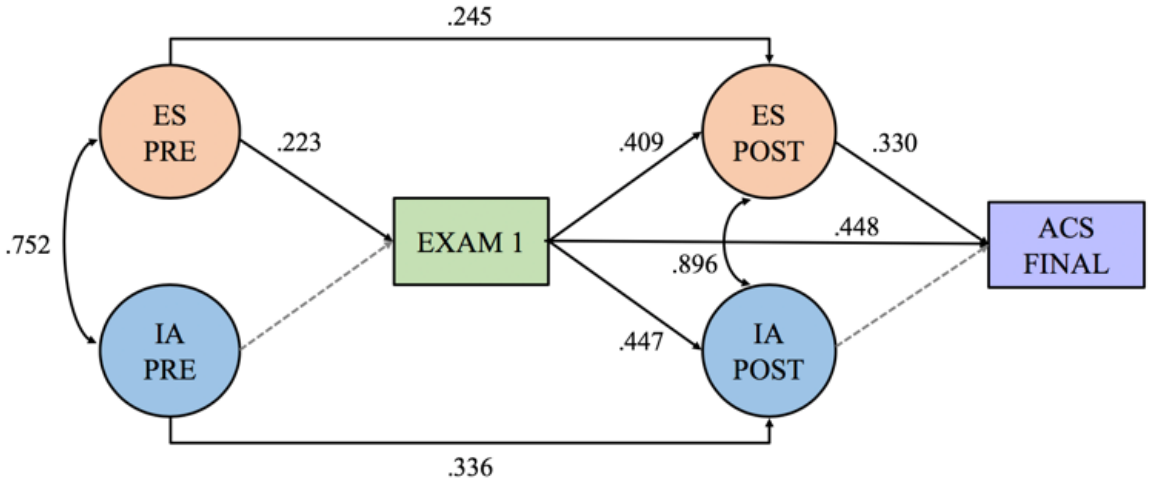


Figure 3.1. SEM Model A for Organic Chemistry I students in the flipped classroom. All values are significant at the .05 level. Dashed arrows mean non-significant paths in the model. $N = 194$.

Conclusions and Implications

As a gatekeeper course, organic chemistry prevails as one of the most feared courses in undergraduate education (Rowe, 1983; Grove, Hershberger and Bretz, 2008; Barr *et al*., 2010; Flynn, 2015). Given this reputation, students from all backgrounds struggle to understand and successfully complete the requirements to pass this course based on the high attrition rates observed in most classrooms (Zoller, 1990; Grove and Bretz, 2010; Gasiewski *et al*., 2012). Although in this study we did not address attrition rates directly, when looking at a specific group of students who have historically been underrepresented in the sciences, we see that Black female students do worse than their peers. While all of the students in the flipped class experienced attitude gains over the course of the semester, Black female students began and ended lower and also did worse on exams than their peers (Table 3.3). Additionally, we observed no evidence of the attitude gap closing for these students. This concerning issue is a compelling reason to investigate this particular group in greater depth and to ensure that the interventions we implement and outcomes we measure extend to Black female students.

As many studies have demonstrated, demographic background can play a role in differential student outcomes (Baumgartner and Johnson-Bailey, 2008; Charleston *et al*., 2014; Owo and Ikwut, 2015). In this study we have found that Black female students begin organic chemistry with lower attitudes than the rest of their peers and although their attitude improves in a flipped classroom environment, the attitude gap does not close. Every student has a distinctive experience in the classroom due to their unique set of identities such as race, ethnicity, gender, orientation, socio-economic status, academic goals, and so forth. This phenomenon is understood

as a consequence of intersectionality (Crenshaw, 1989; Litzler, Samuelson and Lorah, 2014; Ireland *et al*., 2018). Although the investigation of unique narratives of students is beyond the scope of the present study, we have tried to set the stage for this work by acknowledging the fact that Black female students are caught in the "double bind" described by Ong and colleagues (2011). While much of the research in general pays attention to either gender or race/ethnicity, students who belong to two or more minority groups simultaneously are sometimes forced into one or the other (Crenshaw, 1989; Ong *et al*., 2011). The interventions we implement or outcomes we measure do not often extend to considering complex identities (Crenshaw, 1989; Ong *et al*., 2011; Litzler, Samuelson and Lorah, 2014; Ireland *et al*., 2018). Thus, evaluations of interventions in our classrooms, including those designed to improve attitudes for underrepresented minority groups found at different intersections, should consider those groups particularly. While we encourage researchers and practitioners to utilize published instruments, or develop new instruments, to assess the effects of interventions, we urge practitioners to work with researchers to test instruments in their own specific classrooms as we have done here. We cannot assume that instruments lead to valid inferences irrespective of context, and we must acknowledge that students within a classroom can be experiencing different contexts. Every classroom is a unique setting in which intersectional identities exist. We can develop a greater understanding of diverse experiences when we approach a larger variety of classrooms across the nation. This understanding will aid in the design of appropriate interventions to address achievement and retention for all students. As we move toward more diverse and inclusive environments in our chemistry courses, a commitment to considering whether interventions have positive results for different groups of students is of great consequence.

This study is the first within chemistry education to consider measurement invariance for Black female students compared to all other students. Confirmatory factor analysis and measurement invariance testing revealed that Black female students display similar patterns while answering items on the ASCIv3 to all of their peers, thus score comparisons were appropriate for these two groups. Had we not conducted measurement invariance testing, we would lack the evidence that allows us to discount score differences between groups as artifacts of the instrument. We encourage other researchers to go back to their data and check whether their group comparisons are appropriately supported, just like we did with these data. Moving forward, we encourage researchers to utilize these techniques whenever possible, to ensure proper and meaningful comparisons and to bring more awareness of the range of experiences for diverse students in our classrooms.

It is natural for educators and researchers to want to design content-focused interventions that directly impact achievement outcomes. This practice is important; for example, in this study the relationship between Exam 1 and the final ACS exam is a strong relationship. However, it is also vital not to dismiss the role that affective constructs, such as attitude, can have on achievement. As we have investigated in this study, the effect of attitude on achievement, particularly *emotional satisfaction* on achievement, although small, it is a significant effect and will only help improve achievement outcomes. Some studies done in organic chemistry have observed positive attitude outcomes in student-centered active-learning environments (Tien, Roth and Kampmeier, 2002; Overton, Byers and Seery, 2009; Ültay and Çalik, 2012; Richards-Babb *et al*., 2015). Much like in those studies, we observed that all students in an active learning environment saw positive changes in attitude from the beginning to the end of the semester;

however, there was no evidence of closing an "attitude gap." Therefore, on the basis of this study and others it seems reasonable to recommend that faculty consider implementing a flipped class approach (Mooring *et al*., 2016) or other effective approaches such as formative online homework assessments (Richards-Babb *et al*., 2015), peer-led team learning (Tien, Roth and Kampmeier, 2002), or context-based learning (Overton, Byers and Seery, 2009; Ültay and Çalik, 2012; Mahaffy *et al*., 2017) to help foster more positive student attitudes in organic chemistry courses while remaining alert to the possibility that attitude gaps may remain.

Replicating this study with other student populations and extending the work to additional measures is desirable in order to create a clear map of the landscape relating attitude, achievement, and identity to achievement and retention in chemistry courses. We encourage researchers to utilize the techniques outlined herein for group comparisons, and to continue to build best practices for measurement in order to advance in this field of knowledge. We also call for specific research to look for more ways to improve both attitude and achievement for Black female students in organic chemistry classrooms, and to describe the diverse experiences of students who encounter chemistry courses as part of a larger undergraduate curriculum.

Limitations

A set of limitations arise from having a convenience sample. Students from two Organic Chemistry I classes at a large public research university in the southeastern United States are represented in this data set. Although in this sample the largest subgroup is Black female students,

the results of this study are not evidence that the instrument used herein is applicable for all Black female student groups everywhere. Rather it is evidence that for the data collected, the instrument functions properly for the student groups on which we focused. We encourage all researchers and practitioners to check for population biases when using instruments in their classrooms following the process outlined in this study or other processes that check whether the instrument functions as intended for different populations within a sample. Future endeavors in this area should include the exploration of model invariance and attitude changes among diverse populations with different demographic groups, as well as a more focused comparison between two specific groups (*i.e.*, Black female and White female) when possible. Due to the specifics of the sample in this study, we were limited to comparing Black females with all other students.

We were limited in our ability to address the issue of intersectionality with a purely quantitative study. This issue cannot be fully studied *en masse*, since this concept arises from the unique experiences each individual has at the intersection of all of the identities each person possesses (Crenshaw, 1989; Ong *et al.*, 2011; Litzler, Samuelson and Lorah, 2014; Ireland *et al.*, 2018). This issue would be best addressed in a thoughtful qualitative study where individual narratives can be brought to light in a meaningful way, but we hope we have demonstrated here that quantitative research can include an intersectional awareness.

Moreover, the need to explore further the semantic meaning of each item word pair in this instrument has come to our attention as we have worked on this project. Research that includes cognitive interviews with students regarding the ASCIv2 items would be warranted and timely as

a means to gather more validity evidence for this instrument from a respondent's perspective (Arjoon, Xu and Lewis, 2013; AERA *et al*., 2014).

References

AERA, APA, and NCME, (2014), *Standards for Educational and Psychological Testing.* American Psychological Association: Washington, DC.

Anderson T. L. and Bodner G. M., (2008), What can we do about 'Parker'? A case study of a good student who didn't 'get' organic chemistry, *Chem. Educ. Res. Pract.,* **9,** 93-101. DOI: 10.1039/B806223B

Anzovino M. E. and Bretz S. L., (2015), Organic chemistry students' ideas about nucleophiles and electrophiles: The role of charges and mechanisms, *Chem. Educ. Res. Pract.,* **16**, 797-810. DOI: 10.1039/C5RP00113G

Archer L., Dewitt J. and Osborne J., (2015), Is science for us? Black students' and parents' views of science and science careers, *J. Sci. Educ.*, **99**(2), 199-237. DOI: 10.1002/sce.21146

Arjoon J. A., Xu X. and Lewis J., (2013), Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence, *J. Chem. Educ.*, **90**, 536-545. DOI: 10.1021/ed3002013

Banerjee M., Schenke K., Lam,A. and Eccles J. S., (2018), The roles of teachers, classroom experiences, and finding balance: A qualitative perspective on the experiences and expectations of females within STEM and non-STEM careers, *Int. J. Gender Sci. Tech,.* **10**(2), 287-307.

Barr D., Matsui J., Wanat S. F. and Gonzalez M., (2010), Chemistry courses as the turning point for premedical students, *Adv. Health Sci. Educ*., **15**, 45-54. DOI:10.1007/s10459-009-9165-3

Bauer C. F., (2008), Attitude towards chemistry: A semantic differential instrument for assessing curriculum impacts, *J. Chem. Educ.*, **85**, 1440-1445. DOI: 10.1021/ed085p1440

Baumgartner L. M. and Johnson-Bailey J., (2008), Fostering awareness of diversity and multiculturalism in adult and higher education, *New Dir. Adult Cont. Educ.,* **120**, 45-53. DOI:10.1002/ace.315

Brandriet A. R., Ward R. M. and Bretz S. L., (2013), Modeling meaningful learning in chemistry using structural equation modeling, *Chem. Educ. Res. Pract.,* **14,** 421-430. DOI: 10.1039/C3RP00043E

Bulmer M. G., (1979), *Principles of Statistics*. New York: Dover.

Carlone H. B. and Johnson A., (2007), Understanding the science experiences of successful Women of Color: Science identity as an analytic lens, *J. Res. Sci. Teach.,* **44**(8), 1187-1218. DOI: 10.1002/tea.20237

Catsambis S., (1995), Gender, race, ethnicity, and science education in the middle grades, *J. Res. Sci. Teach*., **32**(2), 243-257. DOI: 10.1002/tea.3660320305

Charleston L. J., George P. L., Jackson J. F. L., Berhanu J. and Amechi M. H. J., (2014), Navigating underrepresented STEM spaces: Experiences of Black women in U.S.

computing science higher education programs who actualize success, *Divers. High. Educ.,* **7**(3), 166-176. DOI: 10.1037/a0036632
Chen F. F., (2007), Sensitivity of goodness of fit indexes to lack of measurement invariance, *Struct. Equ. Modeling*, **14**(3), 464-504. DOI: 10.1080/10705510701301834
Cho E., (2016), Making reliability reliable: A systematic approach to reliability coefficients, *Organ. Res. Methods,* **19**(4), 651-682. DOI: 10.1177/1094428116656239
Cohen J., (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ.
Cooper M. M., Grove N. and Underwood S. M., (2010), Lost in Lewis structures: An investigation of student difficulties in developing representational competence. *J. Chem. Educ.,* **87**(8), 869-874. DOI: 10.1021/ed900004y
Cortina J. M., (1999), What is coefficient alpha? An examination of theory and applications, *J. App. Psychol.,* **78,** 98-104. DOI: 10.1037/0021-9010.78.1.98
Crenshaw K., (1989), Demargenalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics, *University of Chicago Legal Forum*, 139-168.
Cronbach L. J., (1951), Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**, 297-334. DOI: 10.1007/BF02310555
DiStefano C., Zhu M. and Mîndrila D., (2009), Understanding and using factor scores: Considerations for the applied researcher, *Pract. Assess. Res. Eval.,* **14**(20), 1-11. DOI:10.7275/da8t-4g52
Else-Quest N. M., Mineo C. C. and Higgins A., (2013), Math and science attitudes and achievement at the intersection of gender and ethnicity, *Psychol. Women Quarterly,* **37**(3), 293-309. DOI: 10.1177/0361684313480694
Flynn A. B., (2015), Structure and evaluation of flipped chemistry courses: Organic & spectroscopy, large and small, first to third years, English and French, *Chem. Educ. Res. Pract.,* **16**, 198-211. DOI: 10.1039/C4RP00224E
Fordham S. and Ogbu J. U., (1986), Black students' school success: Coping with the "burden of acting White", *Urban Rev.*, **18***,* 176-206.
Gasiewski J. A., Eagan M. K., Garcia G. A., Hurtado S. and Chang M. J., (2012), From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory STEM courses, *Res. High. Educ.,* **53**, 229-261. DOI:10.1007/s11162-011-9247-y
Gibbons R. E. and Raker J. R., (2018), Self-beliefs in organic chemistry: Evaluation of a reciprocal causation, cross-lagged model*, J. Res. Sci. Teach.,* **56**(5), 598-615. DOI:10.1002/tea.21515
Gibbons R. E., Xu X., Villafañe S. M. and Raker J. R., (2018), Testing a reciprocal causation model between anxiety, enjoyment and academic performance in postsecondary organic chemistry, *Educ. Psychol.*, **38**(6), 838-856. DOI: 10.1080/01443410.2018.1447649
Gregorich S. E., (2006), Do self-report instruments allow meaningful comparisons across diverse population groups?, *Medical Care,* **44** (11 Supl 3), S78-S94. DOI: 10.1097/01.mlr.0000245454.12228.8f
Grove N. P. and Bretz S. L., (2010), Perry's scheme of intellectual and epistemological development as a framework for describing student difficulties in learning organic chemistry, *Chem. Educ. Res. Pract.,* **11,** 207-211. DOI: 10.1039/C005469K

Grove N. P., Hershberger J. W. and Bretz S. L., (2008), Impact of a spiral organic curriculum on student attrition and learning, *Chem. Educ. Res. Pract.,* **9,** 157-162. DOI: 10.1039/B806232N

Hu L. T. and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Struct. Equ. Modeling,* **6**(1), 283-292. DOI: 10.1080/10705519909540118

Hurtado S., Newman C. B., Tran M. C. and Chang M. J., (2011), Improving the rate of success for underrepresented racial minorities in STEM fields: Insights from a national project, *New Dir. Inst. Res.*, **148,** 5-15. DOI: 10.1002/ir.357

Ireland D. T., Freeman K. E., Winston-Proctor C. E., DeLaine K. D., McDonald Lowe S. and Woodson K. M., (2018), (Un)hidden figures: A synthesis of research examining the intersectional experiences of Black women and girls in STEM, *Rev.Res. Educ.***, 42,** 226-254. DOI: 10.3102/0091732X18759072

Jackson K. M.; and Winfield L. L., (2014), Realigning the crooked room: Spelman claims a space for African American women in STEM, *Peer Rev.,* **16**(2), 9-12.

Kenny D. A., Kaniskan B. and McCoach D. B., (2015), The performance of RMSEA in models with small degrees of freedom, *Sociol. Methods Res.*, **44**(3), 486-507. DOI:10.1177/0049124114543236

Khaveci A., (2015), Assessing high school student's attitudes toward chemistry with a shortened semantic differential, *Chem. Educ. Res. Pract.*, **16,** 283-292. DOI: 10.1039/C4RP00186A

Klein A. and Moosbrugger H., (2000), Maximum likelihood estimation of latent interaction effects with the LMS method, *Psychometrika,* **65**(4), 457-474.

Kline R. B., (2015), *Principles and Practice of Structural Equation Modeling*, 3rd ed., Guilford Press: New York.

Komperda R., Pentecost T. C. and Barbera J., (2018), Moving beyond alpha: A primer on alternative sources of single-administrations reliability evidence for quantitative chemistry education research, *J. Chem. Educ.*, **95**, 1477-1491. DOI: 10.1021/acs.jchemed.8b00220

Kraft A., Strickland A. M. and Bhattacharyya G., (2010), Reasonable reasoning: Multi-variate problem-solving in organic chemistry, *Chem. Educ. Res. Pract.,* **11**, 281-292. DOI: 10.1039/C0RP90003F

Leslie S. J., Cimpian A., Meyer M. and Freeland E., (2015), Expectations of brilliance underlie gender distributions across academic disciplines, *Science*, **347**(6219), 262-265. DOI:10.1126/science.1261375

Litzler E., Samuelson C. C. and Lorah J. A., (2014), Breaking it down: Engineering students STEM confidence at the intersection of race/ethnicity and gender, *Res. High. Educ.*, **55,** 810-832. DOI 10.1007/s11162-014-9333-z

Mahaffy P. G., Holme T. A., Martin-Visscher L., Martin B. E., Versprille A., Kirchhoff M., McKenzie L. and Towns M., (2017), Beyond "inert" ideas to teaching general chemistry from rich contexts: Visualizing the chemistry of climate change (VC3), *J. Chem. Educ.,* **94**, 1027-1035. DOI: 10.1021/acs.jchemed.6b01009

Marsh H. W., Hau K. T., Arlet C., Baumert J. and Peshar J. L., (2006), OECD's brief self-report measure of educational psychology's most useful sffective constructs: Cross-cultural, psychometric comparison across 25 countries, *Int. J. Test.*, **6**(4), 311-360. DOI:10.1207/s15327574ijt0604_1

Marsh H. W., Trautwein U., Lüdtke O., Köller O. and Baumert J., (2005), Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering, *Child Dev.,* **76**(2), 397-416. DOI: 10.1111/j.1467-8624.2005.00853.x

Montes L. H., Ferreira R. A., Rodriguez C., (2018), Explaining secondary school student's attitudes towards chemistry in Chile, *Chem. Educ. Res. Pract.,* **19**(2), 533-542. DOI: 10.1039/C8RP00003D

Mooring S. R., Mitchell C. E. and Burrows N. L., (2016), Evaluation of a flipped, large enrollment organic chemistry course on student attitude and achievement, *J. Chem. Educ.,* **93**, 1972-1883. DOI: 10.1021/acs.jchemed.6b00367

Muthén B. and Asparouhov T., (2003), Modeling interactions between latent and observed continuous variables using maximum-likelihood estimation in Mplus, *Mplus Web Notes.* **6** (1).

Muthén L. K. and Muthén B. O., (1998-2007), *Mplus User's Guide*, 5th ed., Muthén & Muthén: Los Angeles, CA.

Ong M., Wright C., Espinosa L. L. and Orfield G., (2011), Inside the double bind: A synthesis of empirical research on undergraduate and graduate Women of Color in science, technology, engineering, and mathematics, *Harvard Educ. Rev*., **81**(2), 172-208. DOI:10.17763/haer.81.2.t022245n7x4752v2

Overton T. L., Byers B. and Seery M. K., (2009), Context and problem-based learning in higher level chemistry education, In *Innovative Methods of Teaching and Learning Chemistry in Higher Education,* Royal Society of Chemistry: Cambridge, pp 43-59.

Owo W. J. and Ikwut E. F., (2015), Relationship between metacognition, attitude and achievement of secondary school chemistry students in Port Harcourt, River State. *IOSR – J. Res. Methods,* **5** (6), 6-12. e-ISSN: 2320–7388

Pekrun R., (2006), The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice, *Educ. Psychol. Rev.,* **18***,* 315-341. DOI:10.1007/s10648-006-9029-9

Pekrun R., Hall N. C., Goetz T. and Perry R. P., (2014), Boredom and academic achievement: Testing a model of reciprocal causation, *J. Educ. Psychol.,* **106**, 696-710. DOI:10.1037/a0036006

Pekrun R., Maier M. A. and Elliot A. J., (2009), Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance, *J. Educ. Psychol.,* **101**(1), 115-135. DOI: 10.1037/a0013383

Pinder J. P. and Blackwell E. L., (2013), The "Black girl turn" in research on gender, race and science education: Toward exploring and understanding the early experiences of Black females in science, a literature review, *J. Afr. Am. Stud.,* **18**, 63-71. DOI: 10.1007/s12111-013-9255-4

Richards-Babb M., Curtis R., Georgieva Z. and Penn J. H., (2015), Student perceptions of online homework use for formative assessment of learning organic chemistry, *J. Chem. Educ.,* **92***,* 1813-1819. DOI: 10.1021/acs.jchemed.5b00294

Rowe M. B., (1983), Getting chemistry off the killer course list, *J. Chem. Educ*., **60**, 954-956. DOI: 10.1021/ed060p954

Sass D., (2011), Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework, *J. Psychoeduc.l Assess.*. **29**(4), 347-363. DOI:10.1177/0734282911406661

Seymour E. and Hewitt N., (1997), *Talking About Leaving: Why Undergraduates Leave the Science,*. Westview Press: Boulder, CO.

Simon R. A., Aulls M. W., Dedic H., Hubbard K. and Hall N. C., (2015), Exploring student persistence in STEM programs: A motivational model, *Can. J. Educ.,* **38**(1), 1-27. https://www.jstor.org/stable/10.2307/canajeducrevucan.38.1.09

Thompson M. S. and Green S. B., (2013), Evaluating between-group differences in latent variable means, In *Structural Equation Modeling: A Second Course,* Hancock, G. R., Mueller, R. O. (Eds.), Information Age, Greenwich, pp 163-218.

Tien L. T., Roth V. and Kampmeier J. A., (2002), Implementation of a peer-led team learning instructional approach in an undergraduate organic chemistry course, *J. Res. Sci. Teach.,* **39**(7), 606-632. DOI: 10.1002/tea.10038

Ültay N. and Çalik M. A., (2012), Thematic review of studies into the effectiveness of context-based chemistry curricula, *J. Sci. Educ. Technol.,* **21***,* 686-701. DOI: 10.1007/s10956-011-9357-5

Villafañe S. M. and Lewis J. E., (2016), Exploring a measure of science attitude for different groups of students enrolled in introductory college chemistry, *Chem. Educ. Res. Pract.,* **17,** 731-742. DOI: 10.1039/C5RP00185D

Villafañe S. M., Xu, X. and Raker J. R., (2016), Self-efficacy and academic performance in first-semester organic chemistry: Testing a model of reciprocal causation, *Chem. Educ. Res. Pract.,* **17**, 973-984. DOI: 10.1039/C6RP00119J

Wang J. and Wang X., (2012), *Structural Equation Modeling: Applications Using Mplus*, Wiley: Chichester, West Sussex, UK.

Wicherts J. M. and Dolan C. V., (2010), Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities, *Educ. Meas-Issues Pract.*. **29**(3), 39-47. DOI: 10.1111/j.1745-3992.2010.00182.x

Wyer M., (2003), Intending to stay: Images of scientists, attitudes toward women, and gender as influences on persistence among science and engineering majors, *J. Women Minor. Sci. Eng.,* **9**(1), 10-26. DOI: 10.1615/JWomenMinorScienEng.v9.i1.10

Xu X. (2010), *Refinement of a Chemistry Attitude Measure for College Students*, Dissertation, University of South Florida, Tampa, FL.

Xu X. and Lewis J., (2011), Refinement of a chemistry attitude measure for college students, *J. Chem. Educ.,* **88,** 561-568. DOI: 10.1021/ed900071q

Xu X., Alhoosani K., Southam D. and Lewis J., (2015), Gathering psychometric evidence for ASCIv2 to support cross-cultural attitudinal studies for college chemistry programs, In *Affective Dimensions in Chemistry*, Springer-Verlag Berlin Heidelberg, pp 177-194.

Xu X., Kim E. S. and Lewis, J. E., (2016), Sex difference in spatial ability for college students and exploration of measurement invariance, *Learn. Individ. Differ.,* **45**, 176-184. DOI:10.1016/j.lindif.2015.11.015

Xu X., Southam D. and Lewis J., (2012), Attitude towards the subject of chemistry in Australia: An ALIUS and POGIL collaboration to promote cross-national comparisons, *Aus. J. Educ. Chem.,* **72**, 32-36. ISSN: 14459698

Xu,X., Villafañe S. M. and Lewis J. E., (2013), College students' attitudes toward chemistry, conceptual knowledge and achievement: Structural equation model analysis, *Chem. Educ. Res. Pract.,* **14,** 188-200. DOI: 10.1039/C3RP20170H

Zoller U., (1990), Students' misunderstandings and misconceptions in college freshman chemistry (general and organic), *J. Res. Sci. Teach.,* **27**(10), 1053-1065. DOI:10.1002/tea.3660271011

CHAPTER 4:

ADDRESSING DIVERSITY AND SOCIAL INCLUSION THROUGH GROUP COMPARISONS: A PRIMER ON MEASUREMENT INVARIANCE TESTING

Note to Reader

This chapter is a published manuscript in *Chemistry Education Research and Practice.* The chapter was reproduced from:

Rocabado G. A., Komperda R., Lewis J. E. and Barbera J., (2020), Addressing diversity and social inclusion through group comparisons: A primer on measurement invariance testing, *Chem. Educ. Res. Pract.,* **21,** 969-988. DOI:10.1039/D0RP00025F

with permission from the Royal Society of Chemistry. Further copyright information can be found in Appendix B.

Introduction

Diversity and social inclusion are popular terms in science education at present. In the past few decades, numerous research endeavors have focused on studying diverse populations of students within science, technology, engineering, and mathematics (STEM; e.g., Hong and Page,

2004; Tsui, 2007; Hurtado *et al*., 2010). Due to a directive to increase minority representation in STEM fields in the United States (Seadler, 2012), colleges and universities have launched initiatives to attract underrepresented minority (URM) students. These initiatives can help to initially increase diversity representation; however, simply admitting students is not enough if they feel unvalued or unwelcome in their college communities (Puritty *et al.,* 2017). Thus, diversity initiatives may fail to retain these students without attention to creating inclusive environments where students of all backgrounds feel they have a voice and that they matter (Puritty *et al.,* 2017). Attaining a diverse STEM workforce, then, means promoting social inclusion and social justice in our classrooms and in our research (O'Shea *et al.,* 2016).

Critical Race Theory (CRT) has become a central framework to study issues of inclusion and social justice, particularly for members of marginalized racial groups (Crenshaw, 1995; Solórzano, 1997, 1998; Delgado and Stefanic, 2001; Yosso, 2005; Dixson and Anderson, 2018). Although CRT was born in the legal realm, it has permeated the educational field as well (Crenshaw, 1995; Delgado and Stefanic, 2001). This theory has been linked to five guiding tenets that inform research, curriculum, pedagogy, and policy (Solórzano, 1997; Yosso, 2005). Three of these tenets seem particularly well suited to investigations utilizing quantitative methodology. First, an acknowledgment of the centrality of race and racism in the power relations that underpin society requires that race be explicitly considered rather than ignored in educational research. Second, the de facto existence of 'dominant ideology' informed by race and racism requires us to cast aside naive beliefs that research and researchers are neutral and objective (Yosso, 2005) and work to safeguard against systemic biases and the propagation of social inequities in educational research (García, López, and Vélez, 2017; Gillborn *et al*., 2018). And third, answering CRT's call

for a commitment to social justice requires us to privilege research that works to uncover social inequities and moves toward the eradication of racial and other forms of marginalization (Solórzano, 1997). CRT is a framework well equipped to investigate issues of racism and social inequities in educational settings at the individual as well as at the institutional level. For example, Fernández (2002) uses CRT as a framework and takes an individual approach to display a successful educational experience of one immigrant Latino student in a public school in Chicago via qualitative methods. On the other hand, Solórzano and Ornelas (2004) use CRT to investigate the access and availability of Advanced Placement (AP) courses in California high schools and how they affect African American and Latina/o students' admission to college. This quantitative study exhibits an institutional approach that documents cumulative impacts on individuals and groups of students from minority racial and ethnic populations. Likewise, CRT and quantitative methods can be utilized at the institutional level to investigate achievement gaps in educational systems, providing a wider lens for these investigations (García, López, and Vélez 2017; López *et al.,* 2018), rather than merely grade comparisons. Whenever possible, studies of this nature benefit from a comprehensive investigation with appropriate categories for investigating achievement gaps, such as race-gender-class intersections (Crenshaw, 1989; Covarrubias, 2011, 2013; Litzler, Samuelson and Lorah, 2014; García, López, and Vélez, 2017; Ireland *et al.,* 2018; López *et al.,* 2018) as a movement to achieve a more complete view of the investigation and avoid reproduction of widespread inequities in educational settings (García, López, and Vélez 2017; Gillborn *et al.,* 2018).

In an effort to combat against racism and other societal inequities, these issues have long been studied with qualitative methodologies (Gillborn *et al.,* 2017; García, López and Vélez,

2018). Quantitative methods have been criticized for an inability to speak to the details of lived experiences of diverse populations (García, López, and Vélez, 2018) and thus been deemed inappropriate to study these issues in educational settings due to these everyday experiences having deep roots in social relationships (Apple, 2001). Although qualitative methods are more appropriate to capture nuances of societal processes as experienced by individuals, quantitative methods can explore wider structures in which individual and collective experiences are lived, revealing wider structural issues that affect these diverse groups on a larger scale (Gillborn *et al.,* 2017). With this tension between qualitative and quantitative methodologies attending to issues of social inequities, we encourage the use of either or both types of methods when appropriate, following the tenets of CRT. Therefore, in an effort to promote inclusion and equity in our classrooms, appropriate qualitative and quantitative methods can be used in research, with the premise that our methods must be reflexive and safeguarded against systemic racial, ethnic, gender, and other biases favoring the majority groups (Gillborn *et al.,* 2017).

Much of the critique about using quantitative methods to investigate these issues comes from the problem that numbers are positioned as 'neutral' and audiences may believe 'data speaks for itself.' Critical theorists argue that these claims of neutrality are far from the truth (Gillborn *et al.,* 2017). However, researchers, practitioners, and policy-makers tend to put great emphasis in numbers, as these are the data by which policies are justified and schools and districts are labeled successes or failures (Gillborn *et al.,* 2017). Thus, to rise above these critiques in favor of continuing to use quantitative approaches to investigate social inequities, a process of ongoing self-reflexivity and engagement with historical, social, and political structures of the groups under investigation must be present (García, López and Vélez, 2018). Additionally, because numbers

carry such important consequences, we must use them with caution and systematically interrogate the validity of the inferences we make with these numbers, particularly as it relates to consequential validity (AERA, NCME and APA, 2014). According to Messick (1995) the social consequences of score interpretation may be positive or negative, intentional or unintentional. Thus, in the interest of advancing inclusion and social justice, researchers must engage in collecting evidence of positive consequences while minimizing adverse effects. As an example of unintentional, negative effect, one could imagine that a subgroup of students misinterpret items on an assessment instrument based on unfamiliar words in the item, which may lead to confounding results in the data for that subgroup. This source of invalidity can potentially lead to erroneous decisions that may have adverse consequences for this subgroup of students (Shephard, 1993; Messick, 1995). Therefore, raising the bar for quantitative methods in our field will require taking steps to safeguard against consequential validity threats that may be present when making group comparisons.

Quantitative Standards for Group Comparisons in CER

In CER, investigations of efforts to broaden participation of diverse student populations have been a focus of multiple studies (i.e., Richards-Babb and Jackson, 2011; Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Nawarathne, 2019; Shortlidge *et al.,* 2019). Many of these studies have aimed to investigate differential outcomes of URM students by performing group comparisons with various statistical analyses (Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Shortlidge *et al.,* 2019). For instance, Fink and colleagues (2018) proposed a strategy to promote improved general chemistry performance for women and minorities through a growth

mindset intervention. The results of the study report higher performance overall favoring the White students; however, post-hoc Tukey tests confirmed an intervention effect for minority students, who ultimately earned more than 5 percentage points higher on average in the mindset intervention condition (Fink *et al.,* 2018). Similarly, Stanich and colleagues (2018) implemented a supplementary instruction (SI) course that aimed to narrow achievement gaps by showing that URM students who participated in the SI course had lower failure rates in general chemistry than URM students who did not take the course. Additionally, this study also aimed to narrow affect gaps by increasing perception of relevance, sense of belonging, and emotional satisfaction toward the subject of chemistry (Stanich *et al.,* 2018). While studies such as these are a positive sign that diversity and social inclusion are being taken seriously, there is still work to be done with respect to developing guidelines for quantitative research on these issues.

The next important step in developing research standards is to critically examine the collection, analysis, and representation of quantitative data and results for threats to the validity of inferences when group comparisons are to be made. CER has a long history of assessment design to probe student understanding of concepts taught in the classroom (i.e., Tobin and Capie, 1981; Roadrangka, Yeany and Padilla, 1983; Loertscher, 2010; Villafañe, *et al.,* 2011; Kendhammer, Holme and Murphy, 2013; Wren and Barbera, 2013; Brandriet and Bretz, 2014; Bretz, 2014; Kendhammer and Murphy, 2014; Xu, Kim and Lewis, 2016). These, and other, assessment instruments have been used by researchers and practitioners to evaluate the success of classroom interventions and curricular changes. Furthermore, in the last few decades, CER as a field has moved toward an increased interest in affect and motivation in educational settings (Xu, Villafañe and Lewis, 2013; Ferrell and Barbera, 2015; Salta and Koulougliotis, 2015; Ferrell, Phillips and

Barbera, 2016; Liu *et al.,* 2017; Gibbons and Raker, 2018; Gibbons, *et al.,* 2018; Hensen and Barbera, 2019; Rocabado *et al.,* 2019). Thus, assessment instruments may be used in CER to determine research agendas, report findings, evaluate interventions or curricular design and much more.

Given the current interest in measuring affect in the classroom, there is an added concern that many cognitive and emotional factors might have different effects among diverse populations, particularly disfavoring URM groups (Ceci, Williams and Barnett, 2009; Villafañe, García and Lewis, 2014; Rocabado, *et al.,* 2019). However, some of the differences noted in these data could be an artifact of the assessment instrument (Jiang, García and Lewis, 2010); thereby resulting in a potential threat to the validity of the inferences drawn from the instrument-derived data (Arjoon, Xu and Lewis, 2013; AERA, APA and NCME, 2014). Therefore, in the interest of promoting social inclusion in the classroom, it is important to know that when an instrument functions well for the whole class, the functionality extends to any subgroups of interest. Nevertheless, simply comparing observed scores for subgroups is not appropriate. As shown by several studies (Khaveci, 2015; Komperda, Hosbein and Barbera, 2018; Montes, Ferreira and Rodriguez, 2018), differences might arise as artifacts of the instrument functioning and not as differences in understanding, ability, or affect.

Goals of This Measurement Invariance Testing Primer

To encourage and support the gathering of evidence to substantiate group comparisons within CER, this manuscript presents the quantitative method of measurement invariance testing for those familiar with factor analysis. A comprehensive review of measurement invariance testing can be found in Vandenberg and Lance (2000). Measurement invariance testing can be used to investigate the degree to which measured student data is represented by the same theoretical model. Prior to introducing the details and meanings of the various levels of measurement invariance testing, we discuss latent variables and data visualization techniques. This introduction provides initial insight into the relations among assessment items as well as providing a basis for understanding the mathematical foundations being tested. We then provide a step-by-step tutorial of measurement invariance testing, discussing what is being tested, how to evaluate if invariance has been achieved, and what (if any) comparisons between groups are supported at each step. Finally, we present a summary of the implications of measurement invariance testing as well as recommendations for researchers, practitioners, reviewers, and journal editors.

Group Comparisons on Latent Constructs

Commonly, the variables of interest in CER are ones that cannot be measured directly, i.e., they are latent traits. Variables such as student self-efficacy, attitude, metacognition, mindset, and understanding of chemistry are all examples of latent traits. Many of these latent traits are multidimensional, that is, they are subdivided into smaller latent units (subconstructs or factors)

that make up the latent trait (Brown 2006 pp.2). To provide an example for our discussion of quantitative data comparison by group, we devised a *fictitious* assessment instrument to measure the latent trait of 'perceived relevance' toward chemistry. Such an instrument might be useful in understanding college students' perceptions of the field of chemistry. For this *fictitious* assessment instrument, it could be expected that students' perceived relevance of chemistry might differ by college major and that a researcher might want to compare data from this instrument by group. While many times the groupings of students we quantitatively investigate are by gender or URM status, these are not the only groupings for which comparisons need to be supported by evidence. For example, with our *fictitious* instrument the comparison groups could be defined as STEM and non-STEM majors. Other groupings could be first-generation college students or community college transfer students for comparison to students not in these groupings. Whatever the chosen comparison groups are, it is imperative that researchers have a directive to investigate those groups and use an appropriate construct for the comparison.

It is important to note that utilizing assessment instruments that have been developed with a strong theoretical background and which have been investigated for forms of validity and reliability evidence delineated by the *Standards for Educational and Psychological Testing* (Arjoon, Xu and Lewis, 2013; AERA, APA and NCME, 2014) is imperative to drawing meaningful insights from studies. Following with the example, and assuming that the instrument was created under these conditions, our *fictitious* assessment instrument is called the Perceived Relevance of Chemistry Questionnaire (PRCQ) and contains three *fictitious* subconstructs: Importance of Chemistry (IC), Connectedness of Chemistry (CC), and Applications of Chemistry (AC). The *fictitious* PRCQ is a 12-item instrument with four items per subconstruct. When student

responses to these 12 items are examined, the expected pattern of bivariate correlations among responses would be that items aligned with the same subconstruct should have stronger correlations with each other, meaning they are highly associated with each other through an underlying subconstruct, and have weaker correlations with other items aligned with different subconstructs. For comparison purposes then, these item-level patterns need to be consistent within each group.

Group Comparisons Through Data Visualization

In addition to using descriptive statistics to investigate data patterns, item-level data can be visually inspected using a variety of methods (e.g., box-plots, violin plots, graphs, charts). To demonstrate ways in which to visualize data, we have created simulated PRCQ datasets that highlight several different data patterns across groups (see Appendix C for additional details). Item correlation values for one of these datasets are plotted in a correlation heatmap shown in Figure 4.1. In this correlation plot the item labels (i.e., I1, I2, etc.) are listed on the diagonal, and the color of each square represents the value for the correlation (i.e., the strength of association) between two items. Pairs of items with stronger correlations are represented with darker squares and pairs of items with weaker correlations are represented with lighter squares. The simulated data used in this example are strongly correlated in four-item sets (I1 to I4, I5 to I8, and I9 to I12); items outside these sets (e.g., I1 and I8) are weakly correlated. As the PRCQ has three subconstructs, another way to represent the relations between the twelve items is with a factor diagram. The intended factor diagram for the 12-item PRCQ instrument has been added above the correlation plot. In a

factor diagram each individual item (called an indicator item and represented by a square) is associated with a subconstruct or factor (represented by a circle). Together, these visual representations of the PRCQ data provide initial visual evidence for the presence of the intended factors (i.e., item set groupings).

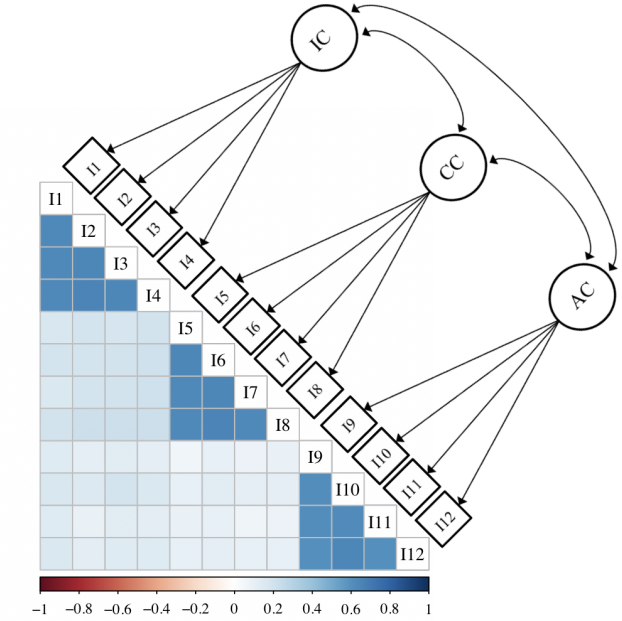


Figure 4.1. A visualization of the lower correlation matrix for the 12-item PRCQ instrument with a factor model overlaid to illustrate how correlations between sets of items implies the presence of an underlying factor structure. We note that, although the covariance matrix is more directly applicable, the correlation matrix is a standardized covariance matrix, and therefore easier to visualize and discuss.

When making measurements that will ultimately be used to compare the outcomes of various groups on an underlying construct (i.e., Importance of Chemistry (IC), Connectedness of Chemistry (CC), and Applications of Chemistry (AC)), it is necessary to provide evidence that the PRCQ instrument is functioning in a similar way for each group being compared. This practice is a way in which the field of CER can meet best practices when making comparisons and provide

evidence to support that any differences between the groups' data are due to true differences in the construct, not a result of systematic bias in the measurement of the construct (Gregorich 2006; Sass 2011). Using our example, as researchers we could be interested in measuring potential differences in the perceived relevance of chemistry (as measured by the PRCQ) between groups. As lower-level chemistry courses serve a range of majors, we could investigate potential differences in perceived relevance between STEM and non-STEM majors, or among multiple groups such as White, African-American, Asian, and Hispanic students. For simplicity in our example, we have simulated response data for a two-group comparison, which will help us visualize the discussion that will proceed. In addition, the data we have simulated is continuous. However, we do understand that much of the data generated in CER is categorical in nature and as such will necessitate a different set of considerations. Thus, we provide explanation and analyses for both continuous and categorical data, in the electronic supplementary information (ESI), along with code (in R and Mplus) for generating the data visualizations as well as the additional analysis steps described later in this manuscript.

If the aggregated PRCQ data in Figure 1 were divided by STEM and non-STEM majors, one step towards examining consistent functioning across groups would be to see if the two groups have similar correlation plots. As shown in Figure 4.2, when visually comparing the correlation plots by group, it can be seen that they are essentially identical. Ways of testing this similarity statistically will be discussed later.

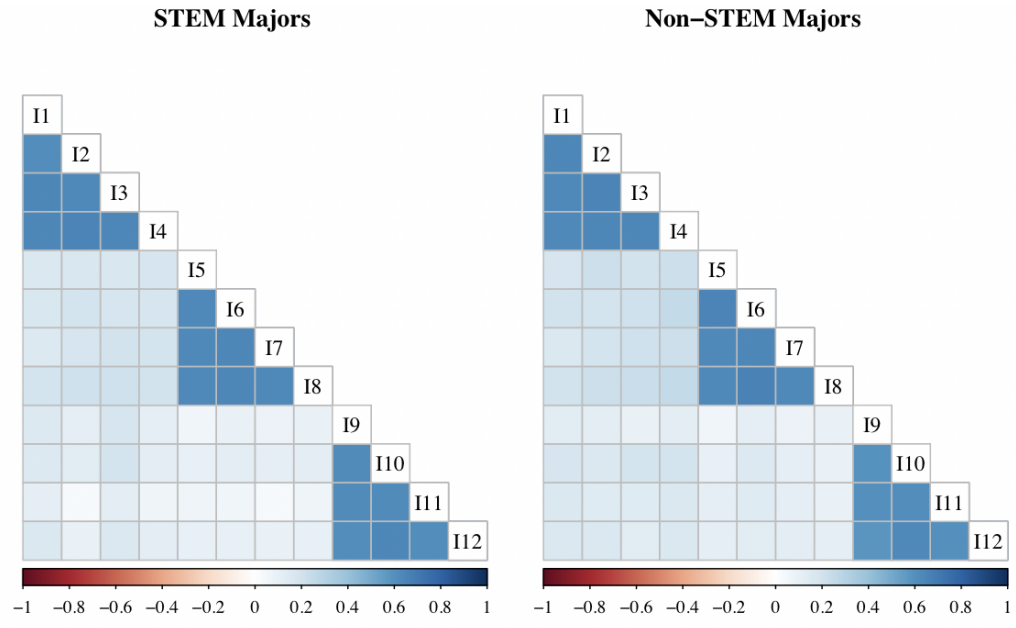


Figure 4.2. Correlation plots for 12 items with similar strength of association for each item and its intended factor for two subgroups (STEM majors and non-STEM majors) within the data set.

While the situation represented in Figure 4.2 is the best possible outcome (i.e., the data are simulated to align with a known factor structure for both groups), it is not always the case that data from students in different groups will show the same strength of association between each item and each intended factor. An example of such a situation is visualized in Figure 4.3 where we simulated a difference in strength of association for one item in one group. In this aggregated PRCQ data set (Figure 4.3a) we can see inconsistencies around I10, where some correlation boxes are lighter. Although, the overall correlation pattern is consistent (i.e., an instrument that measures three distinct factors as hypothesized for the PRCQ), when we disaggregate the data and view the correlation matrix for each group separately, we observe that I10 has a much lower association with the AC factor for non-STEM majors (Figure 4.3c) compared to STEM Majors (Figure 4.3b).

This group difference would not be obvious when looking at the correlations in the aggregated dataset (Figure 4.3a). The situation represented here, dissimilar associations between items and factors across groups, implies that the item is not functioning in similar ways for each group, which could be due to differences in item interpretation for I10. Regardless of the underlying reason, which may never be known for sure, this situation indicates a possible threat to the validity of the potential inferences from the data and needs to be examined more closely to determine whether the data can still be used to compare the groups.

Another type of measurement difference that could occur between the groups is that an item may not have similar response averages in each group. In the next set of simulated data, the strength of association between all items and their intended factor is equivalent, as in Figure 4.3, but the average response for I3 has been modified for the STEM majors group to illustrate this issue. Unlike when the strength of association differed in the previous example, this result is more obviously seen when visualizing the correlations in the aggregated dataset (Figure 4.4a) than in the disaggregated sets (Figures 4.4b and 4.4c).

To further visualize the distribution of values for each item within each group, Figure 4.5 plots the means for each item in the two groups using a boxplot. It can be clearly seen that the distribution for I3 in the STEM majors group is much different and is shifted to the higher end of the scale. This outcome could occur because there are true differences between the groups or it could be due to improper item functioning for one group. However, a quantitative analysis does not differentiate between these two reasons, thus it is appropriate to further investigate the item functioning when this occurs.

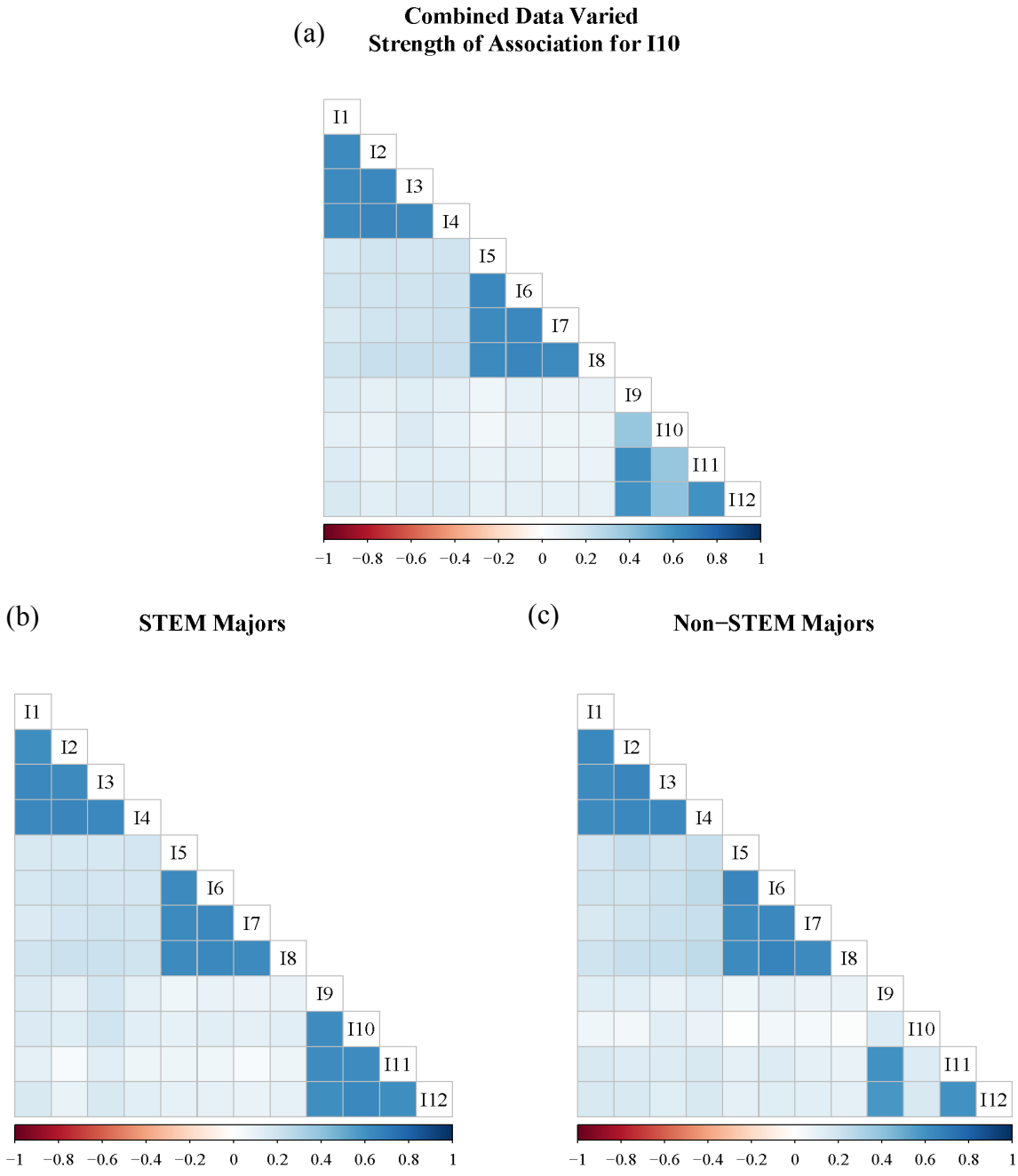


Figure 4.3. (a) Correlation plot for 12 items with combined dataset; (b) Correlation plot with STEM major data; (c) Correlation plot with non-STEM major data with I10 correlation lowered.

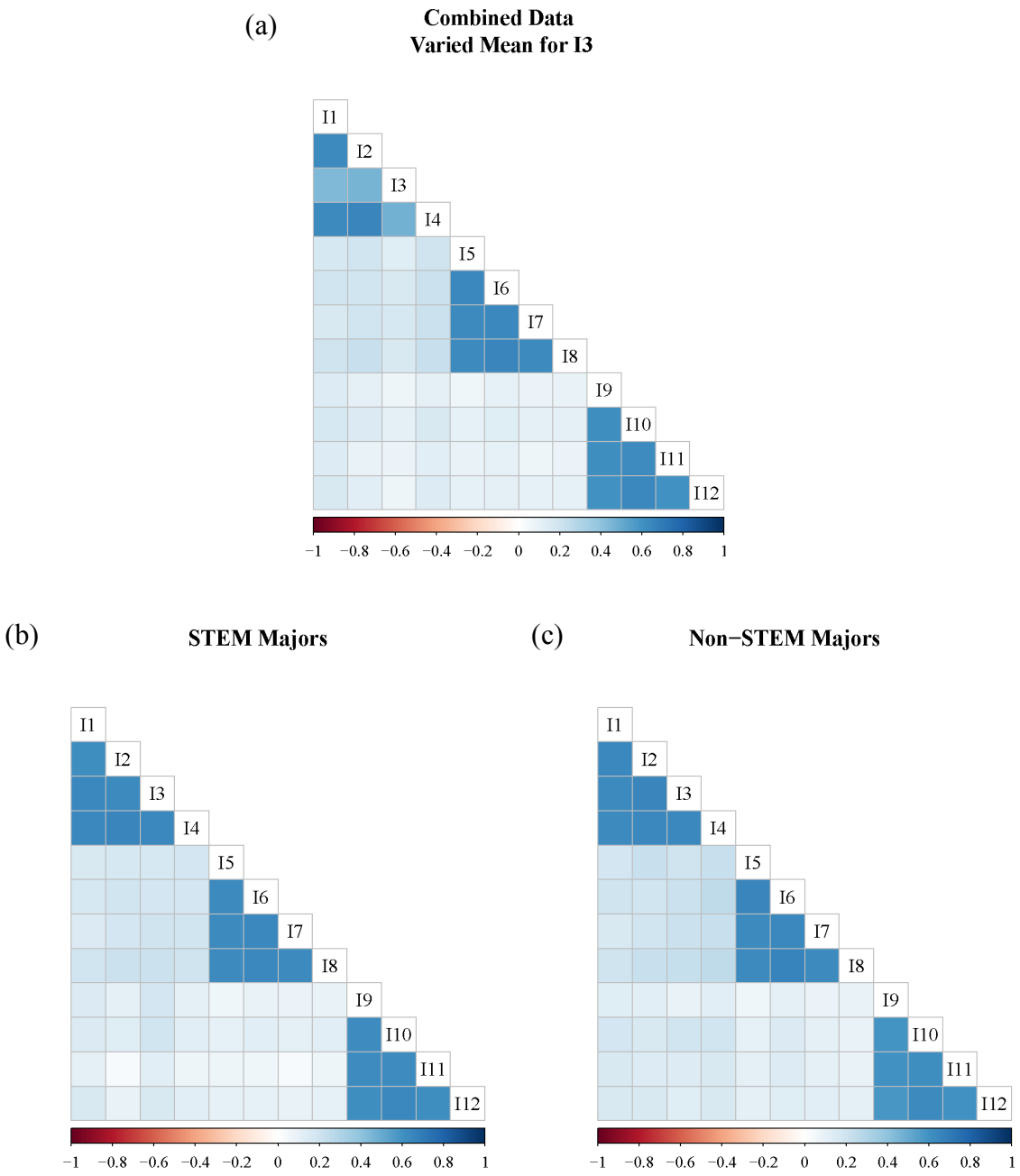


Figure 4.4. (a) Correlation plot for 12 items with combined dataset; (b) Correlation plot of STEM majors with mean of I3 raised; (c) Correlation plot of non-STEM majors.

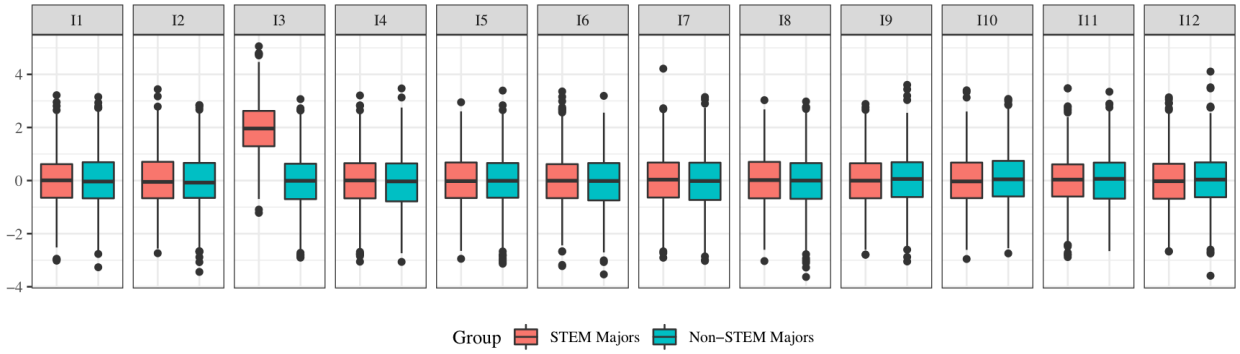


Figure 4.5. Boxplot of item means for each group.

The item-level differences noted in Figures 4.3-4.5 may be due to a variety of issues, which would be worth exploring further in order to understand why they occur. However, in considering whether the data can still be used to make comparisons between groups, the degree to which these differences impact the proposed factor structure need to be evaluated using measurement invariance testing. This quantitative method would indicate if the differences pose a potential issue with how the instrument functions for the different groups, potentially limiting the ability to draw valid conclusions about how the underlying factors of interest differ across groups.

Data Considerations Prior to Performing Measurement Invariance Testing

While we have emphasized the importance of visualizing data and have shown various ways it can be useful, we acknowledge that data visualization is insufficient to address the degree to which item-level differences may impact group comparisons, which necessitates more robust investigations using statistical tests. Additionally, and more often than not, many data issues are not easily visualized, but can become evident in statistical analyses. We encourage all researchers

to visualize their data and compute descriptive statistics, thereby providing initial insights to the data as well as evidence about the characteristics of the data. Understanding data characteristics will aid the researcher in making other decisions about further analyses, such as which tests are appropriate to run or which estimator is appropriate to use when modeling data.

Different types of data such as categorical or continuous, can be analyzed with measurement invariance testing utilizing appropriate estimators for each type of data. For example, ordinal data (e.g., categorical data from items with a 7-point Likert-type scale) with variance ranging the entire scale is often treated as continuous data and can be estimated with a maximum likelihood estimator (Muthén and Muthén, 2010; Hirschfeld and von Brachel, 2014). On the other hand, categorical data (e.g., data from a 'yes or no' type item or items using fewer than 5 response scale categories) are more appropriately analyzed using a weighted least squares estimator (Muthén and Muthén, 2010; Hirschfeld and von Brachel, 2014; Bowen and Masa, 2015). Ensuring the proper estimator for the data-type is of utmost importance. Violations of normality, independence, and homogeneity are also important to note, and should be handled appropriately. Discussion of estimators and assumptions is beyond the scope of this article; however, we provide a few resource references for interested readers here (Stevens, 2007; Garson, 2012) and in Appendix C.

An additional consideration before conducting measurement invariance testing is statistical power (Hancock and French, 2013). To conduct meaningful statistical analyses, one must ensure an appropriate sample size in order to have enough power to draw meaningful inferences. In measurement invariance testing the interest is in finding no evidence of significant difference

between groups, thus, an inappropriate sample size (i.e., too small) can increase the chances of type II error through failing to reject the null hypothesis (of equivalence) when it should have been rejected (Lieber, 1990; Counsell, Cribbie, and Flora, 2019). Recently, work has been done indicating that sample size requirements can be estimated given the number and value of parameters being estimated (Wolf *et al.,* 2013; Mueller and Hancock, 2019).

Confirmatory Factor Analysis Framework

In the previous section we explored visual methods for detecting potential validity threats in our PRCQ data. Though visualizing is an important initial step, more formal statistical methods can and should be employed to evaluate the degree to which differences pose threats to the validity of comparisons. Methods such as Differential Item Functioning have been used to investigate item-level threats in CER (Kendhammer, Holme and Murphy, 2013; Kendhammer and Murphy, 2014), however, the purpose of this paper is to explore threats at the construct, or latent variable level. At this level, various frameworks can be used, including Item Response Theory (IRT; Candell and Drasgow, 1988; Mellenbergh, 1989) and factor analysis (Brown, 2006). As factor analysis methods have become commonplace within CER, and IRT is less frequently utilized in our field, this discussion will focus only on evaluating measurement invariance in a factor analysis framework.

Within a Confirmatory Factor Analysis (CFA) framework (Brown, 2006), measurement invariance testing is a technique that can be used to support that the internal structure of an

assessment instrument holds for different groups people at one time point (Salta and Koulougliotis, 2015; Bunce *et al.,* 2017; Hensen and Barbera, 2019; Rocabado *et al.,* 2019) or over time in longitudinal studies (Keefer, Holden and Parker, 2013; Hosbein and Barbera, 2019; Rocabado *et al.,* 2019). In the previous section, the idea of internal structure was described in terms of the grouping of items with each other to form an underlying factor of interest (as introduced in Figure 4.1). In this section, these associations will be defined more formally using the language of factor analysis.

The CFA framework operates under a network of equations, among which, regression equations link items to latent variables (Brown, 2006). Regression or linear equations (see Equation 4.1) have several components: a dependent (predicted) variable (y), an independent (predictor) variable (x), the slope of the line (m), the intercept (b), and the measurement error (e).

$$y = mx + b + e \qquad \text{[Eq.4.1]}$$

Translating the regression equation to the language of factor analysis, the predicted variables are the observed variables (i.e., items), the predictor variables are the factors or latent variables, and the slope is the factor loading. In Figure 4.6a we write out the regression equation for an item from the PRCQ and in Figure 4.6b display the model that underlays the PRCQ using common statistical notations in the CFA framework, which we will use for the remainder of the discussion in this manuscript. In this 12-item (i.e., I1-I12), 3-factor (i.e., IC, CC, AC), model lower-case lambdas (λ) represent the factor loading of each item to its respective factor, lower-

case taus (τ) represent the intercept of an item, and lower-case epsilons (ε) represent the measurement error of an item. In addition to these parameters, Figure 6b shows the covariance between factors (e.g., double headed arrow between IC and CC) and each individual factor variance (e.g., small curved arrow from IC to IC). While these parameters are part of the overall CFA model for the PRCQ, they do not need to be modified when evaluating for measurement invariance.

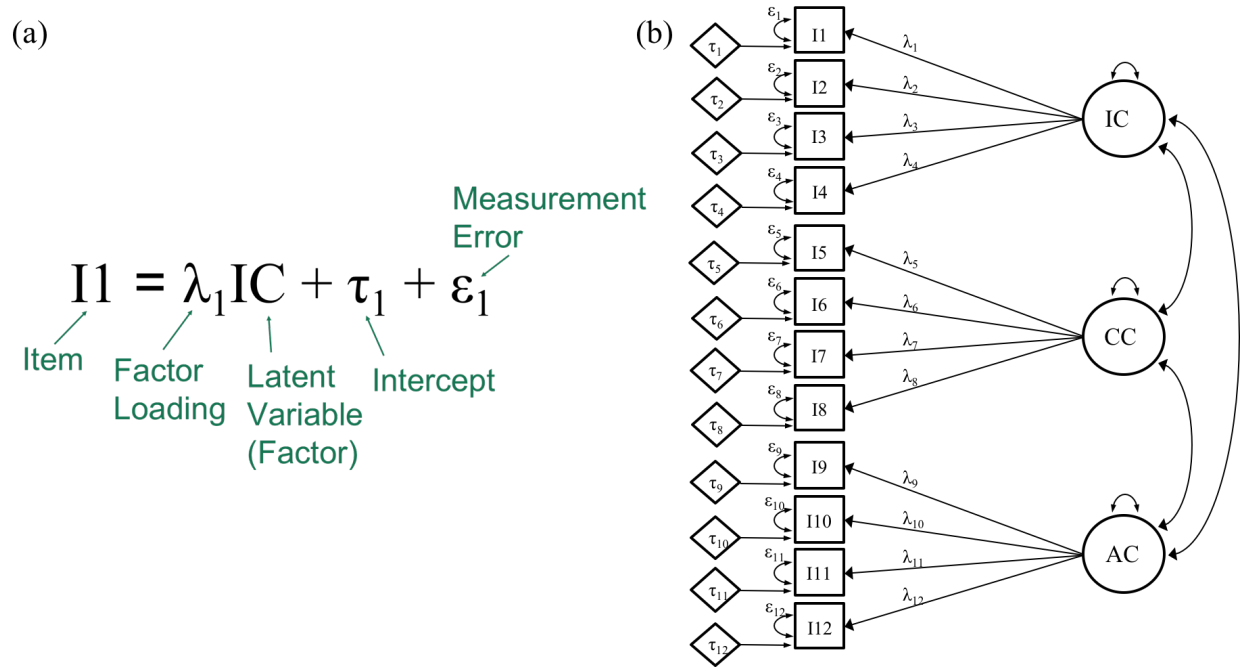


Figure 4.6. (a) Representation of equation components in CFA. Linear equation for I1 and the IC factor with notation and labels corresponding to CFA framework. (b) Factor model displaying the factor analysis notation of the relation between items and their corresponding factors.

Measurement invariance testing within a CFA framework investigates the extent to which the network of equations in a model is similar across group-level data. Therefore, each part of the equation (Eq. 4.1), for each item is tested for evidence of significant differences across groups,

starting with the slopes (loadings), then the intercepts, followed by the measurement error variances. At each stage of measurement invariance testing, evaluation of overall data-model fit occurs.

Data-Model Fit and Fit Indices

The primary goal of measurement invariance testing is to examine how well the data collected fit a proposed model of relations among items and factors as described by a set of regression equations. Continuing with the example, we investigate the PRCQ data (by STEM and non-STEM groupings) for item associations based on the proposed (*a priori*) three-factor model for the PRCQ shown in Figure 4.6b. Mapping the data to this proposed model using a maximum likelihood estimator (the default in most software packages and the one that is appropriate for our simulated continuous data), fit indices are generated and are used to evaluate how well the data fit the model. Regardless of the software package used, it is good practice to review several kinds of fit indices that fall in each of these categories: comparative fit, absolute fit, and parsimony correction. The comparative fit indices evaluate the fit of a specified model solution in relation to a baseline model solution. Absolute fit indices assess how reasonable the model fit is based on the null hypothesis that the data fit the model perfectly. Finally, the parsimony correction indices are similar to the absolute fit but include a penalty for poor model parsimony (Brown, 2006).

With these fit index descriptions in mind, we present several suggested cutoff criteria for fit indices that were simulated by Hu and Bentler (1999) using a maximum likelihood estimator.

Examples of comparative fit are: the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI), both of which have a recommended cutoff of >0.90 as acceptable, but best if >0.95 (Hu and Bentler 1999). For the absolute fit category, the Chi-square (χ^2) test statistic and the standardized root-mean square residual (SRMR) indices can be considered. The χ^2 is a descriptive index utilized to evaluate how closely the data fit the model. However, this test is highly influenced by sample size, thus additional fit indices must be considered to evaluate appropriate data-model fit (Brown, 2006). Hence the SRMR is a valuable index to add in this category and its cutoff criteria is <0.08 as acceptable (Hu and Bentler 1999). Finally, for the parsimony correction, the root mean square of approximation (RMSEA) index can be evaluated with acceptable cutoff criteria of <0.06 (Hu and Bentler 1999). Though these recommended criteria are often considered as firm cutoffs, there are known situations where the strength of the factor loadings can confound interpretation of fit indices (McNeish *et al*., 2018). Therefore, it is up to the researcher to provide as much evidence as possible to support the acceptability of a proposed factor model. It is also important to note that for categorical data a different estimator should be used, thus model fit indices and cutoff criteria are different from the ones noted here for continuous data and the maximum likelihood estimator. A more thorough description of estimator, model fit indices, and their respective cutoffs for categorical data are provided in Appendix C.

In the following section of this manuscript we present measurement invariance testing as the step-by-step evaluation of a series of nested models. Each step in the evaluation adds a constraint to test whether the groups being compared share a similar measurement model and if comparisons can be supported. Therefore, in addition to evaluating the data-model fit at each step of measurement invariance testing, we also calculate and evaluate the change in data-model fit

between nested models. Cheung and Rensvold (1999, 2002) as well as Chen (2007) conducted a series of simulation studies with continuous data to investigate data-model fit criteria, in particular the change in data-model fit at each step of measurement invariance testing. Cheung and Rensvold (2002) focus solely on evaluating the change in Chi-square ($\Delta\chi^2$) between nested models, looking for a nonsignificant value. More recent work finds this practice acceptable (Mueller and Hancock, 2019), as the idea of measurement invariance testing is to find no evidence of significant difference between the models, which provides support for group comparisons. Other researchers, such as Chen (2007) have investigated the change in other fit indices as well, to ensure that there are various indicators that provide further evidence that no significant difference between nested models is observed. Chen (2007) offers a range of values that, based on the simulation studies conducted, offer reasonable cutoff values for the fit indices we have introduced earlier in this section. These values vary by level of invariance being evaluated and therefore will be presented within the appropriate testing step below. However, simulation studies have called into question the exact cutoffs and fit indices to use in the context of invariance testing (Kang *et al*., 2016) so again the researcher must decide what evidence to present to justify interpretation of models.

Steps of Measurement Invariance Testing

In 1997 Widaman and Reise described 4 steps of measurement invariance testing: configural, metric (weak), scalar (strong), and residual (strict, also known as conservative). In this report we focus on this 4-step method, although there are other methods that utilize additional steps

when investigating whether comparisons are supported between groups (for examples see Jöreskog, 1971; Vanderberg and Lance, 2000).

Step 0: Establishing Baseline Model

A preliminary step before conducting measurement invariance testing is to conduct a separate CFA for each group dataset that will be compared. In this step, the CFA is used to investigate that each group's response patterns align with the proposed model to an acceptable level (Gregorich, 2006). The acceptability of the fit between each dataset (i.e., STEM and non-STEM groupings) and the model (Figure 4.6) is checked using the fit indices noted earlier. If the data-model fit for either group's data is deemed unacceptable at this stage, measurement invariance testing is not appropriate and comparisons between the groups would not be supported. At this point, the next step would be to conduct an investigation of the reasons for failing to achieve acceptable data-model fit. However, if the data-model fit reached acceptable criteria for each group, then beginning the measurement invariance testing steps is appropriate.

Step 1: Configural Invariance

Once the independent CFAs for each group are found to have acceptable data-model fit, the first step of measurement invariance testing can begin. In this step, the same model is estimated concurrently for each group, allowing all model parameters to be freely estimated (Gregorich,

2006; Sass, 2011; Putnick and Bornstein, 2016). The point of this unconstrained model is two-fold: 1) to investigate whether items associate with each other in similar ways in all groups (i.e., items belonging to the same factor correlate more highly with each other than to other items); and 2) to establish a baseline of data-model fit, ensuring that subsequent comparisons are conducted utilizing the same network of equations for both groups. This baseline model is called the *configural model*, as it verifies that the general structure (or configuration) of items and factors is similar across groups. Configural invariance is achieved when this model has acceptable data-model fit values (Hu and Bentler, 1999).

The models in Figure 4.7 represent the configural model, for our three-factor PRCQ instrument, for two groups. For discussion purposes, the model parameters for STEM majors (group 1) are labeled with numeric subscripts and those for non-STEM majors (group 2) are labeled with alphabetical subscripts. Take for example the relation between the first factor, IC, and the first item, I1. This relation is symbolized as λ_1 for group 1 and λ_a for group 2. In the configural model, these two relations are free to take on whichever value provides the optimal solution to the system of regression equations.

If the configural model fails to reach acceptable levels of fit, the result suggests that the factors are not associated with the same items for both groups (Gregorich, 2006; Putnick and Bornstein, 2016). Therefore, one can question whether the constructs being measured have the same meaning for these groups (Bornstein, 1995; Putnick and Bornstein, 2016). With this outcome, no further invariance testing is advised. However, we encourage researchers to conduct further investigation to find the source of noninvariance between the groups. Modes of investigation could

be quantitative in nature, such as inspection of covariance or correlation matrices similar to the visuals we provided earlier (Figures 4.1- 4.4). Investigation could also be qualitative in nature, for example conducting cognitive interviews (Willis, 1999) with respondents from both groups to explore the constructs being measured and find the root of the differences between the two groups. These practices can help to ascertain any fundamental differences in construct meaning for different groups, which can provide insight into their lived experiences and interpretation of the construct of study (Komperda, Hosbein and Barbera, 2018).

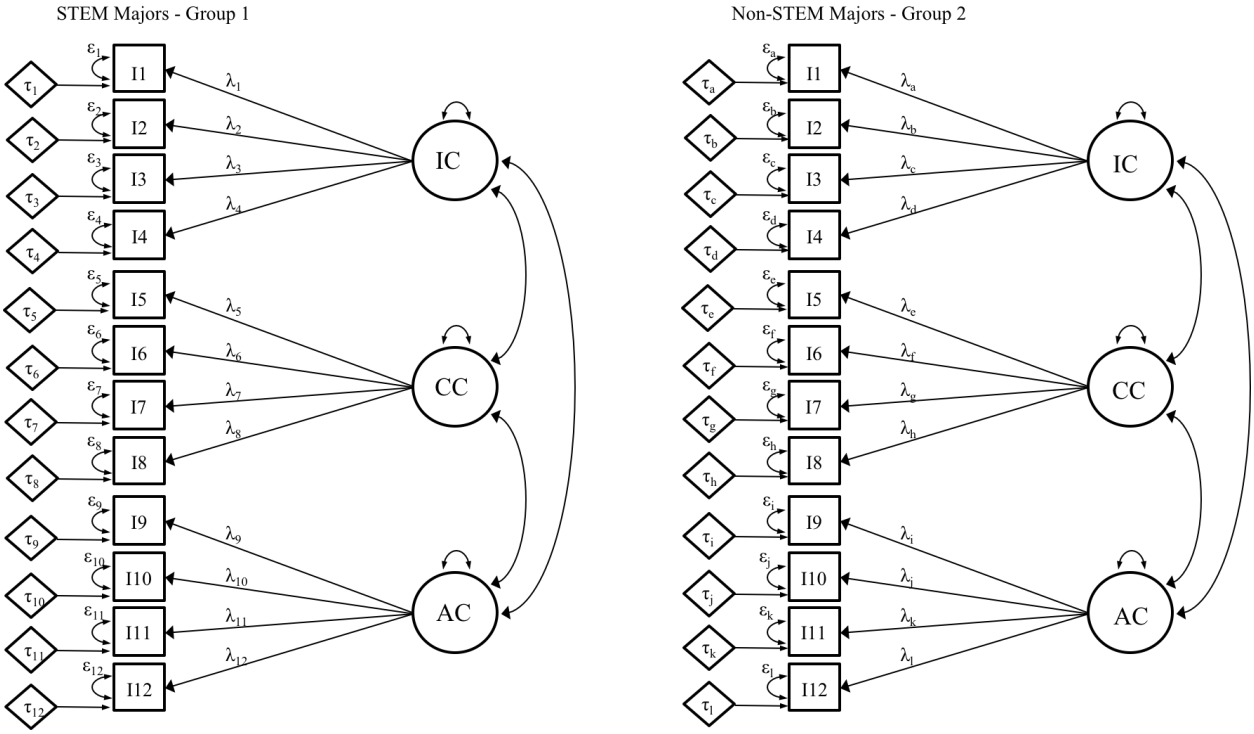


Figure 4.7. Configural invariance model where all parameters are freely estimated for two groups (STEM and non-STEM majors).

Step 2: Metric Invariance (Weak)

If a configural model (Figure 4.7) is observed to have acceptable data-model fit, the next level of establishing equality between the group-level data can be conducted. This step involves applying the first constraint to the baseline model equations, which establish the linear relationship between items (e.g., I1) and factors (e.g., IC). In the *metric model* (Figure 4.8), also called the *weak invariance model* (Meredith, 1993), the constraint of equal unstandardized slopes, or factor loadings (λ), is applied (Gregorich, 2006; Sass, 2011; Putnick and Bornstein, 2016; see Figure 4.8 where loading subscripts match across groups). That is to say that for STEM majors, the factor loadings are freely estimated, but for non-STEM majors, the loadings are set to be equal to the loadings for STEM majors. At this level of invariance testing, we are exploring whether the strength of associations between the items and the latent variables are similar across groups (Byrne, Shavelson and Muthén, 1989; Gregorich, 2006). To achieve metric invariance, first the fit statistics of the metric model (Figure 4.8) are evaluated (Hu and Bentler, 1999), and then they are compared to those of the configural model (Figure 4.7). No evidence of significant difference should be observed between the configural and metric models. To evaluate the comparison between configural and metric models, the change in fit indices between levels is established utilizing the guidelines noted earlier. It is important to note that evaluating model fit is pertinent; however, evaluating the change between the models is essential to establishing invariance between groups. Establishing metric invariance implies that the meaning of the factor (in terms of relative weight of items) is similar across groups (Gregorich, 2006). However, this evidence is not enough to make comparisons between groups. At the very least, another level of constraint is needed before group comparisons can be made, as will be summarized in subsequent steps.

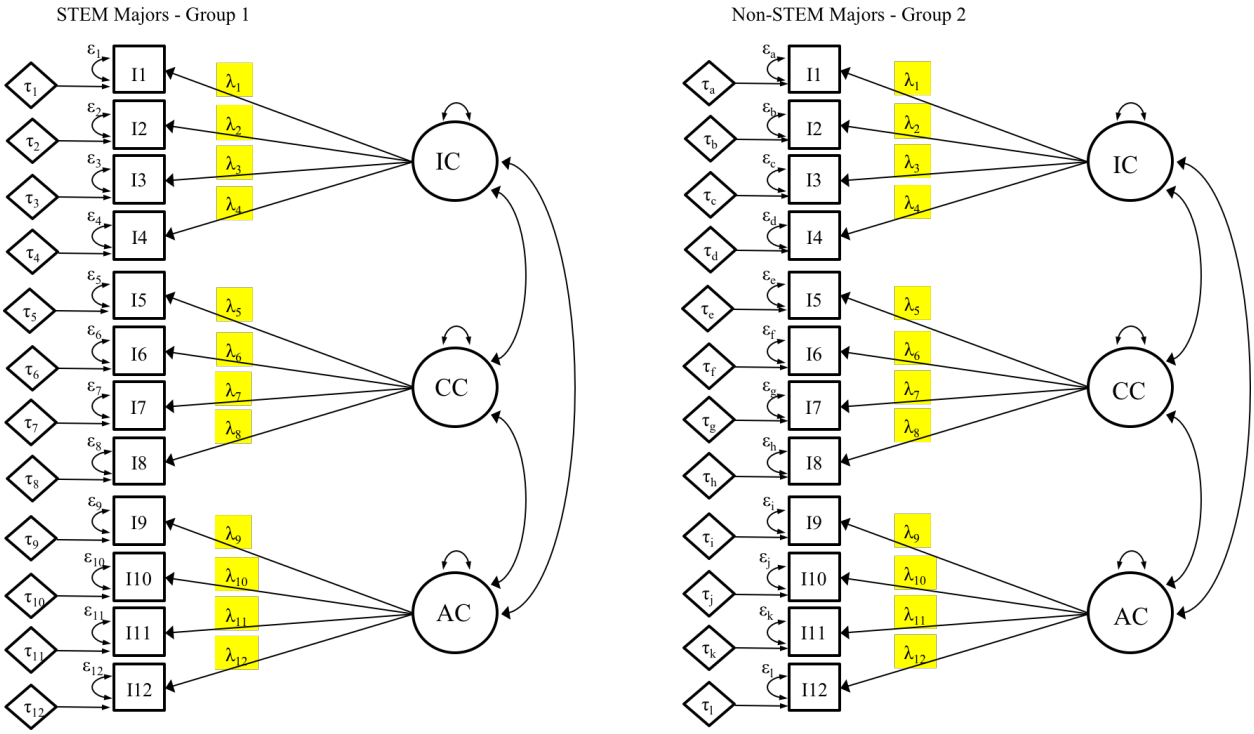


Figure 4.8. Metric model where factor loadings (highlighted) are constrained to be equal for both groups. All other parameters (e.g. intercepts and error variances) are freely estimated.

Failure to reach metric invariance suggests that the strength of association between items and the factor to which they belong are different between the groups. The strength of item association with the factor provides meaning to the factor from the perspective of the respondents (Gregorich, 2006). Therefore, if the item-factor associations are significantly different across groups, then the meaning of the underlying factor is different between groups, or the factor loadings are biased (Gregorich, 2006). Generally, when metric invariance is not achieved, there are one or more items with poor loadings for one of the groups compared to the other group. At this juncture, investigation of the item loadings or modification indices generated by the software can provide meaningful insight about the different ways that respondents may associate items to the underlying construct. After evaluation, researchers may choose to release the constraint of

equal loadings for the problematic item(s) and run the model again for partial measurement invariance (Byrne, Shavelson and Muthén, 1989; Putnick and Bornstein, 2016). If this release of constraints is undertaken, comparisons between groups are cautioned, particularly for the constructs that involve the problematic items. These items might be the subjects of further investigation as to the alignment between items and underlying constructs for the groups of interest.

Step 3: Scalar Invariance (Strong)

Once metric invariance is established (i.e., no evidence of significant difference is found between the metric and configural models), the next constraint can be applied. The *scalar model* (Figure 4.9), also called the *strong (factorial) invariance model* (Meredith, 1993), consists of incorporating unstandardized equal intercepts, in addition to equal loadings, across groups in the model (Gregorich, 2006; Sass, 2011; Putnick and Bornstein, 2016). With this addition, the intercepts (τ) are freely estimated for STEM majors, but for non-STEM majors they are set to be equal to the intercepts for STEM majors (see Figure 4.9). The purpose of this model is to establish evidence of unbiased estimated factor mean differences between groups (Gregorich, 2006), which implies that factor means encompass all mean differences in the shared variance of the items (Putnick and Bornstein, 2016). Factor means are unbiased because the error terms (ε) are not part of them. This is not true for observed item and observed scale means as they are calculated from the observed item scores that include the associated error terms (Putnick and Bornstein, 2016).

Just as with the metric model, first the scalar data-model fit is evaluated (Hu and Bentler, 1999) and then the fit comparison between, now, the metric (Figure 4.8) and scalar (Figure 4.9) models utilizing the appropriate values noted earlier. We reiterate that evaluating data-model fit is an important step of measurement invariance; however, essential to providing sufficient evidence for score comparisons is the change in fit statistics from one model to the next.

Once scalar invariance is achieved, the researcher has established evidence to support the comparison of *factor means* between groups. This evidence helps to rule out that any observed differences arise from variations caused by systematic higher or lower item responses (Gregorich, 2006; Sass, 2011; Putnick and Bornstein, 2016) due to issues like cultural norms.

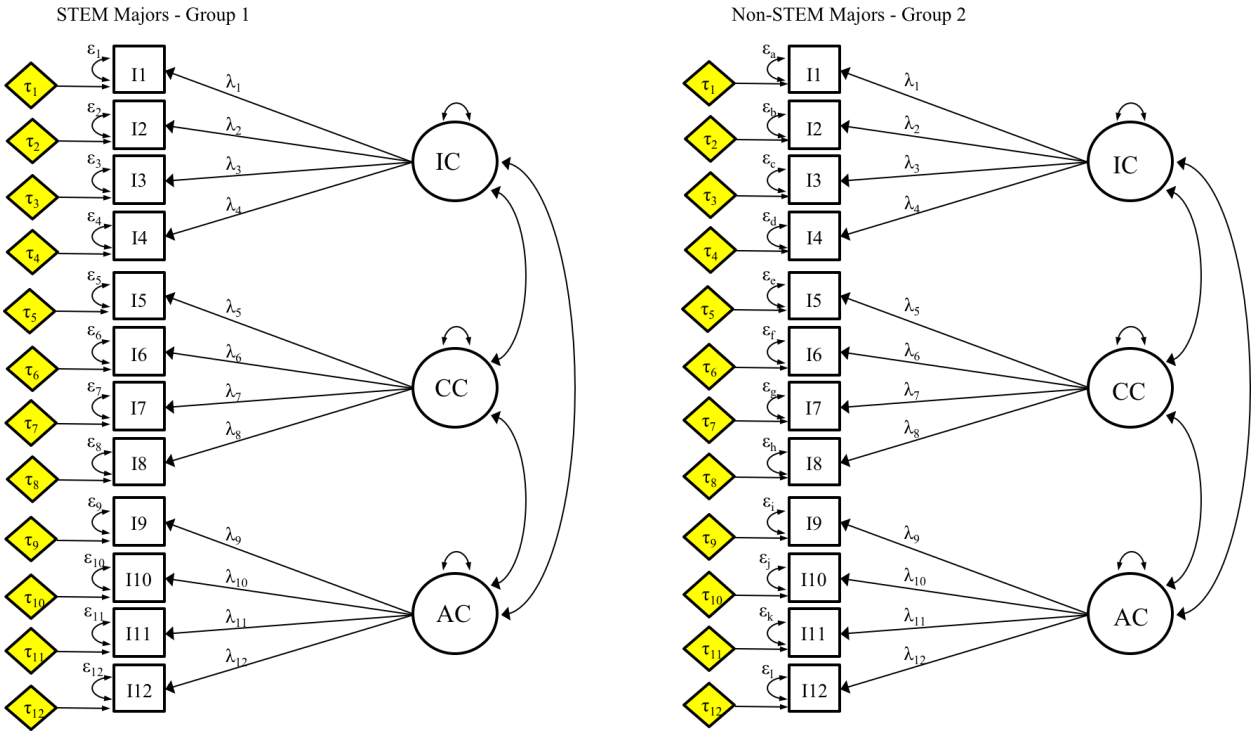


Figure 4.9. Scalar model where factor loadings and intercepts (highlighted) are constrained to be equal for both groups. All other parameters, including error variances are freely estimated.

If the scalar model provides results that are significantly different from the metric model, then scalar invariance has not been achieved and factor mean comparisons between groups are not supported. However, investigation as to the source of mismatch can be conducted. As demonstrated earlier, visualizing the data can be helpful at this juncture. Figure 5 shows item intercepts displayed as boxplots. Although one can choose to visualize data in various ways, Figure 4.5 visually suggests that the intercept for I3 (STEM majors) might be different than the intercept of the same item for the non-STEM majors. As I3 belongs to the IC factor, interpreting the IC factor mean comparisons between groups can be more difficult given this limitation. However, investigation as to the reason for the mismatch between groups is warranted. As previously mentioned, differences in item intercepts can be caused by diverging cultural norms that cause higher or lower item responses in diverse groups (Gregorich, 2006), thus investigating the source of the difference is encouraged. An example of this phenomenon that could cause systematic higher or lower responses is acquiescence bias. For example, one group might not utilize the entire response scale range, rather the response distribution is skewed to either end of the scale or narrowly in the middle.

In this situation, researchers may choose to release the constraint of equal intercepts for I3 only and evaluate the scalar model again. If releasing the constraint for I3 results in scalar model fit that is not significantly different from the metric model, then scalar invariance is established *with limitations*, sometimes described in terms of partial invariance (Putnick and Bornstein, 2016; Fischer and Karl, 2019). However, if an item loading was not held constant between groups in a previous step of invariance testing then the intercepts must also not be held constant as there is no reason to believe items with two different slopes would be expected to have the same intercepts.

There is some evidence that with partial invariance of intercepts comparison of factor means may provide acceptable results (Steinmetz, 2013).

An important distinction at this juncture is that factor means are obtained from the model, not from summing or taking the average of the observed item response values. Factor means are not a 'set' number, rather they are a comparison of latent (unobserved) means between two (or more) groups, where one group serves as the reference, taking the value of zero, and the other group or groups is/are compared to the reference. An effect size of the comparison can also be calculated (Hancock, 2001; Bunce *et al.,* 2017). Although this way of making comparisons is not frequently used in CER, the application of this practice is useful. We encourage researchers to work with factor means more often for two main reasons: 1) As explained earlier, factor means are estimated from the model, capture all mean differences in the shared variance of the items in the factor, and are free from error terms (Putnick and Bornstein, 2016). This cannot be said for observed scale scores, meaning composite scores taken directly as an average or sum of the observed variables (i.e., items), since these scores must include the error terms and do not take into account the strength of the association between items and factors. 2) In order to compare observed scale scores, the conservative invariance test, described in the following section, must be achieved. Meaning, it is harder to provide sufficient evidence for observed scale score comparison between groups than it is to compare factor means. Thus, we encourage researchers to utilize factor means as an effective tool for group comparisons as these values are void of error terms and will lead to more accurate interpretations and more meaningful inferences.

Step 4: Conservative Invariance (Strict)

Once scalar invariance is achieved, comparison of factor means between the groups is possible. However, if researchers desire to compare the observed scale scores of each factor; meaning composite scores taken directly as an average or sum of the observed variables (i.e., items), it is advisable to conduct a *conservative* or *strict* (Meredith, 1993) invariance test first (Gregorich, 2006; Sass, 2011). The conservative test checks the additional condition that measurement error variances are similar across groups. This is done in the same fashion as the prior models, with the final addition being that the STEM majors' error variances (ε) are freely estimated and non-STEM majors' error variances (ε) are constrained to be equal to those of STEM majors (see Figure 4.10). At this point, all loadings, intercepts and error variances are fixed to be equal between the groups to be compared. To establish strict invariance, the data-model fit statistics are first evaluated and then compared between the strict (Figure 4.10) and scalar (Figure 4.9) models and no evidence of significant difference should be found. If strict invariance is established, enough evidence is gathered to warrant observed scale score comparisons between groups (Gregorich, 2006; Sass, 2011). This type of comparison is what most researchers are accustomed to investigating; however, it is important to note that these comparisons require evidence of meeting this highest level of invariance testing. Failure to achieve strict invariance means that observed scale comparisons are not supported. Thus, researchers may investigate scalar invariance (i.e., Step 3) to compare factor scores instead.

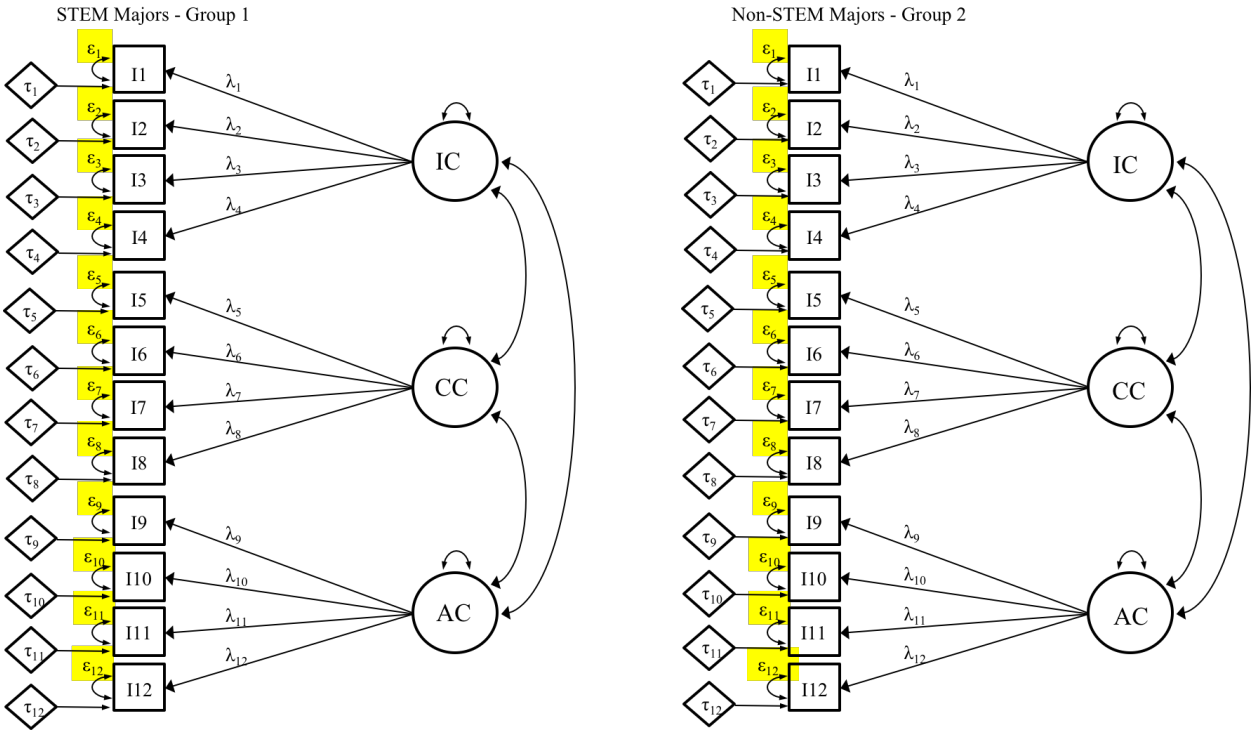


Figure 4.10. Conservative (strict) invariance where loadings, intercepts, and error variances are constrained to be equal for both groups.

Based on the four steps described previously, we provide a summary table (Table 4.1) for readers to reference as they conduct measurement invariance testing in their own studies. This table, while not comprehensive, provides the basic model characteristics, the evidence established, appropriate claims, and supported group comparisons that can be made at each level of invariance testing. This table can also prove useful as reviewers and journal editors review quantitative studies that can benefit from this method to support comparisons between groups or across time.

Table 4.1. Summary of Claims and Evidence Established at Each Stage of Measurement Invariance Testing - Guide for Researchers, Practitioners, and Reviewers.

	Configural	Metric (Weak)	Scalar (Strong)	Conservative (Strict)
Model characteristics	all parameters freely estimated in all groups, no constraints	factor loadings constrained to be the same for all groups	factor loadings and item intercepts constrained to be the same for all groups	factor loadings, item intercepts, and error variances constrained to be the same for all groups
Evidence established	same number of factors, items associated with the same specific factor for all groups	evidence in configural **plus** same strength of association between factors and corresponding items for all groups	evidence in configural and metric **plus** same item intercepts for all groups	evidence in configural, metric, and scalar **plus** same item error variances for all groups
Appropriate claims	items are associated with each other and the underlying factors in similar ways	claims from configural **plus** meaning of the factor (in terms of relative weight of items) is similar across groups	claims from configural and metric **plus** no systematic response biases; differences in factor means are due to a true difference in groups	claims from configural, metric, and scalar **plus** no systematic response biases or difference in error between groups; differences in item and scale means are due to a true difference in groups
Supported comparisons between groups	none	none	factor mean scores (from the model)	observed scale scores

Measurement Invariance Testing Example with Simulated Data

To illustrate the steps of utilizing measurement invariance testing for determining if, and to what degree, group comparisons can be made, we use the simulated dataset that generated Figures 4.4 and 4.5 to work through an example. The data was simulated to be continuous, therefore the maximum likelihood estimator was used for each model. For each step in the process, the data-model fit results as well as the fit comparisons between models are displayed in Table 4.2. It is important to note that while fit indices for each model will be calculated and tabulated by the software being used (i.e., R or Mplus, etc), the change values between models have to be manually calculated with a simple subtraction, with the exception of the *p*-value associated with the $\Delta\chi^2$, which must be retrieved from a χ^2 table that contains degrees of freedom.

At the baseline (Step 0) and configural (Step 1) levels, only the overall data-model fit is investigated. In our PRCQ example, the data at these levels was simulated with essentially perfect data-model fit as noted in Table 4.2. Perfect fit at these levels is unlikely to happen in a real study; thus, expecting a less-than-perfect fit is reasonable. Therefore, evaluating the data-model fit should follow acceptable guidelines, such as those by Hu and Bentler (1999) used here, or others as appropriate based on the data type. As each of our independent baseline models showed acceptable data-model fit and then the combined configural model showed good data-model fit, we can proceed to the next step of invariance testing.

The metric model data (Step 2) exhibits acceptable data-model fit (see Table 4.2). Beginning with these metric level indices, we not only evaluate the data-model fit but also compare

the fit obtained with the metric model to that of the configural model. First, we evaluate the $\Delta\chi^2$ (Cheung and Rensvold, 2002; Mueller and Hancock, 2019) which is a non-significant value, thus providing proof that there is no evidence of significant difference between the models. Then, following the suggestions of Chen (2007), our calculated values of ΔCFI = 0.000, ΔSRMR = 0.001, and ΔRMSEA = 0.003 are within the acceptable change cutoff levels: ΔCFI (< 0.01), ΔSRMR (< 0.03), and ΔRMSEA (< 0.015) to establish metric invariance (Chen, 2007). The comparison between configural and metric models shows that there is no evidence of significant change between these two models, thus metric invariance is achieved based on the comparison and we are warranted in moving to the next step of invariance testing.

For evaluating if scalar invariance is achieved (Step 3), a similar analysis pattern is followed. First, we evaluate the data-model fit. At this point, we observe that the fit indices for the scalar model are no longer within the acceptable ranges (see Table 4.2). This result is problematic because it is an indication that scalar invariance does not hold for the groups. Further evidence is found when we compare the change in fit indices between the metric and scalar models. Here we observe that our value of $\Delta\chi^2$ is significant, and the values for ΔCFI, ΔSRMR, and ΔRMSEA are also not within the recommended fit index cutoffs: ΔCFI (< 0.01), ΔSRMR (< 0.01), and ΔRMSEA (< 0.015) for scalar invariance (Chen, 2007). These additional results confirm that scalar invariance is not reached for these data. As the model at this level is not supported, we do not go on to evaluate the next highest level of invariance (i.e., the strict invariance model at Step 4), as we do not have a supported scalar model to compare it to. However, if the scalar model held and we desired to move on to test for strict invariance, the same guidelines and fit index cutoffs would be used as for scalar invariance (Chen, 2007).

In this simulated data example with the PRCQ, our analysis provided evidence for metric invariance at Step 2 but not for scalar invariance at Step 3. Therefore, these results imply that factor mean comparisons between STEM and non-STEM majors are not supported and should not be performed. Investigating the source of the misfit in the scalar model is warranted. Based on our previous discussion, we know that the I3 intercept is higher for the STEM majors compared to non-STEM majors (see Figures 4.4 and 4.5). At this point, we may choose to qualitatively investigate the difference between these groups for I3. Alternatively, we may choose to release this item's intercept constraint (*i.e.*, allowing the I3 intercept for each group to be freely estimated) and run the scalar model again. If the data-model fit and model comparisons indicate acceptable levels with this modification, *partial* scalar invariance would be achieved. At this point, we would have limited support for factor mean comparisons. However, we would not be able to make any significant claims, particularly for the IC factor, due to the limitation for I3. Based on this limitation, reflection on the consequences of making factor mean comparisons between these groups and the validity of inferences drawn from these comparisons is crucial. Finally, as we were not able to evaluate for scalar invariance, we have no basis for comparing the observed scale scores of the STEM and non-STEM majors using the PRCQ.

As we have described throughout this manuscript, and shown through the example here, measurement invariance testing provides researchers and practitioners with statistical evidence to support (or in this case, refute) comparisons between the groups evaluated (Sass, 2011). Once it has been established that both groups view the items on an instrument in similar ways (i.e., by establishing a certain level of measurement invariance), the interpretation of results becomes validated. Utilizing measurement invariance testing provides support for meaningful inferences

between populations, taking into account response patterns that may arise from a group's background or experiences (Wicherts, Dolan and Hessen, 2005). Furthermore, providing evidence that the data from an assessment instrument does not have validity threats against a comparison group, such as URMs (Gillborn *et al.,* 2017), provides more confidence in the results obtained and may provide increased support for claims of social inclusion for these groups.

Table 4.2. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors with Simulated Data for Illustration

Step	Testing level	$\chi 2$	*df*	*p*-value	CFI	SRMR	RMSEA	$\Delta\chi 2$	Δdf	*p*-value	ΔCFI	ΔSRMR	ΔRMSEA
0	STEM majors Baseline	65	51	0.084	0.998	0.021	0.017	-	-	-	-	-	-
0	Non-STEM majors Baseline	52	51	0.437	1.000	0.016	0.004	-	-	-	-	-	-
1	Configural	117	102	0.142	0.999	0.018	0.012	-	-	-	-	-	-
2	Metric	120	111	0.245	0.999	0.019	0.009	3	9	0.231	0.000	0.001	0.003
3	Scalar	2268	120	< 0.001	0.820	0.191	0.134	2148	9	< 0.001	0.179	0.172	0.125

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Limitations

While we encourage all researchers and practitioners to utilize measurement invariance testing prior to conducting comparisons between groups, we acknowledge there are limitations which may not allow the use of this method. One of these limitations is the sample size required to conduct these model-based tests. Similar to factor analysis techniques, measurement invariance testing requires a large sample size. Although there are no specific rules about the sample size

required, some work indicates that sample size requirements can be calculated given the number and value of the parameters being estimated (Wolf *et al.,* 2013; Mueller and Hancock, 2019). However, we encourage researchers to continue investigating newer methods to determine appropriate sample size that may be more suitable for this technique that is specific to the model parameters, the type of data being analyzed, and other characteristics of their study (Wolf *et al.,* 2013). Therefore, when conducting research and comparisons between groups with small samples, the technique presented in this work is not appropriate. Thus, we encourage researchers, practitioners, reviewers and journal editors to consider other methods of reflexivity such as response process evidence and/or content review by culturally-aware experts.

Additionally, researchers should be aware that the fit index cutoffs we have presented in this manuscript both for evaluating data-model fit and for change in model fit indices are suggested values based on simulation studies. While these guidelines are generally accepted within the field of measurement, this is an area of active investigation and these guidelines could evolve in coming years. As we encourage researchers to follow these guidelines, we also encourage a thoughtful evaluation of the data, model, and data-model fit where the suggested guidelines may not apply (Kang *et al*., 2016; McNeish *et al*., 2018).

Another limitation of measurement invariance testing is that this technique alone does not inform the exact ways in which groups differ in item and factor interpretation. Although this technique can point to the problematic items and factors that are dissimilar between groups, it cannot provide reasoning for the different meaning of items or factors between groups. This

information is best investigated using qualitative methods that can inform the perspective and interpretation from a respondent's point of view.

Finally, as with all statistical inferences, the measurement invariance testing process is built upon a series of assumptions. Without clearly identifying and acknowledging these assumptions, there is little support for the conclusions drawn from invariance testing. Due to the limited focus of this manuscript, only a few of the underlying assumptions for invariance testing were briefly discussed (i.e., theoretical support for the model being tested, quality of data being fit to the model, and acceptability of partial invariance at the metric and scalar stages). However, other assumptions are described more fully in the ESI and other resources (Bontempo and Hofer, 2007; Hancock *et al.*, 2009; Putnick and Bornstein, 2016; Fischer and Karl, 2019).

Discussion

CER is moving in the direction of greater interest in the differential impacts and outcomes of diverse populations (Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Shortlidge *et al.,* 2019). However, efforts to increase diversity by enrolling more URM students are not sustainable unless paired with efforts to increase social inclusion and social justice (O'Shea *et al.,* 2016; Puritty *et al.,* 2017). In an effort to 're-imagine' quantitative approaches to better serve social justice initiatives (García, López and Vélez, 2017) and raise the standards for investigating these issues at different intersections of identity and background (e.g., race and gender; race and math preparation, etc.), we have presented a statistical method which investigates potential validity

threats that could arise when analyzing assessment instrument data. Particular focus has been given at each stage of the analysis to explain some issues that could be evidenced if a given model fails to reach acceptable data-model fit criteria. We have included a few examples that could provide readers some ideas to begin their investigation when measurement invariance is not established at a particular level. Suggestions for circumventing some of these difficulties, such as releasing individual item parameters, have also been presented along with their implications. A summary of each stage of testing, along with the supported claims and evidence established is provided in Table 4.1.

Many recent studies in CER have taken the first step toward raising the research standards by including variables such as gender, race, etc. and appropriate intersections in their studies (Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Shortlidge *et al.,* 2019). However, the next step of investigating the validity of the group comparisons was lacking. Therefore, we encourage researchers to investigate their own data, even the data that has already been published, and consider whether the inferences made were valid for the populations being compared. One recent example of this practice is the study conducted by Rocabado and colleagues (2019), which explored data from a study done in 2016 by Mooring and colleagues who conducted an evaluation of the attitude impact of an organic chemistry flipped classroom compared to a traditional classroom. The researchers found that the flipped classroom showed significant attitude gains when compared to the traditional classroom (Mooring *et al.,* 2016). Rocabado and colleagues (2019), not only investigated whether the original comparison was supported, but also studied whether the attitude gains observed extended to the Black female students in the original sample by utilizing measurement invariance testing to support the investigation and comparisons.

Measurement invariance testing provides opportunities to investigate levels of differences that could arise any time group comparisons are to be made. The 4-Step method presented in this primer is not limited to group comparisons by gender, race, or ethnicity only, it includes groups such as those used in this manuscript (i.e., STEM and non-STEM majors) and to same-group analyses in longitudinal comparisons (e.g., pre-post gain). Regardless of how the groups are defined, at the configural model level (Step 1), an acceptable data-model fit suggests that the groups utilize the same network of equations and the basic measurement model (e.g., number of factors present). At this stage, the claim can be made that item associations are similar between groups, as demonstrated by Figure 4.2. The configural model provides a lens to observe these item associations when the data is disaggregated by the defined groups. Item correlations might not be similar for all groups and therefore, the configural model might not reach acceptable levels of data-model fit, suggesting group-level differences in the constructs being measured. If this level cannot be achieved, comparisons between groups are not fair due to the difference in constructs. This is an important step in measurement invariance testing, as it provides a strong foundational model on which to base the subsequent tests.

The metric model (Step 2) investigates the strength of the association between factors and their corresponding items (Sass, 2011). The strength of these relations indicates the meaning of the factor (Gregorich, 2006). Therefore, when the metric model fails, it is evidence of differences in factor meaning between the groups, which provides grounds for further investigation. These differences are observed when the entire pattern of item loadings differs between groups. As this result does not indicate why the groups differ in meaning, a thorough investigation of construct meaning is advised, data from items and constructs should be reviewed for content validity,

response process validity, and construct validity evidence, keeping in mind the various groups that could be in the target population. Metric non-invariance may also arise when one or more item loadings on a factor differ greatly between groups (see Figure 4.3c), indicating that one group does not associate the item(s) with the construct being measured, while the other group does. For example, in the *fictitious* Applications of Chemistry (AC) scale, a problematic item might ask about the field of Materials Science. As this field is interdisciplinary between Engineering, Physics, and Chemistry, it is likely that STEM students would have been exposed to examples from the field across many courses. However, non-STEM majors may have never been exposed to the ideas and examples of Material Science and the role Chemistry plays. Therefore, when comparing a group of STEM majors, who are more likely to have been exposed to Materials Science, to a group of non-STEM majors, it is possible that this item functions differently between the groups. The non-STEM majors might not view Material Science as being an application on the AC scale because they have not been introduced to this field and its interconnections. Therefore, when an item cannot be explained by the underlying construct for one group, the meaning of the construct is different between the groups.

The scalar model (Step 3) considers whether item averages within the measurement model are similar across groups. As shown in Figure 4.4, item averages may look similar when combined; however, when disaggregated into groups, item means could be different (Figure 4.5) leading to the scalar model not reaching acceptable levels of fit. These differences could arise due to acquiescence biases that affect one group and not the other due to cultural norms not shared between groups (Gregorich, 2006). In the *fictitious* Connectedness of Chemistry (CC) scale, a problematic item might ask about the degree to which chemistry is connected to a specific issue of

global warming, say CO_2 emission. One can think that STEM majors might see stronger ties between the issue and chemistry and therefore score higher on this item than a group of non-STEM majors that may not have been exposed to the idea of light-matter interactions. Therefore, if all the STEM majors score this item high (*i.e.*, a 4 or 5 on a 5-point scale) because they have learned about this phenomenon, then the scale is biased for this item between the two groups in this context. If scalar invariance is not achieved, comparisons between groups beyond the metric model level are not warranted. On the other hand, if scalar invariance is reached, estimated factor mean scores can be computed and compared between groups with evidence that differences between groups are not artifacts of the instrument and construct meaning is similar across the groups. However, if a researcher's goal is to compare observed factor scores (*e.g.*, observed item averages), evidence of conservative invariance (Step 4), in which error variances are constrained to be equal between groups, is required (Sass, 2011).

While conducting measurement invariance testing, each stage provides safeguards and reflexivity (Gillborn *et al.,* 2018) about the groups being compared, rendering this quantitative approach suitable for investigating the differential impacts and outcomes of diverse populations and advancing social justice and equity in CER at the institutional level. We encourage all researchers and practitioners not only to investigate the impact of variables such as race/ethnicity and appropriate intersections (e.g., gender status, language status, socioeconomic status) more often in their research and in their classrooms, but also to employ techniques such as measurement invariance testing in order to safeguard against disguising racism and other social injustices and systemic biases when making comparisons between groups (Gillborn *et al.,* 2018; García, López and Vélez, 2018).

Recommendations and Implications

Measurement invariance testing provides evidence to support or refute quantitative data any time group comparisons are to be made. Although qualitative methodologies are used more often to investigate individuals' and groups' lived experiences, utilizing quantitative methods with reflexivity and safeguards against racial and other biases (Gillborn *et al.,* 2018; García, López and Vélez, 2018) can enhance research and teaching that aims at studying pedagogies and interventions that benefit URMs in chemistry. This quantitative method is not limited to group comparisons by gender, race, or ethnicity. It includes groups such as those defined by academic major, socioeconomic status, transfer status, or other meaningful categories and also extends to same-group analyses in longitudinal comparisons (*e.g.*, pre-post gain). To make the endeavor of utilizing measurement invariance testing as easy and accessible as possible, we have provided code and ample explanation for two common software programs (R and Mplus) in the ESI. Although we provided code for these programs, there are a variety of other programs available that support this technique such as SAS, LISREL, EQS, or the AMOS add-in for SPSS. A helpful comparison of software for structural equation modeling with multiple groups can be found in Narayanan (2012).

For Researchers and Reviewers

Measurement invariance testing is a technique that we encourage all researchers to use when analyzing assessment instrument data for the purpose of group comparison in their studies. Identifying potential validity threats will greatly enhance the interpretation of the results obtained

and claims made, as well as further the answer to the call for increased diversity and social inclusion. At each specific stage of measurement invariance testing, certain model claims can be supported or refuted, which either provide evidence for group comparison (see Table 4.1) or inform the subsequent steps to take in the research. Each of the measurement invariance steps is an opportunity to safeguard against observed and unobserved differences between groups that may be artifacts of the assessment instrument. As researchers, it is our duty to ensure that we present results that have the potential of being transformative; thus, working to minimize artifacts of measurement bias in our analyses is imperative to further the field of CER in more inclusive ways.

Likewise, when reviewing articles for publication, reviewers have the responsibility to ensure that the analyses conducted are held to high standards and that the results and implications are supported by sufficient evidence. In this work, we have highlighted the importance of conducting measurement invariance testing when researchers and practitioners utilize assessment instruments of latent traits on which groups will be compared. The results of these comparisons can have important implications and consequences in CER as the field moves toward greater diversity and social inclusion. Thus, these comparisons have to be made responsibly to properly address the consequential validity of the inferences drawn from studies where group or longitudinal comparisons are made. Particularly, we advocate for safeguards and reflexivity in research methodology that aims to challenge the idea of neutral and objective research in an effort to work toward the abolition of social inequities (Solórzano, 1997; Yosso, 2005). Therefore, we urge reviewers and journal editors to check the conditions necessary for the comparison of outcomes by group. First, ensuring that researchers provide reason to believe it is valuable to compare the noted groups (i.e., the comparisons are not simply because the demographic data

exists) on the variable of interest. Second, that there is reason to believe the construct being compared can be measured appropriately for all groups through establishing the relevant level of measurement invariance. We have shown how measurement invariance testing can provide reflexivity and ample opportunity to check for differences in measurement for groups in studies. Thus we encourage the use of this method whenever possible.

Often, the comparisons made between groups will be done at the observed scale score level. If this is the ultimate goal of a study, then the researchers and reviewers should be aware that observed score comparisons require meeting strict invariance (the most conservative level of invariance) across all groups. If this strict invariance model provides acceptable data-model fit, then researchers and reviewers have evidence that observed scale scores can be compared between groups. Within this primer on measurement invariance testing, we laid out a step-by-step method, working up to establishing strict invariance. However, it is beneficial to mention that if only the strict invariance test is conducted, the investigation at each stage of measurement invariance testing is not provided and the change in data-model fit from one level to the next is not produced. Although valuable step-by-step information is not obtained when choosing to run only the desired test, this practice is sound. However, if the strict invariance test fails to provide acceptable data-model fit, then researchers may benefit from conducting the lower level tests and investigating the source of measurement non-invariance. Table 4.1 provides a summary of appropriate claims and comparisons at each level of measurement invariance.

For Practitioners

We encourage practitioners to use measurement invariance testing, when possible, in any endeavor to inform their practice where group comparisons with assessment instrument data of latent traits are utilized. Safeguarding against threats to the validity of the inferences drawn from group comparison studies is fundamental to the evaluation and success of inclusive pedagogies in the classroom. We acknowledge that sample size is often a limitation in many studies. Thus we advise practitioners to utilize similar processes of reflexivity to safeguard against threats to the validity of inferences against groups that are appropriate for their sample size, such as cognitive interviews (Willis, 1999). This practice will help to ensure that the investigations conducted across individual and institutional levels remain mindful of the tenets of CRT and move toward, rather than away from, equity. Additionally, we recommend the collaboration between practitioners and researchers in analyzing and interpreting quantitative data, particularly when comparing groups. These collaborations can be fruitful and inform a wider variety of settings in which our studies take place, providing the field of CER a broader and more complete view of the field as it advances toward greater diversity and social inclusion.

Lastly, we urge practitioners to review the research literature with a critical lens and hold research findings to a high standard when data is compared by group. Following the steps of measurement invariance testing can inform whether an instrument can be utilized to make meaningful comparisons with diverse groups. For a practical approach, if measurement invariance testing is not feasible, we suggest a careful review of the literature for instruments which have

been appropriately tested with diverse populations, to support appropriate data collection and analyses that lead to meaningful conclusions.

References

AERA, APA, and NCME., (2014), *Standards for Educational and Psychological Testing,* American Psychological Association, Washington, DC.

Apple M. W., (2001), *Educating the 'Right' Way: Markets, Standards, God, and Inequality.* RoutledgeFalmer, New York, NY.

Arjoon J. A., Xu X., and Lewis J. E., (2013), Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence, *J. Chem. Educ.*, **90**, 536-545. DOI: 10.1021/ed3002013

Beier M. E., Kim M. H., Saterbak A., Leautaud V., Bishnoi S., and Gilberto J. M., (2019), The effect of authentic project-based learning on attitudes and career aspirations in STEM, *J. Res. Sci. Teach.,* **56**(1), 3-23. DOI: 10.1002/tea.21465

Bontempo D. E. and Hofer S. M., (2007), Assessing Factorial Invariance in Cross-Sectional and Longitudinal Studies., in Ong A. D. and van Dulmen M. H. M. (eds.), *Series in positive psychology. Oxford handbook of methods in positive psychology*. Oxford University Press, pp. 153–175.

Bornstein M. H., (1995), Form and function: Implications for studies of culture and human development, *Cult. Psychol.,* **1**(1), 123–137. DOI: 10.1177/1354067X9511009

Bowen N. K., and Masa R. D., (2015), Conducting measurement invariance tests with ordinal data: A guide for social work researchers, *J. Soc. Social Work Res.,* **6**(2), 229-249. DOI:10.1086/681607

Brandriet A. R., and Bretz S. L., (2014), The development of the redox concept inventory as a measure of students' symbolic and particular redox understandings and confidence, *J. Chem. Educ.,* **91**, 1132-1144. DOI: 10.1021/ed500051n

Bretz S. L., (2014), Designing assessment tools to measure students' conceptual knowledge of chemistry. In *Tools of Chemistry Education Research,* Bunce D., Cole R., Eds., ACS Symposium Series. DOI: 10.1021/bk-2014-1166.ch009

Brown T. A., (2006), *Confirmatory Factor Analysis for Applied Research*, The Guilford Press, New York, NY.

Bunce D. M., Komperda R., Schroeder M. J., Dillner D. K., Lin S., Teichert M. A., and Hartman J. R., (2017), Differential use of study approaches by students of different achievement levels, *J. Chem. Educ.,* **94**(10), 1415–1424. DOI:10.1021/acs.jchemed.7b00202

Byrne B. M., Shavelson R. J., and Muthén B., (1989), Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance, *Psychol. Bull.,* **105**(3), 456-466. DOI: 10.1037/0033-2909.105.3.456

Candell G. L., and Drasgow F., (1988), An iterative procedure for linking metrics and assessing item bias in item response theory, *Appl. Psychol. Meas.,* **12**(3), 253-260. DOI:10.1177/014662168801200304

Ceci S. J., Williams W. M., and Barnett S. M., (2009), Women's underrepresentation in science: Sociocultural and biological considerations, *Psychol. Bull.,* **135**(2), 218–261. DOI:10.1037/a0014412

Chen F. F., (2007), Sensitivity of goodness of fit indexes to lack of measurement invariance, *Struct. Equ. Modeling*, **14**(3), 464-504. DOI: 10.1080/10705510701301834

Cheung G. W., and Rensvold R. B., (1999), Testing factorial invariance across groups: A reconceptualization and proposed new method, *J. Manage.,* **25**(1), 1-27. DOI:10.1177/014920639902500101

Cheung G. W., and Rensvold R. B., (2002), Evaluating goodness-of-fit indexes for testing measurement invariance, *Struct. Equ. Modeling,* **9**(2), 233-255. DOI:10.1207/S15328007SEM0902_5

Cohen J., (1988), *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ.

Counsell, A., Cribbie, R. A., and Flora, D. B., (2019), Evaluating equivalence testing methods for measurement invariance, *Multivar. Behav. Res.,* **55**(2), 312-329. DOI:10.1080/00273171.2019.1633617

Covarrubias A., (2011), Quantitative intersectionality: A critical race analysis of the Chicana/o educational pipeline, *J. Latinos Educ.,* **10**(2), 86-105. DOI: 10.1080/15348431.2011.556519

Covarrubias A., and Velez, V., (2013), Critical race quantitative intersectionality: An antiracists research paradigm that refuses to 'Let the numbers speak for themselves.' In *Handbook of Critical Race Theory in education,* (Eds.) Dixson A., Lynn, M. New York City, Routledge, pp. 270-285.

Crenshaw K., (1989), Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics, *Univ. Chicago Leg. For.,* 139-168.

Crenshaw K., (1995), Critical Race Theory: The key writings that formed the movement, New York City: State University of New York Press.

Deng X., Doll W. J., Hendrickson A. R., and Scazzero J. A., (2005), A multi-group analysis of structural invariance: An illustration using the technology acceptance model, *Inform. Manage.,* **42**, 745-759. DOI: 10.1016/j.im.2004.08.001

Delgado A., and Stefanic J., (2001), Critical Race Theory: An Introduction, New York City, NYU Press.

Dixson A., and Anderson C. R., (2018), Where are we? Critical Race Theory in education 20 years later, *Peabody J. Educ.,* **93**(1), 121-131. DOI: 10.1080/0161956X.2017.1403194

Fernández L., (2002), Telling stories about school: Using critical race and Latino critical theories to document Latina/Latino education and resistance, *Qual. Inq.,* **8**(1), 45-65. DOI:10.1177/107780040200800104

Ferrell B. and Barbera J., (2015), Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry, *Chem. Educ. Res. Pract.,* **16**, 318-337. DOI: 10.1039/C4RP00152D

Ferrell B., Phillips M. M., and Barbera J., (2016), Connecting achievement motivation to performance in general chemistry, *Chem. Educ. Res. Pract.,* **17**, 1054-1066. DOI: 10.1039/C6RP00148C

Fink A., Cahill M. J., McDaniel M. A., Hoffman A., and Frey R. F., (2018), Improving general chemistry performance through a growth mindset intervention: selective effects on underrepresented minorities, *Chem. Educ. Res. Pract.,* **19***,* 783-806. DOI: 10.1039/C7RP00244K

Finney S. J. and DiStefano C., (2013), Non-normal and categorical data in structural equation modeling., in Hancock G. R. and Mueller R. O. (eds.), *Structural equation modeling: a second course*. Charlotte, NC: Information Age Publishing, pp. 439–492.

Fischer R. and Karl J. A., (2019), A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Front. Psychol.*, **10**, 1–18. DOI: 10.3389/fpsyg.2019.01507

García N. M., López N., and Vélez V. N., (2018), QuantCrit: Rectifying quantitative methods through critical race theory, *Race Ethn. Educ.,* **21**(2), 149-157. DOI:10.1080/13613324.2017.1377675

Garson D., (2012), *Testing statistical assumptions.* Statistical Associates Publishing, Asheboro, NC.

Gibbons R. E., and Raker J. R., (2018), Self-beliefs in organic chemistry: Evaluation of a reciprocal causation, cross-lagged model*, J. Res. Sci. Teach.*, **56**(5), 598-615. DOI:10.1002/tea.21515

Gibbons R. E., Xu X., Villafañe S. M., and Raker J. R., (2018), Testing a reciprocal causation model between anxiety, enjoyment and academic performance in postsecondary organic chemistry, *Educ. Psychol.*, **38** (6), 838-856. DOI: 10.1080/01443410.2018.1447649

Gillborn D., Warmington P., and Demack S., (2018), QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics, *Race Ethn. Educ.,* **21** (2), 158-179. DOI:10.1080/13613324.2017.1377417

Gregorich S. E., (2006), Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework, *Med Care,* **44** (11 Suppl 3), S78-S94. DOI: 10.1097/01.mlr.0000245454.12228.8f

Hancock G. R., (2001), Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct, *Psychometrika,* **66**, 373–388.

Hancock G. R., and French B., (2013), *Power analysis in covariance structure modeling.* In G. R. Hancock and R. O. Mueller (Eds.), Structural equation modeling: A second course (2nd ed.), Information Age Publishing, Charlotte, NC, pp. 117-159.

Hancock G. R., Stapleton L. M., and Arnold-Berkovits I., (2009), The tenuousness of invariance tests within multisample covariance and mean structure models., in *Structural equation modeling in educational research: concepts and applications*., pp. 137–174.

Hensen C. and Barbera J. (2019), Assessing affective differences between a virtual general chemistry experiment and a similar hands-on experiment, *J. Chem. Educ.,* **96**, 2097−2108. DOI:10.1021/acs.jchemed.9b00561

Hirschfeld G., and Von Brachel R., (2014), Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal, *Pract. Assess. Res. Eval.*, **19**(7), 1–11. DOI: 10.7275/qazy-2946

Hong L. and Page S. E., (2004), Groups of diverse problem solvers can outperform groups of high-ability problem solvers, *P.Natl. Acad. Sci.,* **101**(46), 16385-16389. DOI:10.1073/pnas.0403723101

Hosbein K. N., and Barbera J., (2019), Development and valuation of novel science and chemistry identity measures, *Chem. Educ. Res. Pract.*, **21,** 852-877. DOI: 10.1039/C9RP00223E

Hu L. T., and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Struct. Equ. Modeling,* **6**(1), 283−292. DOI: 10.1080/10705519909540118

Hurtado S., Newman C. B., Tran M. C., and Chang M. J., (2010), *Improving the Rate of Success for Underrepresented Racial Minorities in STEM Fields: Insights from a National Project.* In New Directions for Institutional Research, no. 148. Wiley Periodicals, Inc.

Ireland D. T., Freeman K .E., Winston-Proctor C. E., Delaine K. D., McDonald Lowe S., and Woodson K. M., (2018), (Un)hidden figures: A synthesis of research examining the intersectional experiences of Black women and girls in STEM, *Rev. Res. Educ.,* **42***,* 226-254. DOI: 10.3102/0091732X18759072

Jiang B., Xu X., García A., and Lewis J. E., (2010), Comparing two tests of formal reasoning in a college chemistry context, *Chem. Educ. Res. Pract.,* **87**(12), 1430-1437. DOI:10.1021/ed100222v

Jöreskog K. G. (1971). Simultaneous factor analysis in several populations, *Psychometrika*, **36**, 409–426.

Kahveci A., (2015), Assessing high school students' attitudes toward chemistry with a shortened semantic differential, *Chem. Educ. Res. Pract.,* **16**, 283−292. DOI: 10.1039/C4RP00186A

Kang Y., McNeish D. M., and Hancock G. R., (2016), The role of measurement quality on practical guidelines for assessing measurement and structural invariance, *Educ. Psychol. Meas.*, **76**(4), 533–561. DOI: 10.1177/0013164415603764

Keefer K. V., Holden R. R., and Parker J. D. A., (2013), Longitudinal assessment of trait emotional intelligence: Measurement invariance and construct continuity from late childhood to adolescence, *Psychol. Assess.,* **25**(4), 1255-1272. DOI: 10.1037/a0033903

Kendhammer L., Holme T., and Murphy K., (2013), Identifying differential performance in general chemistry: Differential item functioning analysis of ACS general chemistry trial tests, *J. Chem. Educ.,* **90***,* 846-853. DOI: 10.1021/ed4000298

Kendhammer L. K., Murphy K., (2014), General statistical techniques for detecting differential item functioning based on gender subgroups: A comparison of the Mantel-Haenszel procedure, IRT, and logistic regression. In *Innovative Uses of Assessments for Teaching and Research ACS Symposium Series,* American Chemical Society: Washington, DC.

Komperda R., Hosbein K. N. and Barbera J., (2018), Evaluation of the influence of wording changes and course type on motivation instrument functioning in chemistry, *Chem. Educ. Res. Pract.,* **19**, 184-198. DOI: 10.1039/C7RP00181A

Lieber R. L., (1990), Statistical significance and statistical power in hypothesis testing, *J. Orthop. Res.,* **8***,* 304-309. DOI: 10.1002/jor.1100080221

Litzler E., Samuelson C. C., and Lorah J. A., (2014), Breaking it down: Engineering students STEM confidence at the intersection of race/ethnicity and gender, *Res. High. Educ.,* **55**, 810-832. DOI 10.1007/s11162-014-9333-z

Liu Y., Ferrell B., Barbera J., and Lewis J. E., (2017), Development and evaluation of a chemistry-specific version of the academic motivation scale (AMS-Chem), *Chem. Educ. Res. Pract.,* **18***,* 191-213. DOI: 10.1039/C6RP00200E

Loertscher J., (2010), Using assessment to improve learning in the biochemistry classroom, *Biochem. Mol. Biol. Educ.,* **38** (3), 188-189. ISSN-1470-8175

López N., Erwin C., Binder M., and Chavez M. J., (2018), Making the invisible visible: Advancing quantitative methods in higher education using Critical Race Theory and intersectionality, *Race Ethnic. Educ.,* **21**(2), 180-207. DOI: 10.1080/13613324.2017.1375185

McNeish D., An J., and Hancock G. R., (2018), The thorny relation between measurement quality and fit index cutoffs in latent variable models. *J. Pers. Assess.*, **100**(1), 43–52. DOI:10.1080/00223891.2017.1281286

Mellenbergh G. J., (1989), Item bias and item response theory, *Int. J. Educ. Res.,* **13***,* 127-143. DOI: 10.1016/0883-0355(89)90002-5

Meredith W., (1993), Measurement equivalence, factor analysis, and factorial equivalence, *Psychometrika*, **58**, 525–543.

Messick S., (1995)., Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *Am. Psychol.,* **50**(9), 741-749.

Montes L. H., Ferreira R. A., and Rodriguez C., (2018), Explaining secondary school students' attitudes towards chemistry in Chile, *Chem. Educ. Res. Pract.,* **19***,* 533-542. DOI: 10.1039/C8RP00003D

Mooring S. R., Mitchell C. E., and Burrows N. L., (2016), Evaluation of a flipped, large enrollment organic chemistry course on student attitude and achievement, *J. Chem. Educ.,* **93**, 1972-1883. DOI: 10.1021/acs.jchemed.6b00367

Mueller R. O., Hancock G. R., (2019), *Structural Equation Modeling.* In G. R. Hancock, L. M. Stapleton, and R. O. Mueller (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 445-456). Routledge, New York, NY.

Muthén L. K. and Muthén B. O., (2010), *Mplus User's Guide*, 6th ed., Muthén and Muthén: Los Angeles, CA.

Narawathne I. N., (2019), Introducing diversity through an organic approach, *J. Chem. Educ.,* **96** (9), 2042-2049. DOI: 10.1021/acs.jchemed.8b00646

Narayanan A., (2012), A review of eight software packages for structural equation modeling, *Am. Stat.*, **66**(2), 129–138. DOI: 10.1080/00031305.2012.708641

O'Shea S., Lysaght P., Roberts J., and Harwood V., (2016), Shifting the blame in higher education – social inclusion and deficit discourses, *High. Educ. Res. Dev.,* **35** (2), 322-336. DOI:10.1080/07294360.2015.1087388

Puritty C., Strickland L. R., Alia E., Blonder B., Klein E., Kohl M. T., McGee E., Quintana M., Ridley R. E., Tellman B., and Gerber L. R., (2017), Without inclusion, diversity initiatives may not be enough, *Science,* **357** (6356), 1101-1102. DOI: 10.1126/science.aai9054

Putnick D. L., and Bornstein M. H., (2016), Measurement invariance conventions and reporting: The state of the art and future directions for psychological research, *Dev. Rev.,* **47***,* 71–90. DOI:10.1016/j.dr.2016.06.004

Rath K. A., Peterfreund A., Bayliss F., Runquist E., and Simonis U., (2012), Impact of supplemental instruction in entry-level chemistry courses at a midsized public university, *J. Chem. Educ.,* **89***,* 449-455. DOI: 10.1021/ed100337a

Richards-Babb M., and Jackson J. K., (2011), Gendered responses to online homework use in general chemistry, *Chem. Educ. Res. Pract.,* **12***,* 409-419. DOI: 10.1039/C0RP90014A

Roadrangka V., Yeany R. H., and Padilla M. J., (1983), Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX.

Rocabado G. A., Kilpatrick N. A., Mooring S. R., and Lewis J. E., (2019), Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to

support valid comparisons between Black female students and their peers in an organic chemistry course, *J. Chem. Educ.,* **96**(11), 2371-2382. DOI: acs.jchemed.9b00516

Salta K., and Koulougliotis D., (2015), Assessing motivation to learn chemistry: Adaptation and validation of Science Motivation Questionnaire II with Greek secondary school students, *Chem. Educ. Res. Pract.,* **16,** 237-250. DOI: 10.1039/C4RP00196F

Sass D., (2011), Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework, *J. Psychoeduc. Assess.,* **29**(4), 347−363. DOI:10.1177/0734282911406661

Seadler A., (2012), Obama introduces plan to increase U. S. STEM undergraduates, *Earth,* **57**(6), 27.

Shepard L. A., (1993), Evaluating Test Validity, *Am. Educ. Res. Assoc.,* **19,** 405-450. DOI:10.3102/0091732X019001405

Shortlidge E. E., Rain-Griffith L., Shelby C., Shusterman G. P., and Barbera J., (2019), Despite similar perceptions and attitudes, postbaccalaureate students outperform in introductory biology and chemistry courses, *CBE-Life Sci. Educ.,* **18**(3), 1-14. DOI: 10.1187/cbe.17-12-0289

Solórzano D. G., (1997), Images and words that wound: Critical Race Theory, racial stereotyping, and teacher education, *Teach. Educ. Quart.,* **24**(3), 5-19. https://www.jstor.org/stable/23478088

Solórzano D. G., (1998), Critical Race Theory , race and gender microaggressions, and the experiences of Chicana and Chicano scholars, *Int. J. Qual. Stud. Educ.,* **11**(1), 121-136. DOI:10.1080/095183998236926

Solórzano D. G., and Ornelas A., (2004), A critical race analysis of Latina/o and African American advanced placement enrollment in public high schools, *High School J.,* **87**(3), 15-26. http://www.jstor.com/stable/40364293

Stanich C. A., Pelch M. A., Theobald E. J., and Freeman S., (2018), A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first generation students, and women, *Chem. Res. Educ. Pract.,* **19***,* 846-866. DOI: 10.1039/C8RP00044A

Steinmetz H., (2013), Analyzing Observed Composite Differences Across Groups, *Methodology*, **9**(1), 1–12. DOI: 10.1027/1614-2241/a000049

Stevens J. P., (2007), *Intermediate Statistics: A Modern Approach*, 3rd Edition, Routledge Taylor and Francis Group, New York, NY.

Tobin K. G. and Capie W., (1981), The development and validation of a group test of logical thinking, *Educ. Psychol. Meas.*, **41**(2), 413–423. DOI: 10.1177/001316448104100220

Tsui L., (2007), Effective strategies to increase diversity in STEM fields: A review of the research literature, *J. Negro Educ.,* **76**(4), 555-581. http://www.jstor.com/stable/40037228

Vandenberg R. J., and Lance C. E., (2000), A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research, *Organ. Res. Methods,* **2**, 4–69. DOI: 10.1177/109442810031002

Villafañe S. M., Bailey C. P., Loertscher J., Minderhout V., and Lewis J. E., (2011), Development and analysis of an instrument to assess student understanding of foundational concepts before biochemistry coursework, *Biochem. Mol. Biol. Educ.,* **39**(2), 102-109. DOI:10.1002/bmb.20464

Villafañe S. M., García C. A., and Lewis J. E., (2014), Exploring diverse students' trends in chemistry self-efficacy throughout a semester of college-level preparatory chemistry, *Chem. Educ. Res. Pract.,* **15**(2), 114-127. DOI: 10.1039/C3RP00141E

Wicherts J. M., Nolan C. V., and Heesen D. J., (2005), Stereotype threat and group differences in test performance: A question of measurement invariance, *J. Pers. Soc. Psychol.,* **89**(5), 686-716. DOI: 10.1037/0022-3514.89.5.696

Widaman K. F., and Reise S. P., (1997), Exploring the measurement invariance of psychological instruments: Applications in the substance use domain, In: Bryant K. J., Windle M.E., West S.G., (Eds), *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research,* American Psychological Association, Washington, DC.

Willis G. B., (1999), Cognitive interviewing: A "how to" guide, *Meeting of the American Statistical Association,* Research Triangle Institute.

Wolf E. J., Harrington K. M., Clark S. L., and Miller M. W., (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety, *Educ. Psychol.,* **73**(6), 913-934. DOI: 10.1177/0013164413495237

Wren D., and Barbera J., (2013), Gathering evidence for validity during the design, development, and qualitative evaluation of the thermochemistry concept inventory, *J. Chem. Educ.,* **90**, 1590-1601. DOI: 10.1021/ed400384g

Xu X., Kim E. S., and Lewis J. E., (2016), Sex difference in spatial ability for college students and exploration of measurement invariance, *Learn. Individ. Differ.,* **45**, 176–184. DOI:10.1016/j.lindif.2015.11.015

Xu X., Villafañe S. M., and Lewis J. E., (2013), College students' attitudes toward chemistry, conceptual knowledge and achievement: structural equation model analysis, *Chem. Educ. Res. Pract.,* **14**(2), 188-200. DOI: 10.1039/C3RP20170H

Yosso T., (2005), Whose culture has capital? A critical race theory discussion of community cultural health, *Race Ethnic.Educ.,* **8**(1), 69-91. DOI: 10.1080/1361332052000341006

CHAPTER 5:

FROM DEFICIT MINDSET TO ASSET-BASED THINKING: AN EXPLORATION OF HISPANIC FEMALE STUDENTS' ATTITUDES TOWARD CHEMISTRY IN A FIRST SEMESTER ORGANIC CHEMISTRY COURSE

Introduction

When STEM students were asked to identify their most difficult courses in the undergraduate curriculum are, their answer typically included organic chemistry (Rowe, 1983; Barr *et al.*, 2010; Horowitz, Rabin, and Brodale, 2013). Organic chemistry has a reputation of being the most feared and failed course for undergraduate STEM majors (Grove, Hershberger, and Bretz, 2008; Flynn, 2015). This course's high level of difficulty renders it a gatekeeper course for STEM career paths (Seymour and Hewitt, 1997; Gasiewski *et al.*, 2012). Since organic chemistry is required for many STEM and health professions (Barr *et al.,* 2008; Cooper *et al.*, 2010), it is high-stakes. Much of the research on organic chemistry education has focused on understanding the difficulties that students face which might prevent them from succeeding in this course (Cooper *et al.*, 2010; Grove and Bretz, 2010; Kraft, Strickland and Bhattacharyya, 2010; López *et al.,* 2014; Anzovino and Bretz, 2015; Dood *et al.*, 2019; Crandell, Lockhart and Cooper, 2020), and persisting onto subsequent chemistry courses (Anderson and Bodner, 2008). However, very few

studies have focused on studying affective measures in this course (*i.e.*, Black and Deci, 2000; Liu, Raker, and Lewis, 2018; Rocabado *et al.*, 2019).

Taber (2015) argues that it is important to take not only a cognitive approach to teaching and learning, but also an affective approach; meaning that investigations ought to highlight the feelings and perceptions of students as they engage in learning experiences. This notion of investigating the affective domain is especially meaningful in courses such as organic chemistry where students experience a high-demand, high-stakes environment where they might feel stressed, anxious, or overwhelmed throughout the course. Recently, Flaherty (2020) examined 91 studies in the chemistry education research field that investigated affect as it relates to student performance and concluded that it is important to focus on influencing students' affect and not just on the improvement of course performance. Many researchers have focused on establishing links between affect and outcomes. Perceived belonging has been shown to predict achievement and attrition in chemistry courses (Fink, Frey, and Solomon, 2020). Similarly, attitude has also been found to impact achievement over and above prior conceptual knowledge (Xu, Villafane and Lewis, 2013). Halpern *et al.* (2007) concluded that attitudes toward STEM fields have an influence on students' decisions to pursue and continue in their science tracks. Thus, as shown with these studies and many others, the investigation of affective measures is an integral part of the ongoing field of chemistry education research.

In this study we chose to focus on student attitudes toward chemistry in a first semester organic chemistry class by measuring their *intellectual accessibility* (IA) and *emotional satisfaction* (ES) utilizing the Attitude toward the Subject of Chemistry Inventory version 2

(ASCIv2; Xu and Lewis, 2011). The ASCIv2 is a shortened instrument refined from the original ASCI developed by Bauer (2008). It is a 7-point semantic differential scale, and it has been used across the United States (Brandriet *et al*., 2011; Xu and Lewis, 2011; Brandriet, Ward, and Bretz, 2013; Xu, Villafañe, and Lewis, 2013; Cracolice and Busby, 2015; Chan and Bauer, 2014, 2016; Mooring *et al*., 2016; Underwood, Reyes-Gastelum, and Cooper, 2016; Stanich *et al*., 2018; Nenning *et al.,* 2019; Rocabado *et al*., 2019), in several other places in the world (Xu, Southam, and Lewis, 2012; Xu, Alhoosani, Southam, and Lewis, 2015; Vishnumolakala *et al*., 2017; Vishnumolaka *et al*., 2018; Damo and Prudente, 2019) and in different languages (Khaveci, 2015; Sen, Yilmaz, and Temel, 2016; Montes, Ferreira, and Rodriguez, 2018). Since its development, researchers and practitioners have used this instrument in their classrooms for a variety of reasons, including the comparison in attitudes of different demographic groups (Rocabado *et al*., 2019). Yet, the majority of the studies that employed the ASCIv2, particularly those which examine attitude over time, investigate the influence on attitude of various pedagogical interventions in comparison to traditional methods (Mooring *et al*., 2016; Underwood, Reyes-Gastelum and Cooper, 2016; Stanich *et al*., 2018; Vishnumolaka *et al*., 2018). Thus our study supplements the literature with additional information about student attitudes toward chemistry at a baseline level, namely without an intervention. This addition to the literature will aid future researchers to gauge the impact of their work to foster positive student attitudes.

To assist in establishing a basis for comparisons among studies, a practical method of investigation is to conduct a meta-analysis (Smith and Glass, 1977; Lipsey and Wilson, 2001). To date, few meta-analysis studies exist that examine solely the field of chemistry education. Warfa (2016), Leontyev *et al.* (2017), and Raman and Lewis (2019) report on the effectiveness of various

pedagogical interventions in chemistry classrooms utilizing meta-analyses. The present study will contribute a chemistry-specific meta-analysis focused on the use of the ASCIv2 across one semester of instruction. The purpose of this meta-analysis is to capture the range of attitude change within a course that has been observed in the literature, both with and without a pedagogical intervention. Should researchers decide to use the ASCIv2 in their classrooms, these meta-analytic results will be a valuable gauge to frame their own observations. Another contribution of this meta-analysis is to provide a reference point when investigations focus on specific demographic groups of students (*i.e.*, underrepresented minority students). Determining differences in attitude changes between the group as a whole and a specific subgroup can lead to a deeper understandings of chemistry classrooms.

With respect to examining disaggregated results for students from minoritized groups, studies have shown that gender, race, and ethnicity correlate with student perceptions of their educational investments (Fordham and Ogbu, 1986; Baumgartner and Johnson-Bailey, 2008; Banerjee *et al*., 2018), their beliefs about their abilities, and their attitudes toward STEM subjects (Catsambis, 1995; Else-Quest, Mineo and Higgins, 2013; Leslie *et al.,* 2015). Women and girls tend to have lower confidence, more negative attitudes, and greater anxiety in science and math than their male counterparts (Else-Quest, Hyde, and Linn, 2010; Desy, Peterson and Brockman, 2011). Furthermore, students who possess intersectional (Crenshaw, 1989; Litzler, Samuelson and Lorah, 2014; Ireland *et al.*, 2018) minoritized identities, such as Hispanic women, characterize "the double bind" (Ong *et al*., 2011) and may experience in greater depth the effects of negative attitudes toward science. Studies focusing on African American, Hispanic, or Native American (AHN) female students in organic chemistry in the United States are few (*e.g.*, Rocabado *et al*.,

2019); thus, the need for basic research with this focus. Likewise, in countries across the world subgroup comparisons are also few and mostly focused only on gender. For instance, Salta and Koulougliotis (2015) found that girls displayed higher science self-determination than boys, as well as higher science career and intrinsic motivation. In Katsina State, Nigeria, encouraging results that point toward gender equity were found when male and female students science achievement and attitude toward science were compared (Olasehinde and Olatoye, 2014). While these results are promising, they do not address intersectionality (Crenshaw, 1989). Paying attention to students who embody multiple minoritized identities simultaneously in STEM is crucial (Else-Quest, Mineo and Higgins, 2013). This practice aims to illuminate some of the unique challenges for members of intersectional groups which might otherwise be missed. Villafañe, Garcia, and Lewis (2014) conducted a study in which self-efficacy trends were compared between different intersectional groups in a college preparatory chemistry course by measuring this construct several times during the semester. In the present study we investigated the attitude trends of Hispanic female students, with particular interest in placing this trend in the context of our meta-analytic results. This study also considers and problematizes a comparison of attitude trends between Hispanic female and White female students in an organic chemistry course.

One of the major topics of investigation in STEM education that is magnified for AHN students is the issue of retention. For researchers interested in gatekeeping courses such as organic chemistry, retention rates are relevant, since it is at this point that many students decide to leave STEM for good (Zoller, 1990; Seymour & Hewitt, 1997; Grove and Bretz, 2010). While the metaphor of a "leaky pipeline" in STEM fields at all educational levels (Barr, Gonzalez and Wanat, 2008) has been contested (Cannady, Greenwald and Harris, 2014), there is no doubt that

researchers remain concerned about retention. Chen (2014) conducted a national study utilizing Postsecondary Education Transcript Study (PETS) data collected between 2003 and 2009 and concluded that between 48% and 69% percent of students in the United States that intended to major in STEM fields had left these majors some time at the bachelors' or associate's level, respectively. Of these students who left STEM fields, about half switched to fields outside of STEM, and the other half dropped out of school altogether without earning a college degree (Chen, 2014). In a pivotal study, Seymour and Hewitt (1997) investigated reasons why students decide to leave their STEM career paths and concluded that negative emotions play a significant part in these decisions. More recently, Seymour and Hunter (2019) report from a follow-up study that the leading contributors to this decision remain affective in nature, such as feeling overwhelmed, or experiencing a loss of interest in major.

In an effort to characterize the motivations behind leaving STEM fields, Geisinger and Raman (2013) conducted a broad literature search and reported important factors related to students' decisions to leave engineering majors. Some of those factors are classroom climate, conceptual understanding, self-efficacy, and demographic background such as race and gender (Geisinger and Raman, 2013). Other studies have also found that attrition in STEM courses is high, yet there is a difference between men and women's retention, and this difference depends on the major (Rask, 2010). Largely, however, the attrition rates of female and ethnic/racial minority students in STEM disciplines have been disproportionately high relative to that of White males (Seymour and Hewitt, 1997, Tsui, 2007; Barr *et al.*, 2008; Seymour and Hunter, 2019). Studies that have investigated the role of affect in STEM retention have concluded that when students hold positive emotions, these can influence their decision to persist in their chosen careers (Wyer, 2003;

Carlone and Johnson, 2007; Simon *et al*., 2015). Simon and colleagues (2015) found that male students thrived in autonomy-supporting environments, while female students were more likely to perform better and persist in STEM-related courses when they displayed high levels of self-efficacy. Carlone and Johnson (2007) point out that although all students in STEM experience difficult situations, a well-developed and recognized science identity is a strong determinant of persistence for Women of Color. These promising and varied findings related to affect suggest that continuing efforts to investigate affect and STEM retention for students of diverse backgrounds are timely and critical.

Researchers have studied retention with and without interventions (*e.g.*, Grove, Hershberger, and Bretz, 2008; Mitchell, Ippolito and Lewis, 2012; Chen, 2014; Sloane, 2016; Xu, 2016; Fink, Frey and Solomon, 2020), but in many cases investigations of affective measures alongside retention measures were missing. Seymour and Hunter (2019) suggested that retention studies can benefit from an affective lens to explain why students leave even when pedagogical interventions are in place. Studies that have employed the ASCIv2 have investigated the relationships between attitude and performance (*i.e.*, Xu, Villafañe and Lewis, 2013; Mooring *et al*., 2016; Rocabado *et al*., 2019); however, to our knowledge, none have explored retention. Thus, this study aimed to investigate the relationship between attitude and retention for Hispanic and White female students in an organic chemistry course.

Initial Research Questions

This study examined how the attitudes of Hispanic female and White female students in three sections of an Organic Chemistry I (OCI) course in fall 2018 compared to the attitude trends observed in the literature and compared these trends for these two groups of students within the course. We originally designed the study with this purpose in mind, and to that end, we posed two formal research questions:

1. How does the change in attitude of Hispanic female and White female students in OCI compare to the changes in attitude seen in the literature as measured by the ASCIv2?
2. How do attitude trends of Hispanic female students in OCI compare to attitude trends of White female students in OCI?

Telling a Mindset Change Story - Challenging Our Deficit Mindset

When we originally imagined this study, we set out to collect data that would allow us to measure attitude for Hispanic and White female students in organic chemistry. We have explained why we think it important to investigate attitudes of students representing different demographic intersections. During the course of our work we realized that we were utilizing a deficit mindset (Yosso, 2005; Harper, 2010; Gorski, 2011; Yep, 2014). At the conclusion of our initial analyses, we noticed a problem with our approach. Drawing on our understanding of Critical Race Theory

(CRT), and particularly QuantCrit (Solórzano, 1997; Solórzano and Delgado Bernal, 2001; Yosso, 2005; García, López and Vélez, 2018; Gillborn, Warmington and Demack, 2018), we recognized that we needed to think more carefully about how to respect the students from whom we had collected the data.

By checking our approach in this way, we worked to shift our deficit mindset by broadening the lens of our investigation. Walls (2016) has suggested that in order to consider equity in our studies, our research questions should allow the researcher to examine racial/ethnic groups in light of their broader representation within their environments. This suggestion points to, for instance, considering the overall representation of racial/ethnic groups within the university as a whole. Broadening our investigation from the classroom to university, allows for greater understanding of the context that shapes the racial make-up of our classrooms. With this perspective we can begin to appreciate issues of persistence. Persisting can be examined via patterns of enrollment at the institution, enrollment in a target course, and enrollment in the next course in the STEM-major chemistry sequence as well as via traditional examination of patterns in drop rates in a target course.

Theoretical Framework

Since the late 1990's, Critical Race Theory (CRT) has become a guiding framework in the study of race and racial disparities in the pursuit of social justice. In several fields of study, including education, CRT has imparted a lens by which to study marginalized groups (Crenshaw,

1995; Solórzano, 1997, 1998; Delgado and Stefanic, 2001; Solórzano and Delgado Bernal, 2001; Yosso, 2005; Dixson and Anderson, 2018). CRT aims to empower People of Color within marginalized spaces by providing a channel for their voices, and to explore and challenge the ways in which racism permeates social structures (Solórzano, 1997; Yosso, 2005). Although CRT has largely been applied in qualitative work, a few studies have also employed quantitative methodology to investigate issues concerning racial disparities (Solórzano and Ornelas, 2004; Pérez Huber, Vélez, and Solórzano, 2017; Gillborn, Warmington, and Demack, 2018; Baker, 2019; Campbell-Montalvo, 2020). As an emergent branch of CRT developed for the use of quantitative methodologies, QuantCrit expresses principles of CRT in a manner that is intended for direct uptake within quantitative studies. QuantCrit acknowledges that racism is central to society and enhances subordination, recognizes that 'numbers are not neutral' and challenges deficit ideology by promoting counterstorytelling. QuantCrit asks researchers for a commitment to critically scrutinize the forms of analyses we utilize in the pursuit of social justice, stressing that 'data does not speak for itself' and encouraging the collection and understanding of experiential knowledge about the individuals and groups in our studies. Ultimately, QuantCrit argues that statistical analyses need to be carefully carried out to explore wider structural issues with the intention to avoid the propagation of social inequities (García, López and Vélez, 2018; Gillborn, Warmington and Demack, 2018).

QuantCrit has provided an important framework for this study. Although we begin this article by presenting the data collected and analyses performed as intended from the beginning, additional research questions emerged as we worked to adhere to the principles of QuantCrit and internalize them to shape the study. We have formatted this paper in two parts, as we experienced

it, because we desire to take the readers through our evolution as an example of the mindset shift that we hope many researchers experience as they encounter CRT and QuantCrit.

Additional Research Questions

As previously described, additional questions emerged during our analysis process. The second part of the study explores two main ideas. The first one is investigating the overall representation of all racial/ethnic student groups broken down by binary gender, paying particular attention to Hispanic and White female students in OCI and in the university throughout the fall 2018 semester. Additionally, we were interested in exploring drop rates in OCI for these groups as well as the persistence into the next course of the sequence (Organic Chemistry II or Biochemistry). We posed four additional research questions:

3. How does the population of Hispanic female and White female students in OCI compare to the population of these groups at the University throughout the fall semester of 2018?
4. How do the drop rates compare for Hispanic female and White female students in OCI?
5. To what extent do Hispanic female and White female students persist to the next chemistry course in the sequence immediately after passing OCI?
6. What difference did it make to adhere to the QuantCrit tenets in our analyses?

Methods

Students in three sections of an OCI course taught by the same instructor in a large southeastern public research university were given the ASCIv2 prior to each midterm exam in the fall semester of 2018. Data collection in this course followed an IRB approved protocol. The survey was administered via Qualtrics. Students received an email with the link to the survey two days before their exam, and two additional reminder emails were sent for the students who had not yet completed it. The survey was open until the time students were scheduled to take their exam. They were incentivized to complete the survey with two extra-credit points (2% of the grade) toward each of the midterm exams and the final exam. They were not penalized if they did not answer every survey item; however, they were forewarned that completed surveys with no responses would not be awarded the extra-credit points. Student demographic data, such as race/ethnicity, binary gender, and major were collected from the university records and not at the time of the survey nor at any time during the course.

The sections of the OCI were taught by a single instructor on the same two days of the week (*i.e.,* Tuesday and Thursday) and students were also required to attend a recitation session on Friday at a designated time taught by a graduate teaching assistant. The three sections of OCI shared the same syllabus indicating that assignments and exams were the same across the sections. Although no formal pedagogical intervention was implemented in this course, the instructor used clicker questions regularly as formative assessments to keep the class engaged, especially when a concept seemed to not be fully grasped during the lecture. Additionally, the instructor used other

devices to help students learn and remember the material such as 'functional group sudoku,' counting songs, visualization exercises, and other mnemonics.

In this course, the first three midterm exams represent the exams before the withdrawal date in the semester. Although there were four midterm exams and a final exam in this course in the fall semester of 2018, we present only the first three pre-exam survey data before the withdrawal date to represent the attitude trends for the majority of students enrolled in the course including those who eventually withdraw from the course.

Descriptive Statistics

For each of the groups of students in this study (Hispanic female and White female), we computed the mean, standard deviation, skewness, and kurtosis values for each item in the ASCIv2 at each of three pre-exam survey administrations during the semester. We calculated the observed mean scores for the *intellectual accessibility* (IA) and *emotional satisfaction* (ES) factors by taking the average of the item means that belong to each factor. Item 1, 4, 5, and 7 means were reverse coded for ease of interpretation. Thus, negative adjectives are found on the lower end of the scale (*i.e.*, 1), and positive items are found on the upper end of the scale (*i.e.,* 7), with 4 indicating neutral. Thus, if a mean score is less than 4 for items in the IA subscale, it means, for example, that a group of students found chemistry to be hard or confusing, rendering it less intellectually accessible. Similarly, if a mean score is greater than 4 in the ES subscale, it means, for example, that students found chemistry to be pleasant or organized, thus rendering it more emotionally

satisfying. Tables S5.1-S5.3 in Appendix C present the descriptive statistics for the ASCIv2 administered for two days immediately before each exam for Hispanic and White female students. Additionally, the observed mean scores are found in Table 5.1.

Meta-Analysis of ASCIv2 Longitudinal Studies

Meta-analysis can be conceptualized as a way of surveying research studies that have a common theme operationalized consistently among the studies (Lipsey and Wilson, 2001). This systematic methodology of reviewing a chosen theme can be a powerful tool to inform the audience about the extent of impact of this theme within a field. In this article, we have followed the steps to conduct a meta-analysis that were delineated in Rahman and Lewis (2019).

We began by conducting a systematic search of articles that have used the ASCIv2 from the time the instrument was first published in 2011 (Xu and Lewis) through 2019 by first employing the university's libraries general search and typing the different ways the name of the instrument can be reported (*e.g.*, ASCIv2, or Attitude toward the subject of chemistry inventory version 2). This search yielded 14 articles including the original. The next search was done utilizing *google scholar* investigating the articles that cited Xu and Lewis (2011). 24 additional articles were retained from this search with the criteria that only articles that utilize the instrument name (in any form) or the word "attitude" appeared in the title and/or the abstract. A few articles pulled from this search had the word "attitude" in the title or abstract, but clearly listed a different instrument used to measure attitude, therefore these articles were excluded from this initial search.

Finally, one additional exclusion criterion was necessary at this stage, which applied to the few instances where the ASCIv2 was utilized in other fields (*i.e.*, Mathematics) as a tool to measure attitude toward fields other than chemistry.

A total of 38 studies including the original (Xu and Lewis, 2011) were further scrutinized. Additional inclusion criteria were determined to further refine the data corpus to conduct the meta-analysis. Our decisions were made following the condition that the instrument had to be used at least twice in a longitudinal study (*i.e.*, pre-post). From 38 possible articles, we went down to seven articles that met our criteria, however, the data from one of those articles was not independent from another article. Therefore, we removed the article that utilized a subset of data found in another article, leaving six articles in total. These six articles were used in the meta-analysis of longitudinal studies that utilized the ASCIv2. Due to our interest in informing about attitude change in chemistry classrooms, it became apparent we had to make a distinction between classrooms that utilized pedagogical interventions of some kind, such as flipped classroom, and other classrooms that employed traditional pedagogies such as lecture or face-to-face instruction. We conducted two meta-analysis, one for no intervention or control classrooms, and one for intervention or treatment classrooms to inform the readers about the attitude changes that are observed in the different classroom settings. There were three articles that had more than one group in the same category. When carefully examining the groups, we determined these groups were independent from each other. For example, Mooring *et al.* (2016) reports four groups, two traditional lectures and two intervention classes. We determined that the two traditional classrooms were independent since they were two different courses (Organic chemistry I and Organic chemistry II from the same semester and taught by different instructors). The same logic followed for the intervention classes.

Additionally, these separate results can help readers situate their study within one of the two conditions and gauge whether their attitude change observations fall within the range reported in the literature.

To be able to conduct a meta-analysis of the six articles, we decided to calculate the *Standardized Mean Gain* effect size, which "involves the same operationalization of the variable at both times of measurement for each sample" (Lipsey and Wilson, 2001, pp. 44). Equation 5.1 is the equation for the standardized mean gain effect size, where $\bar{X}$ is the mean value reported at time one (T1) or time two (T2). Equation 5.2 is the corresponding equation to calculate standard error, where *r* is the Pearson product-moment correlation (see equation 5.3) between $\bar{X}_{T1}$ and $\bar{X}_{T2}$ and *n* is the sample size (Lipsey and Wilson, 2001).

[Eq. 5.1]

$$ES_{sg} = \frac{\bar{X}_{T2} - \bar{X}_{T1}}{SD_{pooled}}$$

[Eq. 5.2]

$$SE_{sg} = \sqrt{\frac{2(1-r)}{n} + \frac{ES_{sg}^2}{2n}}$$

[Eq. 5.3]

$$r = \frac{\Sigma\,(X_1 - \bar{X}_{T1})(X_2 - \bar{X}_{T2})}{\sqrt{\Sigma\,(X_1 - \bar{X}_{T1})^2 \Sigma\,(X_2 - \bar{X}_{T2})^2}}$$

Confirmatory Factor Analysis

Confirmatory factor analysis was conducted using Mplus (Version 8.2; Muthén and Muthén, 1998-2007) for Hispanic female and White female students separately to check whether the internal structure encompassed the same organization for these groups in each course (see Tables S4 and S5 in the ESI; Rocabado *et al*., 2020). The two-factor model established for this instrument was utilized (Xu and Lewis, 2011) in order to gather internal structure validity evidence (Arjoon, Xu and Lewis, 2013; AERA, APA and NCME, 2014). The seven-point scale data were treated as continuous and a maximum likelihood robust (MLR) estimator was used. This estimator takes into account non-normally distributed data (Cheng-Hsien 2016) with skewness and kurtosis values greater than ± 1.00 (Bulmer, 1979). Missing data values in the data were handled using full-information maximum likelihood (FIML) estimation, which is the default when using a maximum likelihood estimator in Mplus (Klein and Moosbrugger, 2000; Muthén and Asparouhov, 2003). Thus, we utilized all the data obtained without resorting to listwise or pairwise deletion when missing values were found. The model parameters were estimated by fixing the first factor loading on each factor to 1.00 and allowing all of the other loadings, variances and covariances to be freely estimated. Model fit statistics were used to determine whether the data fit the model well. To evaluate model fit we examined the chi-square (χ^2) value. The χ^2 is highly influenced by sample size; thus it becomes critical that we examine additional fit indices, such as, the comparative fit index (CFI), the standardized root-mean square residual (SRMR) and the root mean square of approximation (RMSEA) provided by the software (Brown, 2006). The accepted cutoff criteria for these fit indices are as follows: for CFI > .90 is acceptable, but best if > .95; for SRMR < .08; and for RMSEA < .06 (Hu and Bentler, 1999). The RMSEA has been shown to produce unpredictable

results with a short instrument due to fewer degrees of freedom (Kenny, Kaniskan and McCoach, 2015); therefore, RMSEA values will be provided with the caveat that these values are inconsistent and may lead to idiosyncrasies in interpretation.

Reliability

Reliability of scores was also calculated for each factor, for each group, and at each time point (see Tables S5.4 and S5.5 In Appendix C). Cronbach's alpha is often a reported measure for reliability (Cortina 1999; Cronbach 1951), which is a measure of how closely related the items within a factor are. However, this coefficient works under the assumption that the model in the study is a *tau*-equivalent or essentially *tau*-equivalent model (Komperda, Pentecost and Barbera 2018). Essentially "tau-equivalent" means that the measurement model assumes equal factor loadings for each item in the factor. Rarely a measurement model assumes this type of constraint, thus the model used in this study is not *tau*-equivalent. Rather, we utilize a congeneric measurement model, in which factor loadings, intercepts, and all other parameters are freely estimated. Therefore, following Komperda and colleagues' (2018) suggestion, a more appropriate measure of reliability is coefficient Omega. This coefficient is directly calculated using the parameter estimates obtained from the output of confirmatory factor analysis and it is interpreted much like Cronbach's alpha, with higher values (closer to one) indicating high reliability so long as the model fits the data well. The equation used to calculate the omega coefficient of reliability is as shown in Equation 5.3, where lambda (λ) represents the standardized factor loadings and theta (θ) represents the error variances.

[Eq. 5.4] $$\omega = \frac{(\Sigma\lambda)^2}{(\Sigma\lambda)^2 + \Sigma\theta}$$

Measurement Invariance Testing

Measurement invariance testing was performed for the configural, metric, scalar, and strict models comparing Hispanic female and White female students within the OCI course for the 2-factor model before the first three midterm exams. Additionally, a longitudinal analysis was also performed utilizing measurement invariance testing. Following the steps delineated by Rocabado and colleagues (2020), we began with confirmatory factor analysis for each group and at each time point (see Tables S5.4 and S5.5). Then, moving on to the models in measurement invariance testing, the configural invariance model is the least constrained. In this model, only the pattern of fixed and freely estimated factor loadings must be the same for both groups. If fit indices are within the range of acceptable values, the configural model is considered invariant. The next step is to impose a more rigorous constraint: metric invariance is tested by fixing the factor loadings to be the same for both groups. If the fit indices are not significantly different from those for the configural model, the metric model is considered invariant. Finally, even more stringent constraints are imposed, with scalar invariance tested by extending the constraints to equal thresholds (intercepts) for each item. The fit indices produced by the scalar model are compared to those for the metric model (Sass, 2011; Rocabado *et al*., 2020). At this point, if scalar invariance is established, factor scores can be compared between groups (Rocabado *et al*., 2020). However, one additional constraint is required if we desire to compare observed mean scores, thus advancing to the strict model requires the item error variances to be equal among groups (Sass, 2011) The

process of establishing strict invariance is the same as with the previous levels of invariance testing. Based on Chen (2007) we evaluated $\Delta\chi^2$; however, as noted previously, this value is highly influenced by sample size (Brown 2006). Therefore, in addition, we evaluated our results based on the following fit index cutoffs: ΔCFI (< .01), ΔSRMR (< .03), and ΔRMSEA (< .015) for metric invariance, and ΔCFI (< .01), ΔSRMR(< .01), and ΔRMSEA (< .015) for scalar and strict invariance (Chen, 2007). Once measurement invariance is established, a comparison of attitude scores between the groups can provide meaningful results. Observed mean scores were calculated and compared between Hispanic female and White female students as well as longitudinal comparisons (see Tables S5.6-S5.9).

Multilevel Modeling and Effect Size Comparisons

Multilevel modeling (MLM) is a statistical technique that combines aspects of other significance testing analyses and is appropriate for use when there is one dependent variable, one or more continuous predictors, and one grouping variable in which, in this case, students are nested within test occasions. A longitudinal MLM is used when repeated measures are nested within participants over time. Additionally, a level 2 prediction model can be utilized to analyze data not only at the individual level (level 1), but also at the group level (level 2; Harlow, 2014; Heck and Thomas, 2015). In this study we investigated the longitudinal changes of IA and ES over time (level 1), and whether these changes are significantly different based on group membership (*i.e.*, Hispanic females and White females) utilizing longitudinal and level 2 prediction MLM.

Effect size is a useful tool when performing comparisons between groups or within groups. Cohen (1988) explained that the effect size is another metric by which to test the null hypothesis when conducting statistical group comparisons. Often the null hypothesis implies that there is no evidence of difference between two groups. Cohen (1988) explains that the null hypothesis is "the circumstance in which differences in the independent variable… have no effect (have an effect size of zero) on the means or proportions of the dependent variable" (pp. 9). By this definition it is suggested that when there is a deviation from the null hypothesis, one way to quantify the degree to which this deviation is present is by calculating an effect size (Cohen, 1988). Several advantages exist for calculating the effect size instead of, or in addition to, statistical significance tests. One, as mentioned previously, is that effect size deals with the magnitude or 'the degree' to which the null hypothesis is false (Cohen, 1988; Lipsey and Wilson, 2001). On the other hand, significance testing only informs whether or not to reject the null hypothesis. Two is the effect size has a valence, thus becoming a vector, indicating both the magnitude and direction of the difference when performing comparisons (Lipsey and Wilson, 2001).

Results

All underlying assumptions for the analyses performed in this study were either met or addressed. For example, the normality assumption was violated for several items with skewness or kurtosis values outside of the acceptable range of ±1.00 (Bulmer, 1979; see Tables S5.1-S5.3). To address the violation of normality in our study, we used the MLR estimator, which employs a statistical correction of standard errors and Chi-square statistics (Cheng-Hsien, 2016). All analyses

terminated normally and convergence was attained for all measurement models included in this study.

Descriptive Statistics

On Table 1 we see that the observed mean scores for each factor are different for each group of students under investigation; however, it is apparent that the trend is similar. Figures 5.1a and 5.1b exhibit a downward trend for IA and ES for both groups of students. Although the scale of the instrument is 1 to 7, we chose to zoom-in to the relevant 1.2-point sections of these graphs where the means are found (*i.e.,* 2 - 3.2).

Table 5.1. Observed Mean Scores for Hispanic and White Female Students in OCI

	White Female		**Hispanic Female**	
	Mean	**S.D.**	**Mean**	**S.D**
Intellectual Accessibility				
Pre-Exam 1	2.93	1.36	2.78	1.31
Pre-Exam 2	2.87	1.38	2.58	1.40
Pre-Exam 3	2.68	1.37	2.38	1.34
Emotional Satisfaction				
Pre-Exam 1	3.84	1.62	3.67	1.66
Pre-Exam 2	3.69	1.59	3.40	1.74
Pre-Exam 3	3.29	1.70	2.90	1.65

Through the descriptive statistics shared in Table 5.1 and Figures 5.1a-b, we can see that although the downward trend in IA and ES is shared by all students throughout the course, the White female students display higher levels of attitude from the beginning compared to the Hispanic female students. We see that at the beginning of the semester the IA and ES levels are

similar, but toward the end of the semester the attitude gap widens favoring the White female students.

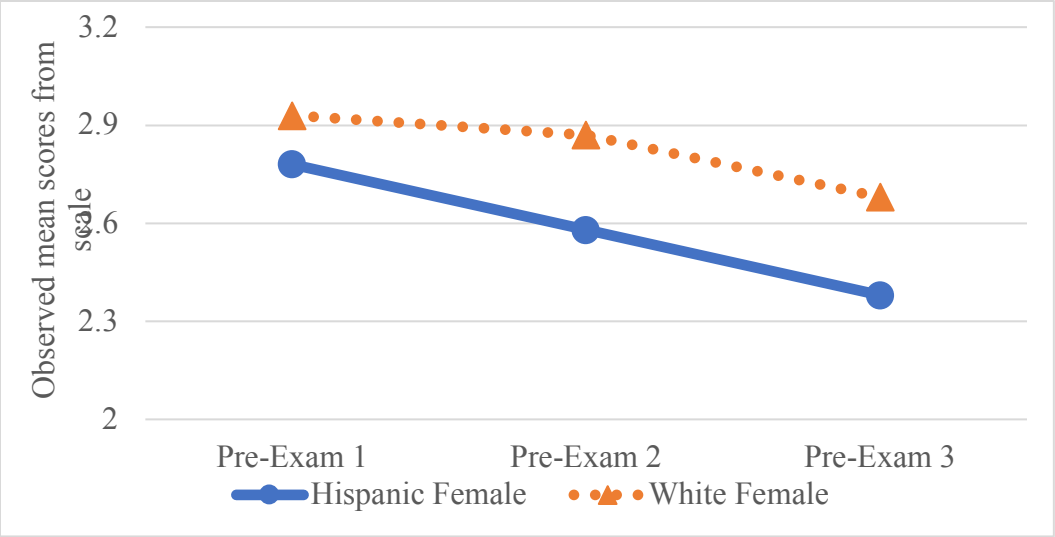


Figure 5.1a. Observed mean score comparison between Hispanic female and White female students in Organic Chemistry I in Fall 2018. These comparisons are for the *Intellectual Accessibility* (IA) subscale in the ASCIv2. The graph displays a downward trend of IA throughout the semester for both groups.

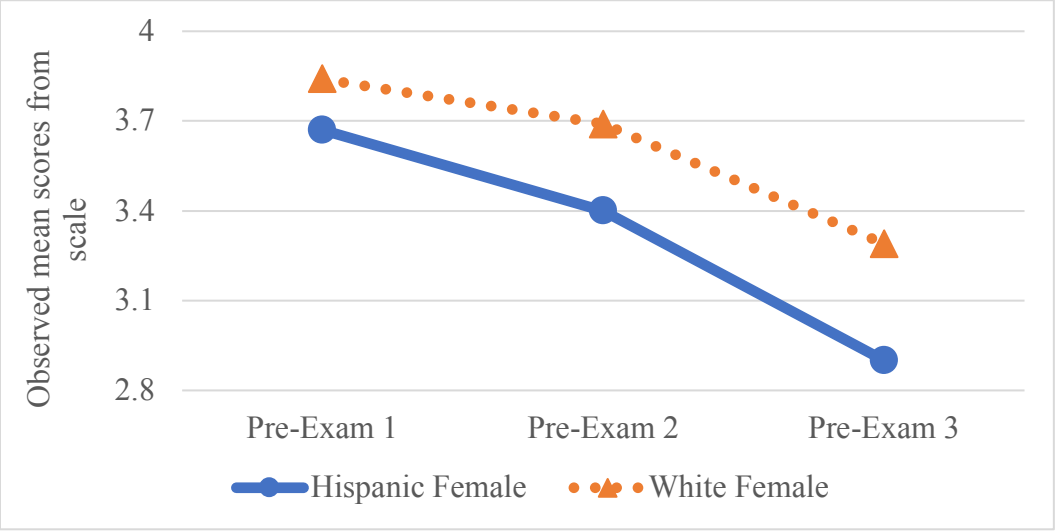


Figure 5.1b. Observed mean score comparison between Hispanic female and White female students in Organic Chemistry I in Fall 2018. These comparisons are for the *Emotional Satisfaction* (ES) subscale in the ASCIv2. The graph displays a downward trend of ES throughout the semester for both groups.

Meta-Analysis of ASCIv2 Longitudinal Studies

For the meta-analysis portion of this study we included six articles that utilized the ASCIv2 in longitudinal research in chemistry classrooms. Tables S5.10 and S5.11 summarize the findings from this analysis for each of the dimensions measured by the ASCIv2, namely IA and ES. Figure 5.2 displays four plots of the calculated effects sizes for the change in attitude throughout the semester of all studies examined including the effect size calculated for the entire OCI class in the present study. In these plots, we have also added the effect size observed for the subgroups of students we have focused on in this study, although these were not part of the meta-analysis. The plots are divided by the two affective dimensions in the ASCIv2, namely IA and ES, as well as intervention and no intervention groups. Figure 5.2a displays the plot for no intervention (control) groups in the IA dimensions, and Figure 5.2b displays the intervention (treatment) groups in the IA dimension. Similarly, Figures 5.2c and 5.2d display the effect sizes found in the ES dimension for no intervention and intervention, respectively. With the addition of the effect sizes for the subgroups of focus to the plots in Figures 5.2a and 5.2c, we can see that the change in attitude for subgroups within a classroom can be quite different from the effect size of the classroom as a whole. We can also clearly observe that although both groups of students' IA and ES drop, the effect size of the drop is greater for the Hispanic female students than for the White female students. Additionally, Tables 5.2 and 5.3 contain the overall average random weighted effect size for IA and ES, respectively.

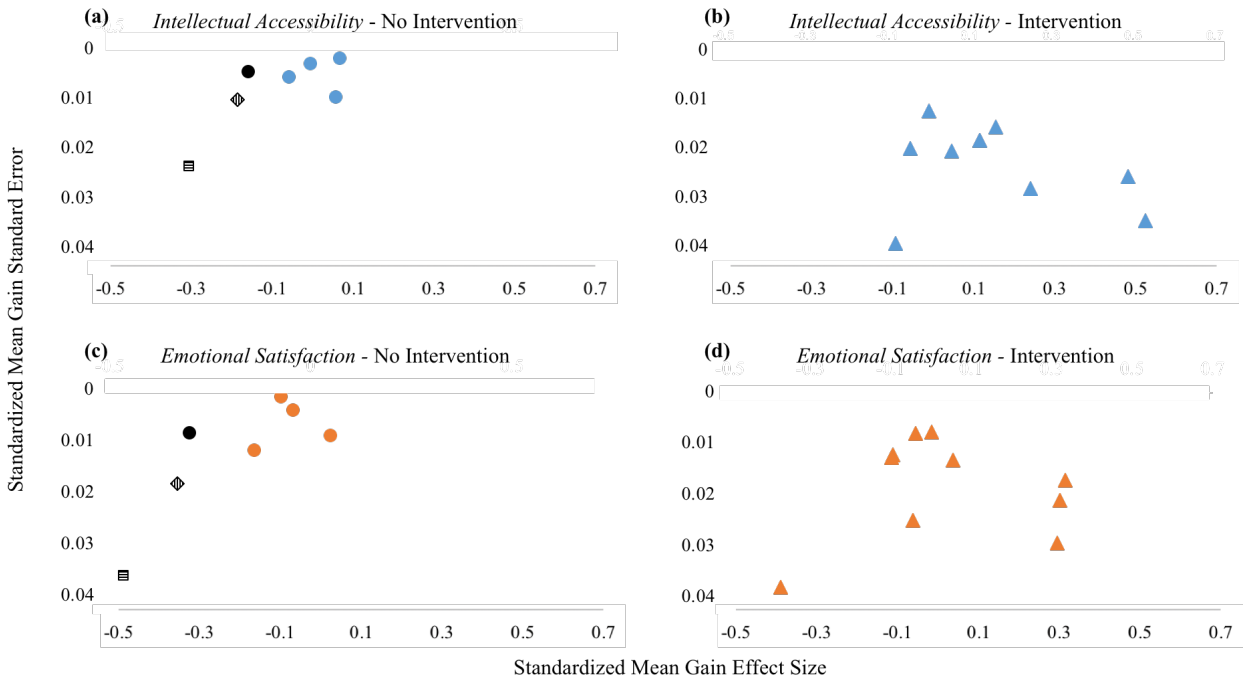


Figure 5.2. Plot of effect size values. No intervention groups (plots a and c) effect sizes in meta-analysis are represented by solid circles (blue for IA and orange for ES). The black circles represent the effect size observed for the entire OCI course in this study and is part of the meta-analysis. The patterned diamond represents the effect size for the White female students and the patterned square represents the effect size for the Hispanic female students. In plots b and d the filled triangles are for intervention groups, blue for IA and orange for ES.

The overall average random weighted effect sizes for each group were calculated taking into account the groups within each study that corresponded to 'no intervention' or 'intervention' and for each affective dimension. For the IA scale and no intervention the overall effect size is -0.01, which is considered negligible. And for the intervention group the overall effect size is 0.16 (see Table 5.2), which is considered negligible to small (Cohen, 1988). These results show first, that there are more intervention groups reported in the literature than no intervention studies, and second, that the intervention studies tend to show more positive effect sizes than the no intervention groups. However, the overall difference between intervention and no intervention is small. This result masks the range of effect sizes that are found in the literature, which may suggest that some

studies show a more positive trend in improving IA than others. IA appears to be a malleable trait that can be positively influenced by certain pedagogical interventions over others. The flipped classroom and the Process-Oriented Guided-Inquiry Learning (POGIL) pedagogies were the most impactful interventions in terms of positive gains over a semester, while other interventions such as "online" sessions were less impactful or showed no difference. The no-intervention classrooms consisted mostly of traditional lecture style teaching (for more information see Appendix C Table S5.10).

Table 5.2. Overall Effect Size of *Intellectual Accessibility* from Control or Treatment Groups

	Average Random Weighted Effect Size	Standard Error	Interval Upper	Interval Lower
No intervention (control)	-0.01	0.04	0.07	-0.09
Intervention (treatment)	0.16	0.05	0.26	0.06

The ES subscale yielded overall similar results in the sense that the intervention groups displayed more positive effect sizes than the no intervention groups. However, the average for both groups hovers over zero indicating that the changes for this construct, with or without intervention, are negligible. Additionally, when we observe the range of change for the ES scale within the studies, ES tends to decline to a greater degree than IA. The overall effect size for the ES subscale for the no intervention group is -0.10. Therefore in the absence of intervention there is a slight decline. And for the intervention group the effect size for ES is 0.03 (see Table 5.3). Although the intervention effect size is more positive than the effect size for no intervention, both of these effect sizes are considered negligible, indicating that while ES might also be a malleable trait than can be positively influenced by certain pedagogical interventions, during the course of a semester the magnitude of the change is not sufficiently large to overcome the general negative trend. Similar

to the result for IA, the interventions that were most impactful for ES were the flipped classroom and the POGIL classrooms.

Table 5.3. Overall Effect Size of *Emotional Satisfaction* from Control or Treatment Groups

	Average Random Weighted Effect Size	Standard Error	Interval Upper	Interval Lower
No intervention (control)	-0.10	0.04	-0.02	-0.17
Intervention (treatment)	0.03	0.05	0.13	-0.05

Confirmatory Factor Analysis

Given the descriptive results, we continued our study to investigate the effect size for the attitude drop observed for each group and situated this study within the report from the meta-analysis conducted. However, first we were required to ensure that the two-factor internal structure of the ASCIv2 as described by Xu and Lewis (2011) held for each group and that meaningful comparisons between the subgroups as well as longitudinal comparisons could be supported (Rocabado *et al*., 2019; Rocabado *et al.*, 2020).

A CFA was performed for the group of students in this investigation following the two-factor structure delineated by Xu and Lewis (2011). The data-model fit was not at the acceptable cutoffs initially; however, the data-model fit improved with the addition of modifications as seen in other studies where this instrument was used (Rocabado *et al.,* 2019) and all models achieved acceptable fit. The modification was found to be appropriate and was added to each of the models for both groups at each time point. The detailed process for the CFA can be found in Appendix C.

Reliability

Additionally, reliability was calculated using the McDonald's Omega value as described by Komperda *et al.* (2018). These values are best when they approach 1.000 and much like Cronbach's alpha (Cronbach, 1951; Cortina, 1999), values above 0.700 indicate good reliability. In Tables S5.4 and S5.5 we indicate the factor reliability at each time point and for each group. The values shown in these tables indicate strong reliability for each group with a range of reliability values of 0.751-0.866 for IA and 0.869-0.911 for ES.

Measurement Invariance Testing

Given the CFA results, we proceeded to conduct measurement invariance testing between White female and Hispanic female students and longitudinal comparisons. Although we investigated descriptive statistics in the previous section, we had not yet investigated whether comparisons between groups were supported. Measurement invariance testing gathers additional evidence to support the comparisons between these two groups as well as longitudinal comparisons at each level of testing (Gregorich, 2006; Sass, 2011; Rocabado *et al*., 2020). Tables S5.6-S5.9 in the SI indicate that comparisons are supported between the Hispanic female and White female students at each time point as well as longitudinal comparisons at the strict level, meaning that comparisons of the observed mean scores are supported.

Multilevel Modeling and Effect Size Comparisons

A longitudinal MLM analysis was explored to investigate level 1 individual effects on IA and ES across the semester. First, intraclass correlation values of 0.706 for IA and 0.520 for ES, which were > 0.05 were obtained from the level 1 model providing evidence that the repeated measures of IA and ES were nested within the students' test occasions over time and that MLM is an appropriate technique to investigate the downward trend in IA and ES for this group of students (Harlow, 2014). Next, we investigated whether the slopes of IA and ES changes over time were significant. Slopes of -0.333 for IA and -0.157 for ES were both significant indicating that Hispanic and White female students' attitude declined significantly throughout the semester in OCI. Finally, level 2 predictor variables were added to the model in which we examined the effect of group membership (*i.e.,* Hispanic female or White female) on the level 1 model described previously. For both IA and ES there was no evidence of statistically significant differences between these two groups.

When making comparisons and intending to measure the magnitude of the difference between scores, in this case attitude scores, a helpful value to investigate is the effect size (Cohen 1988; Lipsey and Wilson 2001). Table 5.4 displays the effect size of the IA and ES mean score comparisons using Hedge's *g*, a similar effect size calculation to Cohen's *d* (Cohen 1988), except more appropriate when the comparisons are between groups of different sample size (Hedges and Olkin, 1985). Following the criteria: $\geq$ 0.2 small effect size, $\geq$0.5 medium effect size, and $\geq$ 0.8 large effect size (Cohen, 1988) we can see that the comparison between Hispanic and White female students, although not significant as demonstrated previously, yield effect sizes that indicate the

differences might be negligible at the beginning of the semester; however, the differences in both IA and ES become more prevalent with a small effect size by exam 2 and later exam 3.

Table 5.4. Effect Size Using Hedge's *g* for *Intellectual Accessibility* and *Emotional Satisfaction* Observed Mean Score Comparisons Between Hispanic and White Female Students

	IA	**ES**
Pre-exam 1	0.111	0.104
Pre-exam 2	0.210	0.180
Pre-exam 3	0.221	0.232

Additionally, the longitudinal comparisons in IA and ES observed mean scores throughout the semester, from pre-exam 1 to pre-exam 3 are presented in Table 5.5. This time we utilize Cohen's *d* to compute the effect size since the sample size is consistent in these comparisons (Cohen, 1988). The drop observed for each group is noted with small to medium effect sizes, accompanying the statistically significant downward slopes reported previously. For Hispanic female students, a noticeable drop in IA and ES display effect sizes of small to medium, respectively. For White female students the drop in IA might not be noticeable (effect size <0.2), yet the drop in ES is noticeable with a small effect size.

Table 5.5. Effect Size Using Cohen's *d* for *Intellectual Accessibility* and *Emotional Satisfaction* Observed Mean Scores Longitudinal Comparisons for Hispanic and White Female Students

	IA	**ES**
Hispanic Female	-0.302	-0.465
White Female	-0.183	-0.331

Note. Cohen's *d* values are calculated between pre-exam 1 and pre-exam 3 for each group. The negative valence indicates a drop between pre-exam 1 and pre-exam 3.

Mindset Change Additional Findings

Given the previous findings we were concerned about Hispanic female retention to the next course given their less positive attitude toward chemistry than their White female peers. These findings were not surprising given the literature that suggests that Women of Color have less positive attitudes in STEM (Catsambis, 1995; Else-Quest, Mineo and Higgins, 2013), which might be a contributing factor in lower retention in these fields (Seymour and Hunter, 2019). After observing an attitude gap between our two groups in favor of White female students, we believed we were ready to compare retention for these groups. At this juncture in our analyses we realized that we had been structuring our study with a deficit mindset and were preparing to report that less positive attitude toward chemistry was a deficit that could lead to lower retention rates, etc. We were about to inform the readers that Hispanic female students in this course needed to be 'fixed' because of their deficit in attitude that could potentially be related to lower retention and overall underrepresentation in chemistry. This deficit story has been told countless times before, suggesting that somehow Hispanic female students 'lack' ability, or motivation, etc. (Bourdieu and Passeron, 1977; Sullivan, 2001). This story usually concludes with recommendations to 'fix' the students' deficiencies to match the central group standard (*i.e.,* White female students). However, we realized that these recommendations did not attempt to capture a bigger picture. Once we realized our narrow lens, we committed to more closely adhere to our theoretical framework by utilizing the tenets described in QuantCrit (Solórzano, 1997; Solórzano and Delgado Bernal, 2001; Yosso, 2005; García, López and Vélez, 2018; Gillborn, Warmington and Demack, 2018). We set out to explore our data further with the prospect to learn more about Hispanic female students in OCI in both representation and persistence to the next course. We followed suggestions

from the literature to consider the racial breakdown of the participants and their institutional enrollment to broaden the lens in order to drive equity to be at the forefront of our study (Walls, 2016).

Broadening the Lens

Focusing our efforts on the representation of the groups chosen for this study, we examined the proportion of all groups of students at the beginning and at the end of the semester both at the classroom and university levels. This investigation allowed us to better understand the make-up of the classroom in relation to the undergraduate students enrolled at the university. Figure 5.3 shows the comparison of the proportion of White and Hispanic female students across the semester. In fall 2018 there were 31,217 undergraduate students at the university, and there were 650 students enrolled in three sections of OCI. From the figure we can see that 25.5% of the total undergraduate enrollment is White female students and 12.4% is Hispanic female. The proportional enrollment in the OCI course is slightly higher for these two groups, with 26% and 12.8% White and Hispanic female, respectively. In addition, it is common to see a portion of students drop a course such as organic chemistry (Zoller, 1990; Seymour and Hewitt, 1997; Grove and Bretz, 2010), and accordingly we see that the number of students in the course drops to 541 total enrolled students at the end of the semester with a rate of 19.0% for Hispanic female and 19.8% for White female students. In Figure 5.3, we observe that the proportion of White and Hispanic female students remains similar at the beginning and end of the semester, suggesting that these two groups of students drop-trends are in line with the overall drop-tends at the university and course levels. This

is an encouraging result given the literature that suggests Women of Color, such as Hispanic women, tend to leave STEM courses at greater rates than their peers (Lubinski and Benbow, 2006; Carter-Sowell and Zimmerman, 2015).

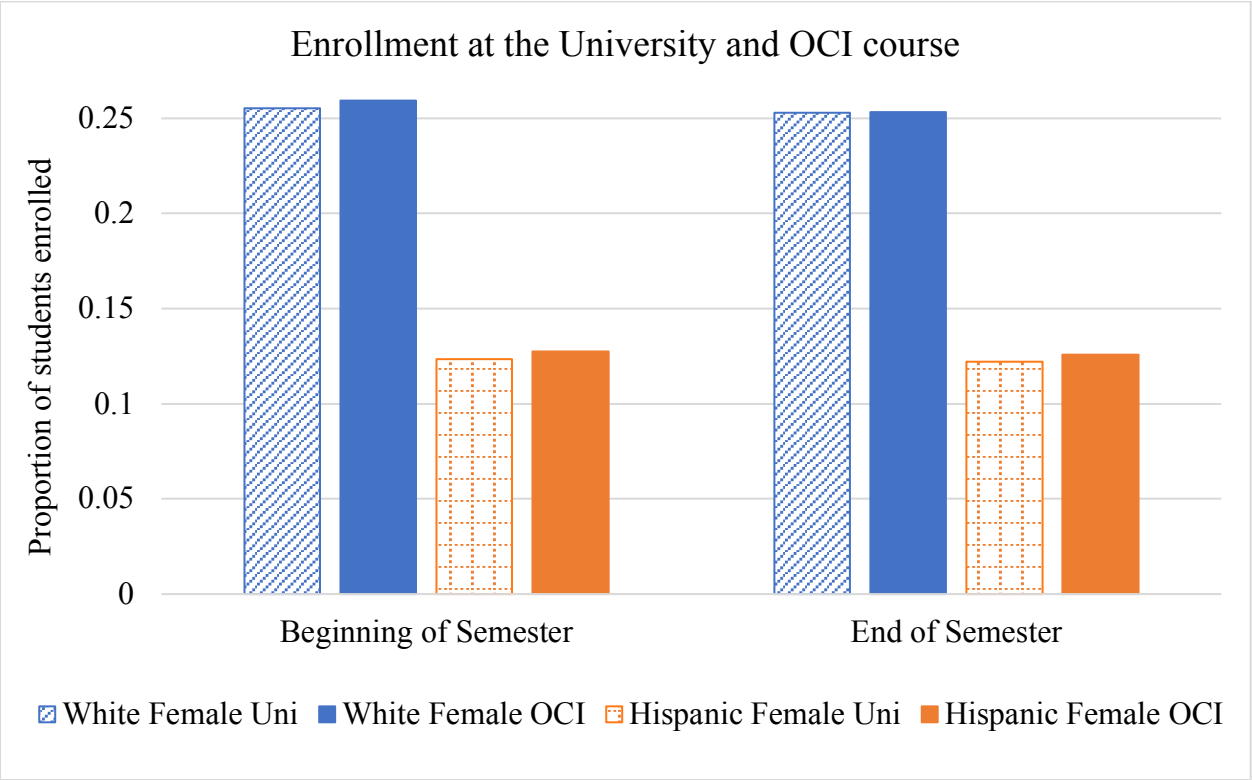


Figure 5.3. Comparison of enrolled students at the university and the OCI classroom at the beginning and end of semester.

Persistence to the Next Course in Chemistry Pathway

There are various reasons why students might choose to remain enrolled in a course through the semester; therefore, we followed our study with analyses on whether students persist to the next course in the sequence. First, we gathered information on students' major the term they

took the course and investigated whether their major required the students to take the next course in the sequence (Organic Chemistry II or Biochemistry). Table 5.6 shows the breakdown of the majors and whether students in those majors passed or failed the course, and Table 5.7 shows whether the students who passed enrolled in the next chemistry course the following term. It is important to note that the STEM and health majors all required the year-long sequence of organic chemistry and many required biochemistry, but none of the non-STEM majors required any chemistry course aside from a science elective requirement. We see that 22.6% of Hispanic female students failed the course (obtained a grade of C- or less) and 19.0% dropped the course (obtained a W) compared to 14.6% of White female students who failed, and 19.8% who dropped. Although the drop percentages were similar between the two groups, the fail rates of the Hispanic female students appear to be higher than the White female students, yet this difference is only due to seven students. However, due to the small sample size, particularly of the Hispanic female students, this small difference is amplified in percentage units.

Table 5.6. Drop, Pass, and Fail Rates for Hispanic and White Female Students

	Hispanic Female			White Female		
Majors	**Drop[a]**	**Fail[b]**	**Pass[c]**	**Drop[a]**	**Fail[b]**	**Pass[c]**
STEM and Health	15	18	48	33	25	105
Non-STEM	1	1	1	1	0	7
TOTAL	**16**	**19**	**49**	**34**	**25**	**112**
%	**19.0%**	**22.6%**	**58.3%**	**19.8%**	**14.6%**	**65.5%**

[a]Drop is designated for student who withdrew the course and earned a W. [b]Fail constitutes students who earned a grade of C- or less in the course, which does not allow them to enroll in the next course of the chemistry sequence. Not included in this group are students who withdrew. [c]Pass constitutes students who earned a passing grade of C or better

Table 5.7. Enrollment Rates to Next Course in the Sequence (Organic Chem II or Biochemistry)

	Hispanic Female	**White Female**
STEM and Health	37	78
Non-STEM	0	4
TOTAL	**37**	**82**
%	**75.5%**	**73.2%**

Furthermore, Table 5.6 provides the percentage of students in each group that passed the course and were eligible to take Organic Chemistry II or Biochemistry. While none of the non-STEM majors required any chemistry courses, it might be reasonable to think that the non-STEM students were enrolled in chemistry courses to fulfill other requirements such as entrance to medical school or the pursuit of a chemistry minor. Table 5.7 contains the number of students who passed the course and also enrolled in the next chemistry course the following semester.

Similarly, Figures 5.4a and b display the number of students in each subgroup that began the OCI course and their pathway. Students who passed OCI could either enroll immediately in Organic Chemistry II, which many students did, or they could enroll directly in Biochemistry. A few students in each subgroup chose to enroll in Biochemistry after OCI. There were also students who passed OCI and chose not to enroll in the next course in the sequence the next semester.

Previously we showed that the fail rates (dark orange bars in Figure 5.4a and b) appear to be higher for Hispanic female students, although we mentioned that the sample size is small, thus the percentage difference is magnified. We also saw that both of these groups of students enrolled in the next courses in the sequence (dark green and light orange bars in Figure 5.4a and b) in similar

percentages. This result indicates that given success in OCI, Hispanic female students move to the next course in the sequence at similar levels to their White female peers; a result which is encouraging given the historical notion that Women of Color continue their STEM tracks at lower rates than their White peers (Smyth and McArdle, 2004; Johnson, 2011).

Figure 5.4. Sankey plots of student retention in chemistry pathway. a) Plot of White female students in OCI. b) Plot of Hispanic female student in OCI.

Discussion

In this study we have strived to inspect our analyses looking for ways to foster practices that support equity. Although this study takes place in the United States, researchers in countries around the world have also used data to examine inequities, particularly gendered differences in science (*e.g.*, Olasehinde and Olatoye, 2014; Salta and Koulougliotis, 2015). Furthermore, as the field of CER moves toward greater diversity, inclusion, and equity, we hope researchers consider implementing similar approaches to their studies as we have demonstrated here. To this end, we utilized data collected with the ASCIv2 which theoretically and empirically measured two latent constructs labeled IA and ES. Following the previous literature that confirmed the 2-factor model (*e.g.*, Xu and Lewis, 2011; Brandriet, Ward, and Bretz, 2013, Xu, Villafañe, and Lewis, 2013), we investigated whether this model functioned for the two subgroups of students (Hispanic and White female) by employing confirmatory factor analysis techniques. Furthermore, we utilized measurement invariance testing as outlined in Rocabado *et al.* (2020) to provide evidence and support for group comparisons as well as for longitudinal comparisons. These steps to collect evidence for appropriate comparisons between groups are closely aligned with our commitment for social justice and the careful scrutiny of quantitative analyses used to inform about AHN students that are found in the tenets of QuantCrit (Solórzano, 1997; Solórzano and Delgado Bernal, 2001; Yosso, 2005; Walls, 2016; García, López, and Vélez, 2018; Gillborn, Warmington, and Demack, 2018).

In this study we have made efforts to contribute to the literature base in several ways. First, we have searched the literature for articles that have utilized the ASCIv2 in chemistry classrooms

longitudinally. This effort culminated in four meta-analyses of six articles plus this present study that informed the readers of the average effect size attitude change observed throughout a semester for two instructional styles, one with an intervention (*i.e.,* implementation of active learning pedagogy in the classroom) and one without an intervention. For the intervention studies we observed an average random weighted effect size of 0.16 for IA and 0.04 for ES. These results suggest that overall the chemistry classrooms that underwent instructional interventions experienced a small but positive gain in IA, and a negligible change in ES. On the other hand, the no treatment studies, including the present study, demonstrated an average random weighted effect size of -0.01 for IA and -0.10 for ES. Although these changes are negative, they are also considered negligible (Cohen, 1988). With these findings, we can conclude that these two subconstructs of attitude toward chemistry appear to be malleable and can be positively affected by certain pedagogical interventions, of which some might be more effective than others. For instance, Mooring *et al*., (2016) reported a positive gain in attitude with an effect size of 0.53 for IA and 0.32 for ES in the flipped classroom condition, while no change was observed in the traditional classroom. Similarly, Vishnumolakala and colleagues (2017) reported small to medium gains in POGIL classrooms in two different semesters in IA and ES. Additionally, it appears that these interventions during a semester make a greater positive impact on IA, while the impact on ES appears to be muffled by the general negative trend during the semester.

Second, we have likewise calculated the effect size for two subgroups in the OCI course, namely Hispanic and White female students. Our study reported effect sizes of -0.30 and -0.18 for Hispanic and White female IA, respectively. Although these effect sizes are small and negative, they are much larger in magnitude than the average effect size reported in the meta-analysis for no

intervention IA. Similarly, our study reported effect sizes of -0.46 and -0.33 for Hispanic and White female ES, respectively. These are small to medium effect sizes that are also much larger in magnitude than the average effect size reported in the meta-analysis for ES. This finding suggests that while it is important to examine attitude for the classroom overall, the trends of diverse groups within that classroom can be different, therefore subgroup investigations are important to learn about the students in our classrooms. Additionally, we compared IA and ES between these two groups at each time point. Both IA and ES for the Hispanic female students was lower than for the White female students at each time point. Although no evidence of significant difference was found between the two subgroups, the gap widened toward the end of the semester with a difference that represented a small effect size of 0.22 and 0.23 for IA and ES, respectively. This discouraging result, together with the observation that IA and ES drop more dramatically for the Hispanic female students in the course elicited a concern for this group of students. However, we realized that this result alone was short-sighted and did not adhere to the tenets of QuantCrit. We experienced a realization that our analysis approach and subsequent conclusions at this juncture propagated a deficit mindset, which then led us to search for ways in which we could continue our analyses and structure our study adhering more closely to our theoretical framework. Thus, recognizing that 'numbers are not neutral' and replacing deficit approaches for counterstorytelling, we broadened our analytical lens in search for evidence of counterstories particularly pertaining to the Hispanic female students in this study.

By broadening the lens of our analyses we utilized tenets of QuantCrit, namely that 'numbers are not neutral' and 'data does not speak for itself' (García, López, and Vélez, 2018; Gillborn, Warmington, and Demack, 2018). First, we investigated student representation in the

classroom based on the overall enrollment at the university. We saw that both White and Hispanic female students representation in the classroom matched that of the university. The drop rate of was similar between the two groups, namely 19.0% for Hispanic female and 19.8% for White female students. We also saw that Hispanic female students appear to fail the course at greater rates to their White female peers (22.6% and 14.8%, respectively). However, as we discussed, this difference of seven students is magnified due to small sample size; nonetheless we have concern for these students. Second, we observed that Hispanic female students who passed the course persisted into the next chemistry course in the sequence (OCII or Biochemistry) at similar rates to their White female peers. From this investigation we concluded that although we were about to succumb to a deficit ideology to search for ways in which Hispanic female students required "fixing," through a QuantCrit lens, we found instead similar retention to the next course for students who passed despite a declining attitude toward chemistry. These results suggest evidence of asset use, particularly for Hispanic female students who represent the "double bind" described by Ong and colleagues (2011) due to their minoritized intersectional identities. Persistence even in the face of opposition is an asset that has been investigated particularly for Hispanic women, who have demonstrated the use aspirational, familial, and linguistic assets (Yosso, 2005; Peralta, Caspary, and Boothe, 2013). Studies have revealed that one way in which Latina students navigate marginalization is to work to "prove others wrong" and persist and succeed in their aspirational goals (Rodriguez, Cunningham, and Jordan, 2019, pp. 268). In a collection of Chicana autobiographies, Cantú (2012) highlighted the importance of the roles of parents, family, and community in their success stories, which could be a point of further studies. Therefore, the findings of this study should be used to plan and design future research studies centered on Hispanic female student's asset use in organic chemistry classrooms.

Challenging our own deficit mindset through the use of QuantCrit, we were able to observe evidence of Hispanic female student persistence that we would otherwise not have inquired about and therefore not uncovered. This discovery was a direct result of our mindset shift we experienced. The constant inspection of our adherence to our framework allowed us to broaden our analytical lens and uncover initial evidence of counterstories, even in a quantitative study, and a spark to further investigations. Although the stories were not collected by individual narratives, we were able to observe the asset of persistence for the Hispanic female student group that corroborates narrative evidence from the literature (Cantú, 2012; Peralta, Caspary, and Boothe, 2013; Rodriguez, Cunningham, and Jordan, 2019). By centering our investigations on women, particularly Hispanic women, we affirm a commitment to social justice and herein demonstrated our efforts to utilize the QuantCrit framework (Solórzano, 1997; Solórzano and Delgado Bernal, 2001; Yosso, 2005; García, López, and Vélez, 2018; Gillborn, Warmington, and Demack, 2018).

Implications

Organic chemistry is a difficult course (Rowe, 1983; Barr *et al.*, 2010; Horowitz, Rabin, and Brodale, 2013), and it holds the status of one of the most feared and failed courses in the undergraduate curriculum (Grove, Hershberger, and Bretz, 2008; Flynn, 2015). It is no wonder that student attitudes decline over the course of the semester if there's no attempt to intervene. Declining attitudes are a contributing factor to the issue of underrepresentation of women, particularly women of color, despite the research that indicates that women and men can and do perform similarly in the sciences (Else-Quest, Mineo, and Higgins, 2013). How we investigate and

tackle this issue is the question that perhaps is the most important. It is common to make comparisons between groups on performance or affect measures, yet frequently we confuse differences for deficit (Gorski, 2011). When these mistakes occur, we fall into the trap of intending to 'fix' the students, often of underrepresented backgrounds, who are deficient in performance or affect, instead of focusing on their abilities to navigate marginalized spaces and their success in doing so. The idea and efforts to 'close the gaps,' which permeate the educational system is, by definition, a symptom of deficit ideology if we are not careful with the way in which we utilize the information gained by group comparisons (Gorski, 2011). In CER investigating performance or affect for People of Color sometimes takes the form of group comparisons, such as in this study we compared Hispanic and White female student attitude and retention. The issue is not in the comparison itself, rather it is in how we interpret and then proceed with this information as well as what we assume based on the results obtained.

For Researchers

In this study we have conducted a meta-analysis of studies that utilized the ASCIv2 longitudinally. Additionally, we compared the longitudinal effect size of the two subgroups in this study to the results of the meta-analysis. First, we observed that the effect sizes of IA and ES in a no intervention classroom were negligible, yet in an intervention classroom the IA effect size was small and the effect size for ES remained negligible. These results indicate that IA can be malleable, and certain pedagogical interventions may be beneficial for students, not only as they perceive the *intellectual accessibility* of field of chemistry, but also for the impact attitude can

have on performance and retention. On the other hand, although the overall ES effect size for intervention is more positive than for the no interventions group, it is still a negligible change. This result suggests that perhaps the magnitude of the change is not sufficient in one semester of intervention to overcome the typical negative trend in ES. Perhaps researchers could consider longer studies that may elucidate a greater positive impact of a series of interventions on students' attitude and in turn on performance and retention. Second, we observed that the subgroup effect size was negative and larger in magnitude than the effect size for the entire class and the overall effect size from all of the studies. These results again call for more longitudinal studies of the impact of interventions on students' attitude particularly for subgroups within a classroom. As the field of CER moves toward implementing greater diversity and inclusion initiatives, subgroup examination and comparisons will be necessary. Researchers should make efforts to check for differential experiences of student subgroups; however, these investigations must be performed with careful scrutiny to prevent propagation of social injustice (García *et al*., 2018; Gillborn *et al*., 2018). We call for adherence to frameworks such as CRT and QuantCrit to guide these research efforts, particularly when investigating AHN and other marginalized groups.

By utilizing QuantCrit as a framework we chose to center Hispanic female students and focus our efforts on driving equity to the forefront of our study (Walls, 2016). When we recognized that our study had fallen to the snare of deficit ideology, we looked for ways in which we could re-focus our analyses to more closely align with our framework and with the literature that exemplifies this shift. A particularly important article that helped our shift was Yosso's (2005) critique on Bourdieu's cultural capital framework, which essentially describes that certain cultures (*i.e*., People of Color) approach the classroom with deficiencies and lack the cultural capital for

social mobility that could aid in the rectification of these deficiencies. According to Bourdieu and Passeron (1977) framing the efforts to address social inequalities from the point of view of cultural capital would help People of Color achieve the desired outcomes, which in turn reifies deficit ideology. Although much of our educational system operates under the assumption that students come endowed with cultural capital from their various backgrounds, little has been accomplished to address the systemic issues that impede the advancement of diversity, inclusion, and equity effectively. Our research and practice is saturated with deficit ideology from almost every angle, thus actively challenging this systemic issue is an iterative process, as demonstrated in this study by the shift in mindset and closer adherence to our framework. Yosso (2005) challenges the cultural capital framework and presents the alternative concept of community cultural wealth, which focuses instead on assets that AHN students utilize in counterspaces, rather than assume these students lack cultural capital and are somehow doomed to remain in their deficient state. Ong, Smith, and Ko (2018) have investigated some counterspaces that are key for Hispanic women's persistence in STEM, such as national STEM diversity conferences, campus students clubs and organizations, STEM departments, and peer-to-peer relationships. By shifting the focus to assets instead of deficiencies, these and many other researchers have been able to display numerous ways in which People of Color, particularly Hispanic women, have shown resilience and persisted in STEM spaces based on their individual and collective experiences (Gallard-Martinez *et al.,* 2019). In this study, Hispanic female students display evidence of persistence as an asset to continue to the next course despite of less positive attitude than their White female peers. The use of the persistence asset, perhaps 'to prove others wrong' (Rodriguez, Cunningham, and Jordan, 2019, pp. 268), could be a contributing factor for the ultimate success of many Hispanic female students that navigated marginalized spaces in this OCI course, and it may be a

worthwhile focus of investigation in future studies. Investigating this interesting result further with complimentary data sources both quantitative and qualitative could yield fruitful evidence of asset use as well as a deeper understanding of the experiences of these students in this course.

For Practitioners

An important initial step for practitioners in challenging deficit mindset is to recognize the various ways in which this ideology permeates our own set of teaching and learning values and work diligently to find ways in which to replace this ideology with more useful practices that further social justice in more effective ways. One way to challenge deficit ideology is to look for evidence of student asset use. It is important to note that not everyone in the classroom has similar experiences as demonstrated with the subgroup investigation of attitude change over the course of the semester in this study. Through the meta-analysis of the IA and ES constructs, we demonstrated how the two subgroups of students' attitude toward chemistry change differed from the class and from other examples in the literature. While both groups' IA and ES effect size of the change throughout the semester were not only negative but larger in magnitude than the classroom overall, we observed that Hispanic female students' attitude had a more dramatic drop. Knowing that students can experience our classrooms differently, the pedagogies that we choose to enact should be carefully selected to positively impact the array of diverse backgrounds that exist within our classrooms. For instance, Mooring *et al.* (2016) observed small to medium gains in a flipped classroom for ES and IA, respectively. Later, Rocabado et al. (2019) investigated these same data to check whether the positive results extended to the Black female students in the course, and

found that while the attitude trend was positive for this subgroup, it began and ended lower than the rest of the students in the course. Therefore, while implementing flipped classroom pedagogies might hold promise, it is important that we check that the desired outcomes extend to groups of students of different intersectional backgrounds. Additionally, Seymour and Hunter (2019, pp. 246-254) note that all students appreciate teachers who show they care about their students, deliver engaging courses, and use humor efficiently, which in turn can have a positive impact on students' persistence. However, the most important practice for teachers is to challenge a deficit ideology in whichever method they choose to enact. For instance, Cohen *et al.,* (1999) described the implementation of cooperative learning to create equitable classrooms in which each student had a contribution to make. Other student-centered pedagogies, such as process-oriented guided-inquiry learning (Farrell, Moog and Spencer, 1999; Moog, 2014), may also prove effective when implemented away from a deficit mindset. To this end, instructors should closely investigate the most appropriate pedagogies that have been designed with diverse groups of students in mind and have been ratified with these groups while constantly check that these interventions are supporting these student populations. These efforts together with a conscious determination to challenge deficit ideology in support for asset use can help improve the experiences of the diverse groups of students in our classrooms resulting in further positive achievement and retention outcomes.

Decades of research have demonstrated that there is an association with race and ethnicity in the achievement and persistence pattern that disfavors Students of Color (*e.g.*, Seymour and Hunter, 2019). However, by utilizing CRT as our framework and challenging the deficit mindset in this study we have demonstrated that Hispanic female students' persistence into the next course is similar to their White female peers despite less positive attitude. This result indicates that we

must not only examine performance or affective metrics for this group of students; we can also broaden the lens and search for the display of assets. Quantitative as well as qualitative data can reveal the utilization of assets in the classroom (Peralta, Caspary and Boothe, 2013). As practitioners, we must stay attuned to our students' ways of navigating our chemistry classrooms and support the use of assets toward greater student success. We encourage the field of CER to adhere to frameworks such as CRT and to oppose deficit ideology by promoting counterstories in which marginalized groups use individual and/or community assets to combat their challenges (Kretzman and McKnight, 1993; Donaldson and Daugherty, 2011; Cantú, 2012; Peralta, Caspary, and Boothe, 2013; Myende, 2015; Rodriguez, Cunningham, and Jordan, 2019). Only then can we acknowledge strengths and become able to observe how students are not only merely consumers but producers of solutions (Myende, 2015).

Limitations

This quantitative study had a set of limitations that should be acknowledged. First, we have focused our analyses on two subgroups of students in the OCI course. Although investigating many subgroups of students would have resulted in interesting and informative comparisons, we were limited by small sample sizes from other subgroups, such as Hispanic male students and others. Additionally, further investigation about the persistence of students into the next course in the sequence in other semesters, such as summer 2019 or fall 2019, was not conducted. Thus our results and inferences on persistence are limited to students who enrolled in the next course the next semester. This limitation might exclude some students who did persist, but did so in a different

semester. In addition, this investigation was conducted in one university during one semester. Further investigations like this one in other settings and throughout multiple semesters could be fruitful to infer greater generalizability of the results. Another limitation is the relatively few studies that were available at this time for the meta-analysis. It would be worth repeating this investigation in a few years when there might be more data to include. This method of investigation is important to determine how malleable attitude factors are and to resolve under which conditions we can see the largest effects. Finally, investigations of asset use could be explored by utilizing qualitative methodology, particularly at this time when, to our knowledge, no other study has studied asset use for Hispanic female students in organic chemistry classrooms. In this study we did not collect qualitative data; therefore, this result could not be fully investigated. But we suggest that the asset of persistence can be a starting point for future qualitative and quantitative studies that could follow patterns found in Peralta, Caspary and Boothe's (2013).

References

AERA, APA, and NCME., (2014), *Standards for Educational and Psychological Testing,* American Psychological Association, Washington, DC.

Anderson T. L. and Bodner G. M., (2008), What Can We Do About 'Parker'? A Case Study of a Good Student Who Didn't 'Get' Organic Chemistry, *Chem. Educ. Res. Pract.*, **9**, 93−101. DOI: 10.1039/B806223B

Anzovino M. E. and Bretz S. L., (2015), Organic Chemistry Students' Ideas About Nucleophiles and Electrophiles: The Role of Charges and Mechanisms, *Chem. Educ. Res. Pract*, **16**, 797−810. DOI: 10.1039/C5RP00113G

Arjoon J. A., Xu X., and Lewis J. E., (2013), Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence, *J. Chem. Educ.*, **90**, 536−545. DOI: 10.1021/ed3002013

Baker C. A., (2019), A QuantCrit approach: Using Critical Race Theory as a means to evaluate if rate my professor assessments are racially biased, *J. Underrepresented and Minority Progress,* **3**(1), 1-22. DOI: 10.32674/jump.v3i1.1012

Banerjee M., Schenke K., Lam A. and Eccles J. S., (2018), The Roles of Teachers, Classroom Experiences, and Finding Balance: A Qualitative Perspective on the Experiences and Expectations of Females Within STEM and Non-STEM Careers, *Int. J. Gender Sci. Tech.*, **10**(2), 287−307.

Barr D. A., Gonzalez M. E. and Wanat S. F., (2008), The leaky pipeline: Factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students, *Acad. Med.,* **83**(5), 503-511. DOI: 10.1097/ACM.0b013e31816bda16

Barr D., Matsui J., Wanat S. F. and Gonzalez, M., (2010), Chemistry Courses as the Turning Point for Premedical Students, *Adv. Health Sci. Educ*., **15**, 45−54. DOI: 10.1007/s10459-009-9165-3

Bauer C. F., (2008), Attitude Towards Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts, *J. Chem. Educ*., **85**, 1440−1445. DOI: 10.1021/ed085p1440

Baumgartner L. M. and Johnson-Bailey J., (2008), Fostering Awareness of Diversity and Multiculturalism in Adult and Higher Education, *New Dir. Adult Cont. Educ.*, **2008**(120), 45−53. DOI: 10.1002/ace.315

Black A. E. and Deci E. L., (2000), The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective, *Sci. Educ*., **84**, 740-756. DOI: 10.1002/1098-237X(200011)84:6<740::AID-SCE4>3.0.CO;2-3

Bourdieu P. and Passeron J., (1977), *Reproduction in education, society and culture,* SAGE, London.

Brandriet A. R., Ward R. M. and Bretz S. L., (2013), Modeling meaningful learning in chemistry using structural equation modeling, *Chem. Educ. Res. Pract.*, **14**, 421-430. DOI: 10.1039/C3RP00043E

Brandriet A. R., Xu X., Bretz S. L. and Lewis J. E., (2011), Diagnosing changes in attitude in first-year college chemistry students with a shortened version of Bauer's semantic differential, *Chem. Educ. Res. Pract*., **12**, 271-278. DOI: 10.1039/C1RP90032C

Brown T. A., (2006), *Confirmatory Factor Analysis for Applied Research*, New York, NY: The Guilford Press.

Bulmer M. G., (1979), *Principles of Statistics*. New York: Dover.

Campbell-Montalvo R. A., (2020), Being QuantCritical of U.S. K-12 demographic data: Using and reporting race/ethnicity in Florida Heartland schools, *Race Ethn. Educ*., **23**(2), 180-199. DOI: 10.1080/13613324.2019.1679748

Cannady, M. A., Greenwald, E. and Harris, K. N., (2014), Problematizing the STEM pipeline metaphor: Is the STEM pipeline metaphor serving our students and the STEM workforce?, *Sci. Educ.,* **98**(3), 443-460. DOI: 10.1002/sce.21108

Cantú N., (2012), Getting there *Cuando no hay camino* (when there is no path): Paths to discovery *Testimonios* by Chicanas in STEM, *Equity & Excellence Educ.,* **45**(3), 472-487. DOI:10.1080/10665684.2012.698936

Carales V. D. and López R. M., (2020), Challenging deficit views of Latinx students: A strength-based perspective, *New Dir. Commun. Colleges,* 190, 103-113. DOI: 10.1002/cc.20390

Carlone H. B. and Johnson A., (2007), Understanding the science experiences of successful Women of Color: Science identity as an analytic lens, *J. Res. Sci. Teach*., **44**(8), 1187−1218. DOI: 10.1002/tea.20237

Carter-Sowell, A. R. and Zimmerman, C. A., (2015), Hidden in plain sight: Locating, validating, and advocating the stigma experience of Women of Color, *Sex Roles,* **73,** 399-407. DOI:10.1007/s11199-015-0529-2

Catsambis S., (1995), Gender, Race, Ethnicity, and Science Education in the Middle Grades, *J. Res. Sci. Teach*., **32**(2), 243−257. DOI: 10.1002/tea.3660320305

Chan J. Y. K. and Bauer C. F., (2014), Identifying at-risk students in general chemistry via cluster analysis of affective characteristics, *J. Chem. Educ*., **91**, 1417-1425. DOI:10.1021/ed500170x

Chan J. K. and Bauer C. F., (2016), Learning and studying strategies used by general chemistry students with different affective characteristics, *Chem. Educ. Res. Pract*., **17**, 675-684. DOI: 10.1039/C5RP00205B

Chen F. F., (2007), Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance, *Struct. Equ. Modeling*, **14**(3), 464-504. DOI: 10.1080/10705510701301834

Chen X., (2014), *STEM Attrition: College students' paths into and out of STEM fields*. In Attrition in Science, Technology, Engineering, and Mathematics (STEM) Education : Data and Analysis. (Ed.) Valerio J. Nova Science Publishers, Inc. pp. 1-96.

Cheng-Hsien L., (2016), Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares, *Behav. Res.,* **48,** 936-949. DOI:10.3758/s13428-015-0619-7

Cohen E. G., Lotan R. A., Scarloss B. A. and Arellano A. R., (1999), Complex Instruction: Equity in cooperative learning classrooms, **38**(2), 80-86. DOI: 10.1080/00405849909543836

Cohen J., (1988), *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed., Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper M. M., Grove N., Underwood S. M. and Klymkowsky M. W., (2010), Lost in Lewis Structures: An Investigation of Student Difficulties in Developing Representational Competence, *J. Chem. Educ.*, **87**(8), 869−874. DOI: 10.1021/ed900004y

Cortina J. M., (1999), What is Coefficient Alpha? An Examination of Theory and Applications, *J. Appl. Psychol.*, **78**, 98-104. DOI: 10.1037/0021-9010.78.1.98

Cracolice M. S. and Busby B. D., (2015), Preparation for College General Chemistry: More than just a matter of content knowledge acquisition, *J. Chem. Educ.,* **92**, 1790-1797. DOI:10.1021/acs.jchemed.5b00146

Crandell O. M., Lockhart M. A. and Cooper M. M., (2020), Arrows on the page are not a good gauge: Evidence for the importance of causal mechanistic explanations about nucleophilic substitution in organic chemistry, *J. Chem. Educ.,* **97**, 313-327. DOI:10.1021/acs.jchemed.9b00815

Crenshaw K., (1989), Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, *University of Chicago Legal Forum*, 139−168.

Crenshaw K., (1995), *Critical Race Theory: The key writings that formed the movement*, New York City: State University of New York Press.

Cronbach L. J., (1951), Coefficient Alpha and the Internal Structure of Tests, *Psychometrika*, **16**, 297-334. DOI: 10.1007/BF02310555

Damo K. L. and Prudente M. S., (2019), Investigating students' attitude and achievement in organic chemistry using interactive application, *Assoc. Computing Machinery,* 36-41, Tokyo, Japan., DOI:10.1145/3306500.3306562

Delgado A. and Stefanic J., (2001), *Critical Race Theory: An Introduction*, New York City, NYU Press.

Desy E. A., Peterson S. A. and Brockman V., (2011), Gender differences in science-related attitudes and interests among middle school and high school students, *Sci. Educ.,* **20**(2), 23-30. ISSN-1094-3277

Dixson A. and Anderson C. R., (2018), Where are we? Critical Race Theory in education 20 years later, *Peabody J. Educ.,* **93**(1), 121-131. DOI: 10.1080/0161956X.2017.1403194

Donaldson L. P. and Daugherty L., (2011), Introducing asset-based models of social justice into service learning: A social work approach, *J. Community Pract.* **19**(1), 80-99. DOI: 10.1080/10705422.2011.550262

Dood A. J., Fields K. B., Cruz-Ramírez de Arellano D. and Raker J. R., (2019), Development and evaluation of a Lewis acid-base tutorial for use in post-secondary organic chemistry courses, *Can. J. Chem.,* **97***,* 711-721. DOI: 10.1139/cjc-2018-0479

Else-Quest N. M., Hyde J. S. and Linn, M. C., (2010), Cross-national patterns of gender differences in mathematics: A meta-analysis, *Psychol. Bull.,* **136**(1), 103-127. DOI:10.1037/a0018053

Else-Quest N. M., Mineo C. C. and Higgins A., (2013), Math and science attitudes and achievement at the intersection of gender and ethnicity, *Psychol. Women Quart.,* **37**(3), 293-309. DOI: 10.1177/0361684313480694

Farrell J. J., Moog R. S. and Spencer J. N., (1999), A guided inquiry general chemistry course, *J. Chem. Educ.,* **76**(4), 570-574. DOI: 10.1021/ed076p570

Fink A., Frey R. F. and Solomon E. D., (2020), Belonging in general chemistry predicts first-year undergraduates' performance and attrition, *Chem. Educ. Res. Pract.,* (Published). DOI: 10.1039/D0RP00053A

Flaherty A. A., (2020), A review of affective chemistry education research and its implication for future research, *Chem. Educ. Res. Pract.,* **21,** 698-713. DOI: 10.1039/C9RP00200F

Flynn A. B., (2015), Structure and Evaluation of Flipped Chemistry Courses: Organic & Spectroscopy, Large and Small, First to Third Years, English and French, *Chem. Educ. Res. Pract.*, **16**, 198−211. DOI: 10.1039/C4RP00224E

Fordham S. and Ogbu J. U., (1986), Black Students' School Success: Coping with the Burden of Acting White, *Urban Rev.*, **18**, 176−206.

Gallard Martinez A. J., Pitts W., Ramos de Robles S. L., Milton Brkich K. L., Flores Bustos B. and Claeys L., (2019), Discerning contextual complexities in STEM career pathways: insights from successful Latinas, *Cult. Stud. Sci. Educ.,* **14***,* 1079-1103. DOI: 10.1007/s11422-018-9900-2

García N. M., López N. and Vélez V. N., (2018), QuantCrit: Rectifying quantitative methods through critical race theory, *Race Ethn. Educ.,* **21**(2), 149-157. DOI:10.1080/13613324.2017.1377675

Gasiewski J. A., Eagan M. K., Garcia G. A., Hurtado S. and Chang, M. J., (2012), From Gatekeeping to Engagement: A Multicontextual, Mixed Method Study of Student Academic Engagement in Introductory STEM Courses, *Res. High. Educ.*, **53**, 229−261. DOI:10.1007/s11162-011-9247-y

Geisinger B. N. and Raman D. R., (2013), Why they leave: Understanding student attrition from engineering majors, *Int. J. Eng. Educ.*, **29**(4), 914-925.

Gillborn D., Warmington P. and Demack S., (2018), QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics, *Race Ethn. Educ.,* **21**(2), 158-179. DOI:10.1080/13613324.2017.1377417

Gorski P. C., (2011), Unlearning deficit ideology and the scornful gaze: Thoughts on authenticating the class discourse in education, *Counterpoints,* **402***,* 152-173. https://www.jstor.org/stable/42981081

Gregorich S. E., (2006), Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework, *Med Care,* **44** (11 Suppl 3), S78-S94. DOI: 10.1097/01.mlr.0000245454.12228.8f

Grove N. P. and Bretz S. L., (2010), Perry's Scheme of Intellectual and Epistemological Development as a Framework for Describing Student Difficulties in Learning Organic Chemistry, *Chem. Educ. Res. Pract*., **11**, 207−211. DOI: 10.1039/C005469K

Grove N. P., Hershberger J. W. and Bretz S. L., (2008), Impact of a Spiral Organic Curriculum on Student Attrition and Learning, *Chem. Educ. Res. Pract*., **9**, 157−162. DOI: 10.1039/B806232N

Halpern D. F., Benbow C. P., Geary D. C., Gur R., Hyde, J. S. and Gernsbacher M. A. (2007), The science of sex differences in science and mathematics, *Psychol. Sci. Pub. Int.,* **8**, 1–51. DOI:10.1111/j.1529-1006.2007.00032.x

Harlow, L. L., (2014), *The essence of multivariate thinking: Basic themes and methods,* (2nd Eds), New York, NY: Routledge.

Harper S. R., (2010), An anti-deficit achievement framework for research on students of color in STEM, *New Dir. Inst. Res.,* **148***,* 63-74. DOI: 10.1002/ir.362

Heck, R. H. and Thomas, S. L., (2015), *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus,* (3rd Eds), New York, NY: Routledge.

Hedges L. V. and Olkin I., (1985), *Statistical methods for meta-analysis,* New York: Academic Press.

Horowitz G., Rabin L. A. and Brodale D. L., (2013), Improving student performance in organic chemistry: Help seeking behaviors and prior chemistry aptitude, *J. Scholarship Teach. Learn.,* **13**(3), 120-133.

Hu L. T. and Bentler P. M., (1999), Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives, *Struct. Equ. Modeling,* **6**(1), 283-292. DOI: 10.1080/10705519909540118

Ireland D. T., Freeman K. E., Winston-Proctor C. E., DeLaine K. D., McDonald Lowe S. and Woodson K. M., (2018), (Un)hidden Figures: A Synthesis of Research Examining the Intersectional Experiences of Black Women and Girls in STEM, *Rev. Res. Educ*., **42**, 226−254. DOI: 10.3102/0091732X18759072

Johnson D., (2011), Women of color in science, technology, engineering, and mathematics (STEM), *New Dir. Inst. Res.,* **152**, 75-85. DOI: 10.1002/ir.410

Kahveci A., (2015), Assessing High School Students' Attitudes Toward Chemistry with a Shortened Semantic Differential, *Chem. Educ. Res. Pract*., **16**, 283−292. DOI: 10.1039/C4RP00186A

Kenny D. A., Kaniskan B. and McCoach D. B., (2015), The Performance of RMSEA in Models with Small Degrees of Freedom, *Sociol. Method. Res*., **44**(3), 486-507. DOI:10.1177/0049124114543236

Klein A. and Moosbrugger H., (2000), Maximum Likelihood Estimation of Latent Interaction Effects With the LMS Method, *Psychometrika*, **65**(4), 457−474.

Komperda R., Pentecost T. C. and Barbera J., (2018), Moving Beyond Alpha: A Primer on Alternative Sources of Single-Administrations Reliability Evidence for Quantitative Chemistry Education Research, *J. Chem. Educ.*, **95**, 1477-1491. DOI: 10.1021/acs.jchemed.8b00220

Kraft A., Strickland A. M. and Bhattacharyya G., (2010), Reasonable Reasoning: Multi-Variate Problem-Solving in Organic Chemistry, *Chem. Educ. Res. Pract.*, **11**, 281−292. DOI: 10.1039/C0RP90003F

Kretzman J. and McKnight J., (1993), *Building communities from the inside out: A path toward finding and mobilizing a community's assets.* ABCD Institute, Evanston, IL.

Leontyev A., Chase A., Pulos S. and Varma-Nelson P., (2017), *Assessment of the Effectiveness of Instructional Interventions Using a Comprehensive Meta-Analysis Package,* In Computer Aided Data Analysis in Chemical Education Research (CADACER): Advance and Avenues, ACS Symposium Series: American Chemical Society, Washington, DC.

Leslie S. J., Cimpian A., Meyer M. and Freeland E., (2015), Expectations of Brilliance Underlie Gender Distributions Across Academic Disciplines, *Science*, **347**(6219), 262−265. DOI:10.1126/science.1261375

Lipsey M. W., and Wilson D. B., (2001), *Practical Meta-Analysis,* Applied Social Research Method Series (Vol. 49), Thousand Oaks: SAGE Publications.

Litzler E., Samuelson C. C. and Lorah J. A., (2014), Breaking it Down: Engineering Student STEM Confidence at the Intersection of Race/ Ethnicity and Gender, *Res. High. Educ.*, **55**, 810−832. DOI: 10.1007/s11162-014-9333-z

Liu Y., Raker J. R. and 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, *Chem. Educ. Res. Pract.,* **19,** 251-264. DOI: 10.1039/C7RP00153C

López E. J., Shavelson R. J., Nandagopal K., Szu, E. and Penn J., (2014), Ethnically diverse students' knowledge structures in first-semester organic chemistry, *J. Res. Sci. Teach.,* **51**(6), 741-758. DOI: 10.1002/tea.21160

Lubinski, D. and Benbow, C. P., (2006), Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise, *Perspect. Psychol. Sci.,* **1**(4), 316-345. DOI: 10.1111/j.1745-6916.2006.00019.x

Mitchell Y. D., Ippolito J. and Lewis S. E., (2012), Evaluating Peer-Led Team Learning across the two semester General Chemistry sequence, *Chem. Educ. Res. Pract.,* **13**, 378-383. DOI: 10.1039/C2RP20028G

Montes L. H., Ferreira R. A. and Rodriguez C., (2018), Explaining Secondary School Students' Attitudes Towards Chemistry in Chile, *Chem. Educ. Res. Pract*., **19**(2), 533−542. DOI: 10.1039/C8RP00003D

Moog R., (2014), *Process oriented guided inquiry learning*. In M. McDaniel, R. Frey, S. Fitzpatrick, & H.L. Roediger (Eds), In Integrating cognitive science with innovative teaching in STEM disciplines [E-reader version] (pp. xxx-xxx). doi: http://dx.doi.org/10.7936/K7PN93H.

Mooring S. R., Mitchell C. E. and Burrows, N. L., (2016), Evaluation of a Flipped, Large Enrollment Organic Chemistry Course on Student Attitude and Achievement, *J. Chem. Educ.*, **93**, 1972−1883. DOI: 10.1021/acs.jchemed.6b00367

Muthén B. and Asparouhov T., (2003), Modeling Interactions Between Latent and Observed Continuous Variables using Maximum-Likelihood Estimation in Mplus, *Mplus Web Notes*, **6**(1), 1−9.

Muthén L. K. Muthén B. O., (2010), *Mplus User's Guide*, 6th edn, Los Angeles, CA: Muthén and Muthén.
Myende P. E., (2015), Tapping in the asset-based approach to improve academic performance in rural schools, *J. Hum. Ecol,* **50**(1), 31-42. DOI:10.1080/09709274.2015.11906857
Nenning H. T., Idarraga K. L., Salzer L. D., Blaske-Rechek A. and Theisen R. M., (2019), Comparison of student attitudes and performance in an online and face-to-face inorganic chemistry course, *Chem. Educ. Res. Pract*., **21,** 168-177. DOI: 10.1039/C9RP00112C
Olasehinde K. J. and Olatoye R A., (2014), Comparison of male and female senior secondary school students' learning outcomes in science in Katsinia State, Nigeria, *Mediterranean J. Soc. Sci.,* **5**(2), 517-523. DOI: 10.5901/mjss.2014.v5n2p517
Ong M., Smith J. M. and Ko L. T., (2018), Counterspaces for Women of Color in STEM higher education: Marginal and central spaces for persistence and success, *J. Res. Sci. Teach.,* **55**(2), 206-245. DOI: 10.1002/tea.21417
Ong M., Wright C., Espinosa L. L. and Orfield G., (2011), Inside the Double Bind: A Synthesis of Empirical Research on Undergraduate and Graduate Women of Color in Science, Technology, Engineering, and Mathematics, *Harvard Educ. Rev.*, **81**(2), 172− 208. DOI:10.17763/haer.81.2.t022245n7x4752v2
Peralta C., Caspary M. and Boothe D., (2013), Success factors impacting Latina/o persistence in higher education leading to STEM opportunities, *Cult. Stud. Sci. Educ.,* **8***,* 905-918. DOI:10.1007/s11422-013-9520-9
Pérez Huber L., Vélez V. N. and Solórzano D., (2018), More than 'papelitos:' A QuantCrit counterstory to critique Latina/o degree value and occupational prestige, *Race Ethn. Educ.,* **21**(1), 208-230. DOI: 10.1080/13613324.2017.1377416
Rahman T., Lewis S. E., (2019), Evaluating the evidence base for evidence-based instructional practices in chemistry through meta-analysis. *J. Res. Sci. Teach.*, 1– 29. DOI: 10.1002/tea.21610
Rask K., (2010), Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences, *Econ. Educ. Rev.,* **29***,* 892-900. DOI:10.1016/j.econedurev.2010.06.013
Rocabado G. A., Kilpatrick N. A., Mooring S. R., and Lewis J. E., (2019), Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course, *J. Chem. Educ.,* **96**(11), 2371-2382. DOI: 10.1021/acs.jchemed.9b00516
Rocabado G. A., Komperda R., Lewis J. E. and Barbera J., (2020), Addressing diversity and social inclusion through groups comparisons: A primer on measurement invariance testing, *Chem. Educ. Res. Pract.,* **21,** 969-988. DOI: 10.1039/D0RP00025F
Rodríguez S., Cunningham K. and Jordan A., (2019), STEM identity development for Latinas: The role of self- and outside recognition, *J. Hispan. High. Educ.,* **18**(3), 254-272. DOI:10.1177/1538192717739958
Rowe M. B., (1983), Getting Chemistry Off the Killer Course List, *J. Chem. Educ.*, **60**, 954−956. DOI: 10.1021/ed060p954
Salta K. and Koulougliotis D., (2015), Assessing motivation to learn chemistry: Adaptation and validation of science motivation questionnaire II with Greek secondary school students, *Chem. Educ. Res. Pract.,* **16**, 237-250. DOI: 10.1039/C4RP00196F

Sass D., (2011), Testing Measurement Invariance and Comparing Latent Factor Means Within a Confirmatory Factor Analysis Framework, *J. Psychoeduc. Assess.,* **29**(4), 347-363. DOI:10.1177/0734282911406661

Sen S., Yilmaz A. and Temel S., (2016), Adaptation of the Attitude toward the Subject of Chemistry Inventory (ASCI) into Turkish, *J. Educ. Training Stud,* **4**(8), 27-33. ISSN-2324-805X

Seymour E. and Hewitt N., (1997), *Talking About Leaving: Why Undergraduates Leave the Sciences*, Boulder, CO: Westview Press.

Seymour E. and Hunter A-B., (2019), *Talking About Leaving Revisited: Persistence, relocation, and loss in undergraduate STEM education*, Springer Nature Switzerland.

Simon R. A., Aulls M. W., Dedic H., Hubbard K. and Hall N. C., (2015), Exploring Student Persistence in STEM Programs: A Motivational Model, *Can. J. Educ.*, **38**(1), 1−27. https://www.jstor.org/stable/10.2307/canajeducrevucan.38.1.09

Sloane J. D., (2016), *The influence of peer-led team learning on underrepresented minority student achievement in introductory biology and recruitment and retention in science, technology, engineering, and mathematics majors*, Thesis, Syracuse University.

Smith M. L. and Glass G. V., (1977), Meta-analysis of psychotherapy outcome studies, *Am. Psychol.,* **32**, 752-760. DOI: 10.1037/0003-066X.32.9.752

Smyth F. L. and McArdle J. J., (2004), Ethnic and Gender Differences in Science Graduation at Selective Colleges with Implications for Admission Policy and College Choice, *Res. High. Educ*., **45**(4), 353–381. DOI: 10.1023/B:RIHE.0000027391.05986.79

Solórzano D. G., (1997), Images and words that wound: Critical Race Theory, racial stereotyping, and teacher education, *Teach. Educ. Quart.,* **24**(3), 5-19. https://www.jstor.org/stable/23478088

Solórzano D. G., (1998), Critical Race Theory , race and gender microaggressions, and the experiences of Chicana and Chicano scholars, *Int. J. Qual. Stud. Educ.,* **11**(1), 121-136. DOI:10.1080/095183998236926

Solórzano D. and Delgado Bernal D., (2001), Critical race theory, transformational resistance and social justice: Chicana and Chicano students in an urban context, *Urban Educ.*, **36**, 308–342. ISSN: 0042-0859

Solórzano D. G. and Ornelas A., (2004), A critical race analysis of Latina/o and African American advanced placement enrollment in public high schools, *High School J.,* **87**(3), 15-26. https://www.jstor.org/stable/40364293

Stanich C. A., Pelch M. A., Theobald E. J. and Freeman S., (2018), A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women, *Chem. Educ. Res. Pract*., **19**, 846-866. DOI: 10.1039/C8RP00044A

Sullivan A., (2001), Cultural capital and educational attainment, *Sociology,* **35**(4), 893-912. DOI: 10.1017/S0038038501008938

Taber K. S., (2015), *Meeting educational objective in the affective and cognitive domains: Personal and social constructivist perspectives on enjoyment, motivation and learning chemistry*. In Affective Dimensions in Chemistry Education,(eds) Kahveci M., Orgill M. pp. 3-27, Springer Heidelberg

Tsui L., (2007), Effective strategies to increase diversity in STEM fields: A review of the research literature, *J. Negro Educ.,* **76**(4), 555-581. https://www.jstor.org/stable/40037228

Villafañe S. M., Garcia C. A. and Lewis J. E., (2014), Exploring diverse students' trends in chemistry self-efficacy throughout a semester of college-level preparatory chemistry, *Chem. Educ. Res. Pract.,* **15***,* 144-127. DOI: 10.1039/C3RP00141E

Underwood S. M., Reyes-Gastelum D. and Cooper M. M., (2016), When do students recognize relationships between molecular structure and properties? A longitudinal comparison of the impact of traditional and transformed curricula, *Chem. Educ. Res. Pract.*, **17**, 365-380. DOI: 10.1039/C5RP00217F

Vishnumolakala V. R., Qureshi S. S., Treagust D. F., Mocerin, M., Southam D. S. and Ojeil J., (2018), Longitudinal impact of process-oriented guided inquiry learning on the attitudes, self-efficacy and experiences of pre-medical chemistry students, *QScience Connect,* **1**, 1-12. DOI:10.5339/connect.2018.1

Vishnumolakala V. R., Southam D. C., Treagust D. F., Mocerino M. and Qureshi S. (2017), Students' attitudes, self-efficacy and experiences in a modified process-oriented guided inquiry learning undergraduate chemistry classroom, *Chem. Educ. Res. Pract*., **18**, 340-352. DOI: 10.1039/C6RP00233A

Walls L., (2016), Awakening a dialogue: A critical race theory analysis of U.S. nature science research from 1967 to 2013, *J. Res. Sci. Teach.,* **53**(10), 1546-1570. DOI: 10.1002/tea.21266

Warfa A-R. M., (2016), Using cooperative learning to teach chemistry: A meta-analytic review, *J. Chem. Educ.,* **93***,* 248-255. DOI: 10.1021/acs.jchemed.5b00608

Wyer M., (2003), Intending to Stay: Images of Scientists, Attitudes Toward Women, *J. Women Minor. Sci.,* **9**(1), 1-16. DOI: 10.1615/JWomenMinorScienEng.v9.i1.10

Xu X. J., (2016), Attention to retention: Exploring and addressing the needs of college students in STEM majors, *J. Educ. Training Stud.,* **4**(2), 67-76. ISSN-2324-805X

Xu X., Alhoosani K., Southam D. and Lewis J. E., (2015), *Gathering Psychometric Evidence for ASCIv2 to Support Cross-Cultural Attitudinal Studies for College Chemistry Programs*. In Affective Dimensions in Chemistry, Springer-Verlag: Berlin, pp. 177−194.

Xu X. and Lewis J., (2011), Refinement of a Chemistry Attitude Measure for College Students, *J. Chem. Educ.,* **88**, 561-568. DOI: 10.1021/ed900071q

Xu X., Southam D. and Lewis J. E., (2012), Attitude Towards the Subject of Chemistry in Australia: An ALIUS and POGIL Collaboration to Promote Cross-National Comparisons, *Aus. J. Educ. Chem.*, **72**, 32−36. ISSN:14459698

Xu X., Villafañe S. M. and Lewis J. E., (2013), College students' attitudes toward chemistry conceptual knowledge and achievement: structural equation model analysis, *Chem. Educ. Res. Pract*., **14**, 188-200. DOI: 10.1039/C3RP20170H

Yep G. A., (2014), Talking back: Shifting the discourse of deficit to a pedagogy of cultural wealth of international instructors in US classrooms, *New Dir. Teach. Learn.,* **138***,* 83-91. DOI:10.1002/tl.20099

Yosso T., (2005), Whose culture has capital? A critical race theory discussion of community cultural wealth, *Race Ethn. Educ.,* **8**(1), 69-91. DOI: 10.1080/1361332052000341006

Zoller U., (1990), Students' Misunderstandings and Misconceptions in College Freshman Chemistry (General and Organic), *J. Res. Sci. Teach*., **27**(10), 1053−1065. DOI:10.1002/tea.3660271011

CHAPTER 6:

GATHERING VALIDITY EVIDENCE IN THE DEVELOPMENT OF A NEW VERSION OF THE ATTITUDE TOWARD THE SUBJECT OF CHEMISTRY INVENTORY (ASCI-UE)

Note to Reader

This chapter is a study done together with collaborators in Chile: Dr. Roberto Ferreira a professor in the department of education at Universidad Católica de la Santísima Concepción in Chile, his student Lilian Montes, and Dr. Cristina Rodríguez a professor in the department of psychology at the Universidad de la Laguna in Spain.

Introduction

In this chapter I will demonstrate the process of gathering validity evidence in the refining and development of a new version of the Attitude toward the Subject of Chemistry Inventory - Utility and Emotional Satisfaction (ASCI-UE) that surfaced following several rounds of data collection and analysis with the ASCIv2. Over time several idiosyncrasies in item behavior were observed with this instrument, particularly with certain items, such as Item 6 (Challenging – Not Challenging; *i.e.*, Rocabado *et al.*, 2019) in the *Intellectual Accessibility* factor. Some of these

inconsistent behaviors were also observed in different countries and languages where the instrument was administered (Xu et al., 2015; Montes, Ferreira and Rodríguez, 2018). Therefore, I along with collaborators in Chile gathered data for a new version of the ASCI that reflects the perceptions of students toward the subject of chemistry featuring an affective (*Emotional Satisfaction*) factor and a different cognitive (*Utility*) factor than the original ASCIv2 to measure attitude toward chemistry.

In the beginning, the ASCIv2 (Xu and Lewis, 2011) was refined from the original ASCI created by Bauer in 2008 following the logic that attitude is composed of affective and cognitive domains (Krech, Crutchfield, and Ballanchey, 1962). Bauer's instrument had 20 items with 5 proposed factors with only three of those factors displaying item correlations and factor loadings appropriate for the theorized model (Bauer, 2008). One of the factors in the original ASCI was *Interest and Utility*, which gives empirical background to explore a similar factor in this study. I explored whether the items in the ASCIv2 followed the theoretical underpinnings of the cognitive and affective domains while gathering several sources of validity evidence (*i.e.*, response process validity, etc.) as described by *The Standards for Education and Psychological Testing* (AERA *et al.*, 2014). From these sources I refined the existing *Emotional Satisfaction* (ES) scale, the *Intellectual Accessibility* (IA) scale, and added a scale that was prevalent in all students' interviews, which is *Utility* (U).

The refinement and design of the items in these factors were done following theoretical ideas taken from prevalent attitude theories which stem from the notion that attitude is composed of cognitive and affective factors that lead to behavioral intentions (Ajzen and Fishbein, 2000). IA

was conceptualized based on the description by Rosenberg and Hovland (1960) that cognition is perceptive responses and verbal statements that come from conceptual or semantic memory. In this sense, the conceptual memory is understood as organized knowledge and meaning of words or other verbal symbols and the relationship between them. Therefore, IA is the collection of student perceptions on how approachable the object (chemistry) is, derived from thoughts and organized knowledge in order to determine beliefs.

ES was conceptualized as the notion that all emotions are reactions to an external stimuli (Damasio, 1994). Damasio describes a division of emotions between primary and secondary emotions. Primary emotions are "complex, coordinated, and automatic" and secondary emotions are variations of the primary emotions that arise from "evaluative, voluntary and non-automatic mental processes" and come from experience (pp. 149-150). Therefore, ES, described as secondary emotions, is a collection of voluntary responses toward the object of attitude (chemistry), which constitute an evaluation that comes from affective mental processes produced from experience.

Utility value is defined by Wigfield and Eccles (2000) as the notion of how a task contributes to an individual's future. Pintrich and Schrauben (1992) describe how utility is an extrinsic belief about how one thinks a task (or object) will help to achieve future goals. In this case, I apply the lens of attitude to this definition and conceptualize U as the notion of how the attitude object of chemistry contributes to an individual's life in terms of its application toward long-term goals .

Although the original idea was to expand the ASCIv2 to contain three factors by adding the U factor to the instrument, in practice this idea was not successful. Instead, I provide the readers with an alternative instrument to the ASCIv2 measuring ES and U that can help researchers and practitioners answer different research questions and interests in their chemistry classrooms. Additionally, my ongoing interest in the attitude-achievement and attitude-retention relationships were the motivation for collecting longitudinal data. Also, keeping this same focus, investigating attitude for students who successfully passed the course (high-achievement group) compared to students who did not pass the course (low-achievement group) was a worthwhile emphasis of this study. The process of refinement and development of this new instrument (ASCI-UE) as well as its application throughout an OCII course will be detailed in the *Methods* and *Results* sections below.

Research Questions

This project was guided by five research questions, which led to the development of a new instrument (ASCI-UE) in English and Spanish and was administered in the U.S. and in Chile. Additionally, I explored several aspects of validity evidence, such as relation to other variables, guided by *The Standards* (AERA *et al.,* 2014). The research questions were as follows:

1. How were students' evaluations of chemistry congruent with the theoretical underpinnings of the ASCIv2?

2. What perceptions of chemistry emerged out of the student interviews and expert panel review that were not captured by the ASCIv2?
3. To what extent does the internal structure of the new instrument hold in two languages (English and Spanish) and in two countries (Chile and U.S.)?
4. To what extent does attitude measured by ASCI-UE relate to other variables such as *Perceived Competence* and achievement in an Organic Chemistry II (OCII) course?
5. How does attitude measured by ASCI-UE change from the beginning to the end of a semester in an organic chemistry course, for both low- and high-achieving students?

Methods

This work was guided by *The Standards* (AERA *et al*., 2014). Gathering validity and reliability evidence to develop an instrument is required to determine the benefit of using the instrument to make inferences about students in chemistry classrooms. In this chapter, I provide evidence of several aspects of validity and reliability for the ASCI-UE. Figure 1 describes the process of the development of the ASCI-UE culminating in its use in an Organic Chemistry II (OCII) course in the U.S. The data analysis and use of this instrument in Chile will be further discussed in a later publication.

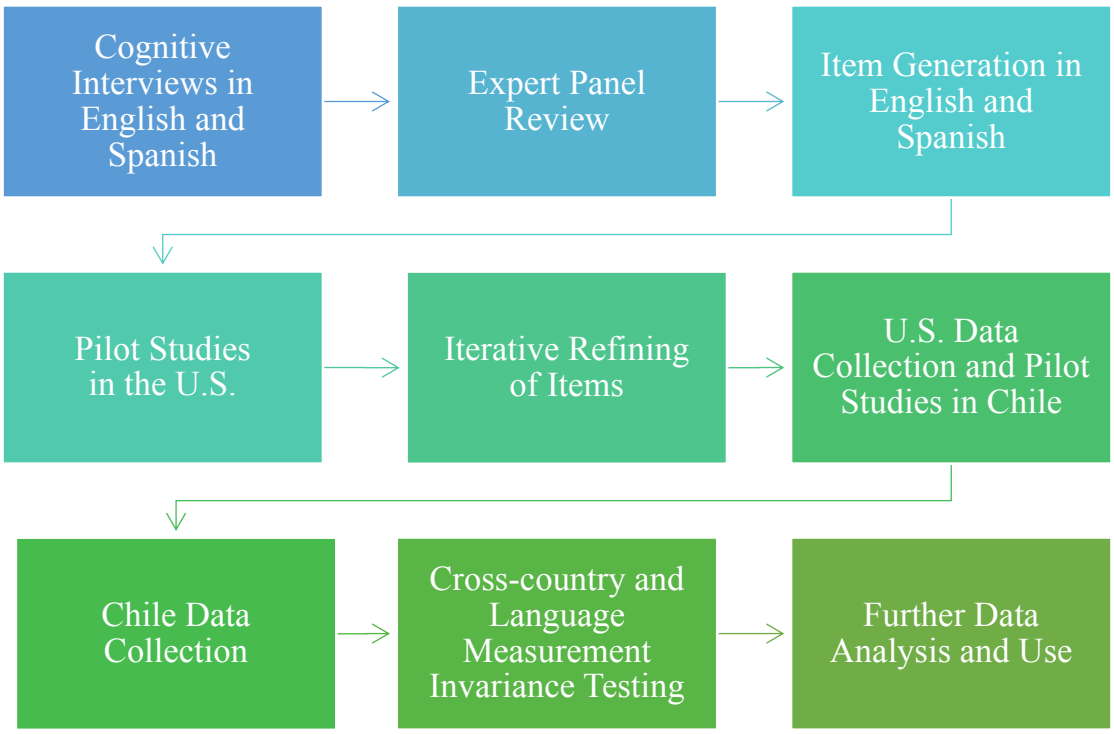


Figure 6.1. Chronology and process of development of ASCI-UE in English and Spanish, in the U.S. and in Chile.

Response Process Validity – Cognitive interviews

Cognitive interviews (Willis, 1999) were conducted following a semi-structured approach (Guba and Lincoln, 1983; Wilkinson, Joffe and Yardley, 2004; Curtis and Curtis, 2017) with students in General Chemistry II and Organic Chemistry I and II following an approved IRB protocol (see Appendix D) at a research intensive institution in the southeastern United States. Eleven students volunteered to be interviewed and were compensated with a $25.00 Amazon gift card upon the completion of the interview. Each student signed a consent form and was told that they would be given a pseudonym and their identity would not be shared outside of the immediate

research team. This team did not include my collaborators in Chile; therefore, only deidentified quotes, such as the ones shared in this document were shared with the team in Chile.

The topic of the interviews was the interpretation of the items on the ASCIv2. The interviews began with a few questions about the students' major and general interest in science to establish rapport. Following these questions, the students were asked to read the items on the ASCIv2 one at a time and describe how they interpreted each item. Finally, the students were asked to provide additional adjectives that could describe their thoughts and feelings toward chemistry and to provide an explanation of each of the items. The interviews were recorded, transcribed, and coded for analysis and interpretation.

During these interviews students indicated opportunities to refine some of the existing items in the ASCIv2 and also provided evidence of another factor of prominence to evaluate the discipline of chemistry, which was *Utility* (U). Adjectives for the factor of *Utility* were tested in some of the interviews that were conducted. It is important to note that all students interviewed both in the U.S. and in Chile mentioned aspects of utility in their responses when asked about their thoughts and feelings toward chemistry.

Content Validity – Expert Panel Review

After conducting, transcribing, coding, and analyzing the student interviews, I consulted with a panel of well-established experts in the fields of chemistry, attitude, chemistry education

research, and psychometrics. The panel was consulted in numerous occasions throughout the development and refinement of the instrument. These experts helped with item generation for the *Utility* factor, refinement of items belonging to the original ASCIv2 factors, as well as the conceptualization of the factors.

Initially, the idea was to add the factor of *Utility* to the ASCIv2 and to refine the original two factors of Intellectual Accessibility and Emotional Satisfaction, making the new instrument a three-factor attitude instrument. Because the Utility factor was new, ten items related to Utility and a second affective factor were added to the original ASCIv2 instrument, plus one possible new item for each of the original factors, with the idea that factor analysis would help to determine the best set of items for each factor. This lengthier version of the instrument was administered to 2000+ students and the data were randomly split in half for exploratory and confirmatory factor analyses. Details of this work are presented in Appendix C. Ultimately, the most tenable instrument was a two-factor instrument comprising a revised *Emotional Satisfaction* factor and the new *Utility* factor. Factor analysis conducted on pilot data collected with a three-factor instrument resulted in lack of convergence of the model. Based on conversations with the expert panel, both *Utility* and *Intellectual Accessibility* are both factors that describe cognitive mental processes, so the decision to go forward with a two-factor instrument measuring *Utility* and *Emotional Satisfaction* maintains the theoretical notion that attitudes have cognitive and affective domains (Krech, Crutchfield, and Ballanchey, 1962).

Descriptive Statistics

In the summer of 2019, the two-factor instrument was administered to a General Chemistry II and an Organic Chemistry II course. Throughout the summer semester, I was able to continue refining the instrument in accordance with the results of the administration and the comments and suggestions of the expert panel. Details regarding the refinement process can be found in Appendix C. Finally, in the Fall of 2019 I administered the final version of the two-factor instrument in two sections of Organic Chemistry II course taught by the same instructor two days before each of the exams throughout the semester. In this work, I focused on the first and fourth (final) administrations of the instrument; however, descriptive statistics and further analyses are reported in Appendix C for the second and third instrument administrations. Students were incentivized to complete the survey for extra-credit points toward their exam score including the final exam (2% of the exam grade). Students who did not complete the survey had no penalty.

A total of 291 out of 304 students participated in the survey from the two sections of OCII at the beginning of the semester. Additionally, the groups we investigated in this study were students who displayed high- and low-achievement in the course. Students in the high-achievement group were those who earned a passing grade (C or better), and the students in the low-achievement group were those who earned a failing grade (C- or worse). There were also 29 students from the initial 291 who withdrew from the course who were counted in the low-achievement group at the beginning of the semester.

The ASCI-UE is a seven-point semantic differential scale containing 9 items; four items belonging to the *Utility* (U) factor and five items belonging to the *Emotional Satisfaction* (ES) factor. Mean, standard deviation, skewness, and kurtosis were computed for each item at each time point utilizing the SPSS v26 software. Additionally, the observed mean scores for U and ES were calculated by taking the average of the item scores which correspond to each factor. Items 1, 4, 6, 7, and 9 were reverse coded so that higher scores could be associated with the positive adjective or a positive attitude toward chemistry. A score of four indicates neutrality; meaning, for instance, that students found chemistry neither relevant nor irrelevant (Item 1). Descriptive statistics were reported for the two subgroups in this study, however, descriptive statistics and other analyses for the entire course were reported in Appendix C. Furthermore, I investigated the difference in scores between low- and high-achievers by conducting a MANOVA along with a measure of effect size with Cohen's *d* (Cohen, 1988) at the beginning and end of the semester.

In addition to the ASCI-UE, students were also asked to complete a four-item *Perceived Competence* scale (PC; Williams and Deci, 1996) in accordance to the constructs investigated in Self-Determination Theory (SDT; Ryan and Deci, 2017). Competence is a cognitive mental process defined as "*feeling effective in one's interactions with the social environment – that is, experiencing opportunities and supports for the exercise, expansion, and expression of one's capacities and talents*" (Ryan and Deci, 2017 pp. 86). *Perceived Competence* (PC) is a self-evaluative process that arises from theoretical currents like social learning theory (Bandura, 1977). This construct allows for a person's discrete evaluative judgement of their competence when engaging in a task (Harter, 1982). PC was chosen as proxy for the original *Intellectual Accessibility* (IA) factor of the ASCIv2 because of the similar conceptualization between these two factors as

cognitive evaluative judgments about a person's ability or accessibility of success in the course. However, the main difference between these factors is that IA evaluates the discipline, and PC is a personal self-evaluation. The PC scale utilizes a seven-point Likert-type scale with 1 indicating "not at all true," 4 indicating "somewhat true," and 7 indicating "very true." Mean, standard deviation, skewness, and kurtosis for each item at each time point were also computed in SPSS v26. The first and final administrations are reported within this document, and results of the second and third administrations are found in Appendix C. Additionally, the two instruments as administered to the course are found in Appendix C.

Internal Structure Validity – Confirmatory Factor Analysis, Reliability, and Measurement Invariance Testing

Each instrument was subject to confirmatory factor analysis (CFA) using Mplus v8.2 (Muthén and Muthén, 1998-2007) at each time point separately to verify the two-factor internal structure of the ASCI-UE and the one-factor internal structure of PC (see Tables 6.2 and 6.3 in this chapter and S6.4 in Appendix C). The seven-point scales were treated as continuous and a Maximum Likelihood Robust (MLR) estimator was used to handle non-normally distributed data (Cheng-Hsien, 2016), such as skewness and kurtosis outside of the +/- 1.00 range (Bulmer, 1979).

The models were identified by fixing the first item loading of each factor to 1.00 and allowing all other parameters to be freely estimated. Additionally, model fit indices were used to determine appropriate data-model fit. To assess model fit I first examined the chi-square (χ^2)

statistic. The χ^2 is influenced by large sample sizes; therefore, it was important to inspect additional fit indices, such as the comparative fit index (CFI), the root mean square of approximation (RMSEA), and the standardized root-mean square residual (SRMR; Brown, 2006). The suggested cutoff criteria for these fit indices are as follows: for CFI > .90 is acceptable, but best if >0.95; for RMSEA <0.06; and for SRMR <0.08 (Hu and Bentler, 1999).

Reliability is a measure of precision of a measurement (Komperda, Pentecost and Barbera, 2018). Often, studies report Cronbach's alpha as a measure of reliability (Cronbach, 1951; Cortina, 1993); however, this coefficient works under the assumption of a *tau*-equivalent model where all factor loadings are constrained to take the same value (Komperda, Pentecost and Barbera, 2018). More often than not, our studies require models that are congeneric, meaning that the factor loadings in the model are freely estimated and are allowed to be different from each other. Thus, Cronbach's alpha was not an appropriate reliability coefficient for a congeneric model. Komperda and colleagues (2018) suggested alternative coefficients of reliability for congeneric models, of which the McDonald's Omega coefficient was the one used in this study (see Tables 6.2 and 6.3 in this chapter and S6.4 in Appendix C). This coefficient was directly calculated using the parameter estimates obtained from the output of the CFA and much like Cronbach's alpha, values closer to one indicated high reliability given a good data-model fit. Equation 6.1 showed how to calculate the Omega coefficient of reliability where lambda (λ) represents the standardized factor loadings and theta (θ) represents the error variances.

[Eq. 6.1]
$$\omega = \frac{(\Sigma\lambda)^2}{(\Sigma\lambda)^2 + \Sigma\theta}$$

In addition to establishing appropriate data-model fit as evidence of internal structure validity at the beginning and end of semester, I also wanted to establish evidence supporting longitudinal comparisons of these measures. I also intended to gather evidence of internal structure validity across two countries and two languages. Measurement invariance testing was an appropriate method to establish evidence of longitudinal comparisons, subgroup comparisons, and cross-country comparisons. I used the steps delineated by Rocabado and colleagues in 2020 (see Tables 6.4, 6.5, 6.6, and 6.8).

In measurement invariance testing the first step after having established a good data-model fit with CFA was to test the configural model where the model with all freely estimated parameters was tested at different time points (longitudinal), or for different groups (*i.e.*, Chile and U.S.) simultaneously. With appropriate fit for the configural model, then in the metric model the constraint of equal loadings was added across time or groups. Evaluation of model fit was again conducted as well as evaluation of the change in model fit. If metric invariance held, then scalar invariance was conducted by adding the constraint of equal intercepts across groups or time. Evaluation of fit and change in model fit was also completed for the scalar model. Finally, the constraint of equal error variances was added for strict invariance, evaluating this model in the same way that the metric and scalar models were scrutinized. Model fit was assessed using the guidelines described by Hu and Bentler (1999) that were used to assess CFA data-model fit. For the change in model fit, Chen (2007) suggested guidelines in addition to calculating the $\Delta\chi^2$. The change in model fit was established based on the following cutoffs: ΔCFI (<0.01), ΔSRMR (<0.03), and ΔRMSEA (<0.015) for metric invariance, and ΔCFI (<0.01), ΔSRMR(<0.01), and ΔRMSEA (<0.015) for scalar and strict invariance (Chen, 2007).

Relationship to Other Variables Validity – Correlation and Structural Equation Modeling (SEM)

Relationship to other variables can be established by investigating correlational interactions between factors. In this case I was interested in investigating the relationship between the factors of the ASCI-UE (U and ES) and PC. Investigating the pattern of these construct relationships can potentially elucidate future areas of investigation for students in OCII. The construct of PC was chosen due to its theoretical relationships with attitude and achievement (Ryan and Deci, 2017). Exploring variable relationships provided validity evidence for the inferences that can be drawn from the constructs in the ASCI-UE, therefore correlational analyses were conducted between U, ES, and PC as well as their respective correlations to achievement.

Furthermore, relationships to variables of achievement, such as exam scores can also be established by utilizing structural equation modeling (Kline, 2015). In this work, I tested a reciprocal causation model (Pekrun, 2006; Pekrun, Maier and Elliott, 2009; Villafañe, Xu and Raker, 2016; Gibbons and Raker, 2018; Gibbons *et al.,* 2018) describing the relationship between the ASCI-UE factors and achievement scores at the beginning and end of the semester. Several models were tested (A-E see Appendix C), with model A (Figure 6.2) displaying the best fit both theoretically and statistically.

Results

The results of this study came mostly from the examination of the data collected in the U.S.; however, one of the analyses was performed together with the data collected in Chile. The complete set of results from the analyses conducted with data collected in Chile will be reported elsewhere. The cognitive interviews and all of the quantitative data presented herein was collected following an approved protocol in the U.S. The data collected in Chile was also collected following an approved protocol. The data from Chile presented in this section was previously deidentified and was only used for the purpose of investigating whether the instrument functions similarly in two languages and in two countries. I did not report the results of the cognitive interviews conducted in Chile in this chapter; however, the entire collection of interviews was used to inform the development of the *Utility* factor as well as the refinement of the *Emotional Satisfaction* factor. The results for the interviews conducted in Chile will be shared elsewhere.

Furthermore, an additional focus of this study was to investigate the effects of attitude for high-achieving students (those who obtained a passing grade) and low-achieving students (those who obtained a failing grade) at the beginning and at the end of the semester. Thus, the results herein displayed comparisons between these two subgroups of students in OCII.

Cognitive Interviews and Expert Panel Review

Cognitive interviews were conducted with eleven volunteer students from General Chemistry and Organic Chemistry courses. The interviews lasted between 30 to 90 minutes with an average of 60 minutes. Each student provided insight into their interpretation of each of the 8 items in the ASCIv2 as well as additional adjectives they thought would be good candidates to add to the instrument. Based on the examination of each item discussed in the interviews I, along with my collaborators, created the new ASCI-UE presented in this chapter. The ASCI-UE is composed of nine items, four that belong to the *Utility* factor, and five that belong to the *Emotional Satisfaction* factor. Appendix C contains a large portion of the interview data that were used to make decisions about the instrument, namely, to refine the ES factor, and select adjectives to add to the U factor. I also presented data about the IA factor items. Throughout this section I presented relevant quotes from students about each of the items in the ASCI-UE as well as instances when the panel of experts helped with item refinement or other decisions throughout the process.

Relevant-Irrelevant (*Utility*)

The adjective "relevant" was prevalent among most students who were interviewed. They emphasized the need for the discipline to be relevant for their individual future goals. To a degree, students also talked about the utility of chemistry in a global and general sense; however, in this study we focus on the students' perceptions at the individual level. Relevant-Irrelevant is one of the items in the U factor because it portrays the idea that chemistry should be useful at some level. Students agreed that this adjective is a good way in which to evaluate the discipline of chemistry.

For instance, student 3 said, *"... but in the future, that might be something that would be relevant or like even converting measurements, that might be something that I'd have to do, I will have to do. Yeah.* ***So yes, I would say [chemistry]'s relevant****."*

Depressing-Exciting (*Emotional Satisfaction*)

This item was settled through iterative tests of different adjectives that elicited strong emotions that students were expressing in the interviews. Student 6, for example, indicated, *"[Comfortable] doesn't make sense within the context of chemistry. When I think of something that's comfortable, I think of me like being home petting my cat, wearing sweatpants and like watching Netflix. That's not chemistry.* ***Chemistry makes me wanna cry****."* When expanding upon this concept, this student as well as others expressed that the original item "Comfortable-Uncomfortable" were not adjectives that they could easily use for chemistry; however, they did portray strong emotions toward the discipline. Student 3 talked about the need to be excited about the discipline. *"Like it's something that you're like into, a field that you're interested in. ...[If] it doesn't interest you at all,* ***it's not exciting to you at all****."* With the help of the expert panel, the item "Depressing-Exciting" was generated to replace "comfortable-uncomfortable" to capture students' strong emotions toward chemistry depicted in the illustrative quotes.

Unnecessary-Essential (*Utility*)

Students voiced the sense that knowledge of chemistry was necessary to advance in the fields of science, medicine, and others. Some students talked about global warming and other

important topics in which an understanding of chemistry was essential to help them be informed citizens. This idea was captured by generating the "Unnecessary-Essential" for the U factor. Student 5 related their view of chemistry as essential knowledge in the field of medicine when they said, *"One is the active role chemistry plays in my life,* ***but also the necessity of it,*** *for not only understanding in terms of medicine but also just in general society, as if we hope to progress through anything we have to be able to understand what we're doing* ***and having that knowledge that lets us progress more****."* The pair "Unnecessary-Necessary" was not chosen because the expert panel suggested finding adjectives that portray true opposing words without resorting to using a negative prefix. The word "essential" was shared by students often and represented a true opposing adjective to unnecessary. However, this practice was not possible with all of the items given the task of providing students with words with which they are easily acquainted.

Pleasant-Unpleasant (*Emotional Satisfaction*)

This item is an original from the ASCIv2. All students agreed this item evoked affective processes, it was well-understood, and interpreted similarly throughout the interviews. Student 1 gave a concise summary of their interpretation of this item that was comparable to other students' views of this item, *"I just really enjoy it. Yeah. So like* ***it feels good*** *when I get something right or when I like connect to things."*

Overwhelming-Manageable (*Emotional Satisfaction*)

Several students observed that studying chemistry, understanding chemistry, and working with chemistry concepts can feel overwhelming often because it takes a lot of time and effort. Student 4 said, *"**I think like the whole subject is overwhelming** because you have to dedicate yourself to it if you want to evolve, especially for someone who's majoring in it. A lot of effort goes into it. Sometimes you might get a little bit behind and it's hard to catch up and **then you feel like your whole world is ending**."* Additionally, as student 4 indicated, sometimes their effort did not pay off as expected and that also felt overwhelming. On the other hand, there were instances that even when chemistry was difficult and time consuming, it could also be manageable. Student 11 indicated, ***"I think I would almost put manageable on the opposite side of overwhelming**. ...Or maybe if you're just thinking about it in the sense of time, time consuming versus manageable."*

Applicable-Not Applicable (*Utility*)

Applicability was a prominent idea students talked about when describing chemistry. Some students argued that chemistry could be applied to their everyday lives, or to a larger global scale, and other students argued that chemistry might only be applicable to people who major in it. When asked if chemistry was applicable, student 1 replied, *"I guess not, or maybe like the basics of chemistry would be like what I said with the salad dressing [earlier] maybe like if you want to have like a more complete understanding of the world, then you could apply chemistry that way. But other than that, **like if I were only a music major, I probably would say it's not applicable**."* This item was added to the U factor.

Satisfying-Frustrating (*Emotional Satisfaction*)

This item is an original item from the ES factor. Students were able to connect with this item and interpret it as part of an affective mental process regarding chemistry. Some students talked about this item describing how they feel after success or failure with a cognitive task in chemistry. Student 7 said, *"So I guess it's, I don't know, it's* ***a feeling that I get after working it out****. Yeah. Or like* ***a feeling that I get after I understand****."*

Important-Not Important (*Utility*)

This item was generated after some students explicitly said it was important to have knowledge of chemistry. This item was linked to the utility of the discipline to solve world problems (*i.e.*, global warming, etc.), as well as students' immediate needs for their major and future career goals. Student 2, a biology (botany) major said, *"I just feel like it's stuff I'll be like, even if I'm not in the chemistry field, I still feel like as I'm still a science major,* ***it's still something I'm going to be using****. It's going to be a groundwork for what I'm going to be using later. So it's still like the stuff I have to know. So like even if it's kind of boring, it's just,* ***it's ground work, you've got to know it****."* This item was added to the U factor.

Enjoyable-Dull (*Emotional Satisfaction*)

Many students shared that they found chemistry to be an enjoyable subject. Student 1 provided a helpful analogy when they shared, *"I think* ***it could be like not fun but still be enjoyable****.*

Like, I don't know, like when you get a massage, like it hurts, but you're like, oh yeah, yeah. But it's enjoyable. Or like ***the end result is enjoyable.****"* Other students found the subject boring and tedious. Student 4 shared, *"****Chemistry is boring when it becomes tedious****, like just doing the same thing over and over again."* After careful consideration for a pair of adjectives that described what the students were sharing, with the help of the expert panel, this item was added to the ES factor. "Dull" was chosen over "boring" because it is a word that can better elicit feelings toward the discipline rather than a specific class.

Descriptive Statistics

Students in two sections of OCII in the fall semester of 2019 were asked to complete the ASCI-UE starting two days before each of the exams in the term including the final exam. The data reported in this chapter comes from Exam 1 and the Final Exam. The other two instances when data was collected associated with exams 2 and 3 can be found in Appendix C. Table 6.1 displayed the observed mean score of the two factors in the ASCI-UE, namely U and ES at the beginning and end of the semester for high- and low-achievement groups.

Based on these observed mean scores, it appears that the high-achievement students displayed more positive ES than the low-achievement students both at the beginning and end of the semester. Interestingly, at the beginning of the semester there was not much difference in the U factor between these two groups; however, that small difference widened toward the end of the semester.

Observing the trends for each group throughout the semester, we saw that for the low-achievement group both the U and ES factors appeared to decline throughout the semester. For the high-achievement group a small increase was observed for ES. The U factor showed a noticeable increase throughout the semester.

Table 6.1. Descriptive Statistics for High- and Low-Achievement Groups at the Beginning and End of the Semester

	U-Pre		ES-Pre		U-Post		ES-Post	
Achievement	Mean	S.D	Mean	SD	Mean	S.D	Mean	S.D
Low[a]	5.79	1.19	3.83	1.28	5.58	1.39	3.44	1.26
High[b]	5.88	1.19	4.45	1.17	6.17	0.99	4.57	1.20

[a]Low-achievement group Pre (n = 157); Post (n = 143).
[b]High-achievement group Pre (n =105); Post (n = 106).

CFA and Measurement Invariance Testing

The results of the CFA for the ASCI-UE and the PC scales for the entire course are displayed in Tables 6.2 and 6.3. All analyses converged and showed acceptable data-model fit. For PC, the RMSEA index did not indicate appropriate data-model fit, however, this fit index has shown to be less reliable with short instruments (Kenny, Kaniskan and McCoach, 2015) such as the PC scale that contains only four items. Furthermore, for the PC factor a large decline in data-model fit was observed at the end of the semester. While the fit remained appropriate, the decline was worth noting.

In addition to the CFA results, Tables 6.2 and 6.3 also displayed the Omega coefficient, a measure of reliability, for each of the scales or subscales of the instruments used in this study. In each instance, the Omega coefficient indicated a strong reliability for each subscale with values equal to or above 0.800.

Table 6.2. Confirmatory Factor Analysis of ASCI-UE factors at the Beginning and End of the Semester in OCII

	N	χ^2	*df*	*p*	CFI	SRMR	RMSEA	Omega U	Omega ES
Pre	291	60.381	26	<0.001	0.955	0.046	0.067	0.800	0.875
Post	249	62.557	26	<0.001	0.947	0.057	0.075	0.869	0.894

U = Utility. ES = Emotional Satisfaction.

Table 6.3. Confirmatory Factor Analysis of Perceived Competence Scale at the Beginning and End of the Semester in OCII

	N	χ^2	*df*	*p*	CFI	SRMR	RMSEA	Omega PC
Pre	291	15.532	2	<0.001	0.978	0.021	0.152	0.918
Post	249	30.497	2	<0.001	0.920	0.032	0.239	0.915

With evidence of appropriate data-model fit for the entire classroom, a series of measurement invariance tests were performed to support longitudinal or group comparisons. Some of the tests were conducted for the entire class (e.g., for longitudinal comparisons for the class), and some were conducted to produce evidence of appropriate comparisons between subgroups (e.g., high- and low-achievement groups). Additionally, in this section, the results of measurement

invariance testing for data collected in Chile and in the US in Spanish and English, respectively were reported.

Longitudinal measurement invariance testing for ASCI-UE held to the strict level for the entire class (see Table 6.4) providing evidence that longitudinal comparisons of observed mean scores were supported. A paired samples *t*-test was conducted for all students, indicating that no evidence of significant difference was observed for either of the factors after a Bonferroni adjustment (see Table S6.6 in Appendix C).

Longitudinal invariance testing was also conducted for PC for the entire class; however, the results indicated that PC holds only to the metric level (see Appendix C), which provided evidence of similar factor meaning across time, but no comparisons were supported.

Table 6.4. Longitudinal Measurement Invariance Testing for ASCI-UE

	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural	249.444	129	<0.001	0.936	0.054	0.056	-	-	-	-	-	-
Metric	273.437	136	<0.001	0.927	0.065	0.058	23.993	7	0.001	0.009	0.011	0.002
Scalar	287.188	143	<0.001	0.924	0.066	0.058	13.751	7	0.056	0.003	0.001	0.000
Strict	299.231	152	<0.001	0.922	0.080	0.057	12.043	9	0.211	0.002	0.014	0.001

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are all students combined (n = 291) for pre-exam 1 and pre-exam 4. The configural model is a comparison model without constraints. The metric model adds the constraint of equal factor loadings. The scalar model adds the constraint of equal intercepts. The strict model adds the constraint of equal error variances. Each constraint was added one at a time. *df* = degrees of freedom.

As previously mentioned, a focus of this study was to compare high- and low-achieving students in OCII at the beginning (Table 6.5) and at the end of the semester (Table 6.6), considering

both U and ES factors simultaneously. The same groups comparison was desired for PC. In order to provide evidence to support these comparisons, measurement invariance testing was conducted for the two groups at each time point. Tables 6.5 and 6.6 show that these comparisons were supported for U and ES. However, Tables S6.8 and S6.9 in Appendix C show that comparisons between the two subgroups were not supported for PC.

Table 6.5. Measurement Invariance Testing for High- and Low-Achievers at the Beginning of the Semester

	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural	97.958	52	<0.001	0.934	0.057	0.082	-	-	-	-	-	-
Metric	109.232	59	<0.001	0.927	0.082	0.081	11.274	7	0.127	0.007	0.025	0.001
Scalar	120.630	66	<0.001	0.921	0.089	0.079	11.398	7	0.122	0.006	0.007	0.002
Strict	118.396	75	0.001	0.937	0.111	0.066	2.234	9	0.987	0.016	0.022	0.013

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups high-achievers (n = 105) and low achievers (n = 157) for pre-exam 1. The configural model is a comparison model without constraints. The metric model adds the constraint of equal factor loadings. The scalar model adds the constraint of equal intercepts. The strict model adds the constraint of equal error variances. Each constraint was added one at a time. *df* = degrees of freedom.

Table 6.6. Measurement Invariance Testing for High- and Low-Achievers at the End of the Semester

	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural	91.591	52	<0.001	0.942	0.071	0.078	-	-	-	-	-	-
Metric	106.927	59	<0.001	0.930	0.098	0.081	15.336	7	0.032	0.012	0.027	0.003
Scalar	116.672	66	<0.001	0.926	0.100	0.079	9.745	7	0.203	0.004	0.002	0.002
Strict	132.612	75	<0.001	0.916	0.113	0.079	15.940	9	0.068	0.010	0.013	0.000

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are high achievers (n = 106) and low achievers (n= 143) for pre-exam 4. The configural model is a comparison model without constraints. The metric model adds the constraint of equal factor loadings. The scalar model adds the constraint of equal intercepts. The strict model adds the constraint of equal error variances. Each constraint was added one at a time. *df* = degrees of freedom.

Since comparisons between subgroups for the ASCI-UE were supported, a MANOVA was conducted in order to determine the differences between these two groups at each time point for U and ES. The results were reported in Table S6.7 in Appendix C, which showed that all comparisons were significant except for U at the beginning of the semester which displayed no evidence of significant difference between the groups. Accompanying these results, Table 6.7 contained the effect sizes calculated for the comparisons. Not surprisingly, a negligible effect size was observed for U-Pre between the groups, but a medium effect size was observed for U-Post favoring the high-achievement group. For ES-Pre a significant difference with a medium effect size was found between groups, and a large effect size for ES-Post, in both instances favoring the high-achievement group.

Table 6.7. Effect Size of the Difference Between High- and Low-Achievement Groups

	U-Pre	ES-Pre	U-Post	ES-Post
Cohen's d	0.08	0.50	0.48	0.92

An important part of this study was to gather evidence that the ASCI-UE could function in similar ways in English and Spanish in two different countries. Data was collected online in Chile with the Spanish version of the ASCI-UE for students in general and organic chemistry courses. 228 complete responses were obtained in the spring of 2020. The deidentified data was joined to the data collected in the U.S. and measurement invariance testing was conducted. The results of this test indicated that metric invariance holds between the groups (see Table 6.8) suggesting that the factor meaning is similar between the groups, although no comparisons were supported at this

level (Sass, 2011; Rocabado *et al*., 2020). Partial scalar invariance was attempted by releasing one item intercept. After careful scrutiny of the results, it was decided that Items 8 and 3 displayed the most differing intercepts between the groups and could be the source of noninvariance. Item 8 was released to be freely estimated first; however, this release made little difference in the model fit. The release of Item 3 to be freely estimated gave similar results. Thus, it was concluded that with metric invariance achieved, we can infer that creating this instrument in both languages simultaneously helped attain similar factor meaning across two countries.

Table 6.8. Measurement Invariance Testing for ASCI-UE Between U.S. and Chile

	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural	116.297	51	<0.001	0.956	0.047	0.070	-	-	-	-	-	-
Metric	132.906	58	<0.001	0.950	0.076	0.070	16.609	7	0.020	0.006	0.029	0.000
Scalar	168.387	65	<0.001	0.930	0.089	0.078	35.481	7	<0.001	0.020	0.013	0.008
Partial Scalar (8)	166.805	64	< 0.001	0.931	0.090	0.079	33.899	6	<0.001	0.019	0.014	0.009
Partial Scalar (3)	166.595	64	<0.001	0.931	0.088	0.079	33.689	6	<0.001	0.019	0.012	0.009

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are OCII students in USA (n = 291) and chemistry students in Chile (n = 228). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. The scalar model adds the constraint of equal intercepts for both groups. Partial scalar models release the constraint of equal intercepts for one item at a time. *df* = degrees of freedom.

Correlation of Utility, Emotional Satisfaction, Perceived Competence, and Achievement

One of the standards of validity is a construct's relation to other variables (Arjoon *et al*., 2013; AERA *et al*., 2014). This standard arises from the idea that in order to gather information

about the construct, examining convergent and discriminant relationships to other well-established constructs could provide evidence about what the construct is (Messick, 1980). In this study, I chose to investigate the relationship between the ASCI-UE factors, PC, and achievement measured by exam scores.

At the beginning of the semester, U displayed significant correlations with ES and PC, but no evidence of a significant correlation with Exam 1 (Table 6.9). However, at the end of the semester, U was significantly correlated to ES, PC, and the Final Exam (Table 6.10). On the other hand, ES displayed significant correlations with U, PC, and exam scores at both times during the semester. Similarly, PC was significantly correlated to exam scores both at the beginning and end of the semester; however, a correlation twice as strong was observed at the end of the semester. PC was also significantly correlated to both U and ES at the beginning and end of the semester as predicted. Not surprisingly, PC displays a stronger correlation to ES than U, since PC was chosen as a proxy of the original *Intellectual Accessibility* factor which has a well-established strong correlation to ES.

Table 6.9. Correlations Between ASCI-UE, PC, and Achievement at the Beginning of the Semester

	U	ES	PC	Exam1
U	1.000			
ES	0.428*	1.000		
PC	0.273*	0.568*	1.000	
Exam1	0.048	0.228*	0.268*	1.000

*significant to the 0.01 level

Table 6.10. Correlations Between ASCI-UE, PC, and Achievement at the End of the Semester

	U	ES	PC	Final
U	1.000			
ES	0.344*	1.000		
PC	0.310*	0.695*	1.000	
Final	0.212*	0.352*	0.431*	1.000

*significant to the 0.01 level

Structural Equation Modeling

As previously discussed, a longitudinal comparison of PC was not supported with these data. Therefore, the subsequent analysis was done only with ASCI-UE and exam scores at the beginning and end of the semester. A reciprocal causation model (Pekrun, 2006; Pekrun, Maier and Elliott, 2009; Gibbons and Raker, 2018; Gibbons *et al.,* 2018) was tested following the logic of a reciprocal relationship between measures of attitude and achievement across time. Structural equation modeling (SEM) was utilized in which several nested models were tested (see Appendix C). Model A showed the best theoretical and statistical results of the models that were explored. Figure 6.2 showed a simplified pictorial representation of the SEM. In this figure it was observed that both ES-Pre and ES-Post had directional linear relationships to the subsequent exam score. These relationships were small, but significant. Conversely, the linear relationships between U-Pre and U-Post to the subsequent exam scores were non-significant. Exam 1 showed small but significant relationships to both ES- and U-post measures. Finally, as expected there were strong relationships between the pre and post measures of ASCI-UE as well as Exam 1 with Final Exam. The model fit statistics for all models are reported in Table 6.11.

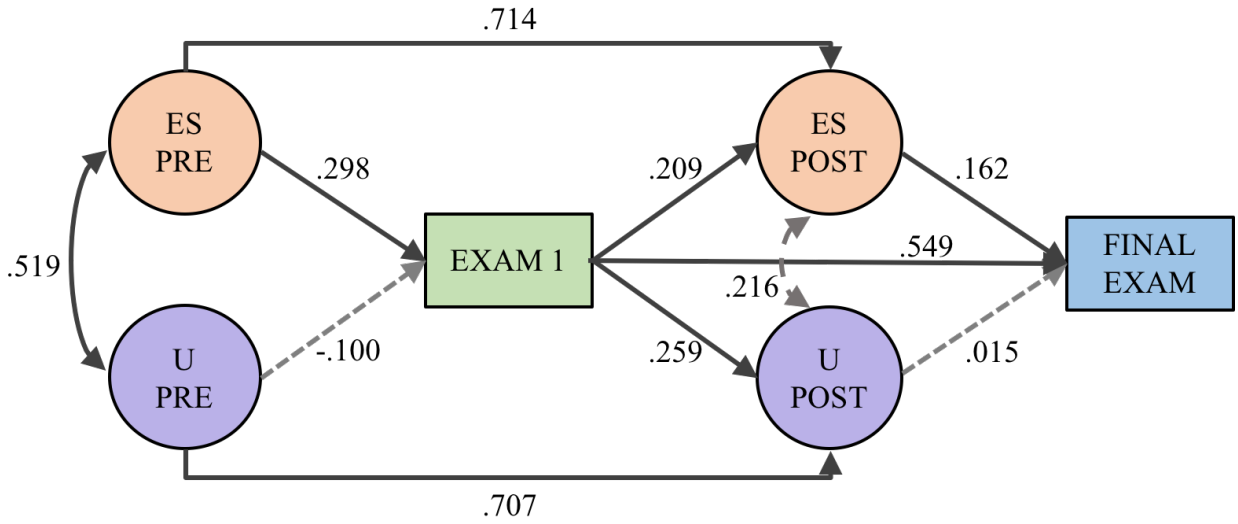


Figure 6.2. Simplified Pictorial Representation of Model A SEM displaying a Reciprocal Causation Model Relationship between ASCI-UE and achievement measures (exams).

Table 6.11. Data-Model Fit Indices for Nested SEM Models

Model	χ^2	***df***	***p***	**CFI**	**SRMR**	**RMSEA**
Model A	302.575	161	<0.001	0.934	0.052	0.054
Model B	320.875	162	<0.001	0.926	0.061	0.057
Model C	315.302	162	<0.001	0.929	0.058	0.056
Model D	314.863	162	<0.001	0.929	0.062	0.056
Model E	308.231	162	<0.001	0.932	0.054	0.054

Discussion

The ASCIv2 is a valuable instrument to measure attitude toward the discipline of chemistry and has been widely and effectively used in many classrooms in the U.S. (Brandriet *et al.*, 2011;

Xu and Lewis, 2011; Brandriet, Ward, and Bretz, 2013; Xu, Villafañe, and Lewis, 2013; Cracolice and Busby, 2015; Chan and Bauer, 2014, 2016; Mooring *et al*., 2016; Underwood, Reyes-Gastelum, and Cooper, 2016; Stanich *et al*., 2018; Nenning *et al.,* 2019; Rocabado *et al*., 2019) and around the world (Xu, Southam, and Lewis, 2012; Xu, Alhoosani, Southam, and Lewis, 2015; Vishnumolakala *et al*., 2017; Vishnumolaka *et al*., 2018; Damo and Prudente, 2019). It has also been translated to several languages (Khaveci, 2015; Sen, Yilmaz, and Temel, 2016; Montes, Ferreira, and Rodriguez, 2018). It is a short instrument and easy to distribute without concern of survey fatigue. Its benefit has been reported in a variety of settings to evaluate the effectiveness of classroom interventions (*i.e*., Mooring *et al*., 2016) or to investigate attitude-achievement relationships (*i.e*., Brandriet, Ward and Bretz, 2013; Xu, Villafañe and Lewis, 2013; Villafañe and Lewis, 2016; Rocabado *et al*., 2019). The theoretical underpinnings of this instrument were based on the theoretical notion that attitude encompasses cognitive and affective domains which lead to behavioral intentions (Ajzen and Fishbein, 2000). However, cognitive interviews had not yet been conducted to investigate students' interpretations of each of the items. This practice was particularly valuable after observing several idiosyncratic behaviors with particular items in a consistent manner across a variety of studies (Xu *et al*., 2015; Montes, Ferreira and Rodríguez, 2018; Rocabado *et al*., 2019).

Cognitive interviews were conducted in English and Spanish with students in the U.S. and in Chile, respectively. These interviews informed on the meaning students attached to the items in the original ASCIv2, some which were consistent with the theory and some which were not. For instance, although students agreed that the item "Comfortable-Uncomfortable" elicited affective evaluation processes, most of them did not think this item was appropriate for the evaluation of

chemistry and many students could not respond with confidence to this item. Therefore, this item was removed and another item "Depressing-Exciting" replaced it to represent the intensity of the feelings that students were expressing towards the discipline.

Another example was the item "Challenging-Not Challenging" which most students agreed elicited cognitive evaluation processes. However, some students could also argue that this item could elicit affective evaluation process as well. Nevertheless, the biggest issue with this item was the fact that students talked about chemistry being 'challenging' as being both positive and negative. Therefore, the meaning of this item could be interpreted in various ways, which could be the reason why this item has been shown to display low factor loadings and cross loadings (Xu *et al*., 2015; Montes, Ferreira and Rodríguez, 2018; Rocabado *et al*., 2019). In this case, the cognitive interviews helped elucidate the multiple connotations of this item and some of the reasons for its idiosyncratic behavior. For more information on this item see Appendix C.

Most importantly, students shared that evaluating the utility of chemistry was a worthwhile endeavor, which was the inspiration behind creating a new instrument that measured *Utility*. Students in Chile and in the U.S. first shared adjectives such as "relevant" or "applicable" in the interviews. Then these items were tested in subsequent interviews. These items resonated well with all students indicating that measuring the *Utility* of chemistry was needed. Bauer (2008) had generated items for an *Interest and Utility* factor in the original ASCI instrument, which indicates that Utility was important for his respondents as well. This practice of involving the respondents in the item-generation process is an underused exercise, perhaps because it requires extensive resources. However, response process interviews along with the expert panel review have been

instrumental for the development of the original ASCI to its current development presented in this chapter leading to successful results.

After conducting the interviews with the students, the idea was to add the U factor as the third factor to the ASCIv2. However, when I tested this three-factor instrument I observed a result which was not statistically allowed due to factor correlations that were too high. This result indicated that the addition of a third factor conflated the relationships between the factors, therefore I decided to keep the U factor with the ES factor in the ASCI-UE. In this sense we highlight the importance of going through a rigorous process of instrument development and checking for validity evidence to support meaningful results and inferences.

In this study I endeavored to gather various aspects of validity evidence based on *The Standards* (AERA *et al.*, 2014). Response process and content validity evidence were shown in the course of instrument development and refinement. The next aspect of validity evidence I have demonstrated was internal structure validity. The data collected with the ASCI-UE were subject to CFA which showed appropriate model fit at each time point indicating a stable factor structure across time. Furthermore, longitudinal measurement invariance testing was conducted which held to the strict level, indicating that longitudinal comparisons were supported. I was also interested in investigating differences in attitude between high- and low-achieving students, therefore measurement invariance testing was also performed at each time point for both subgroups, which also held to the strict level. All of these analyses provided ample evidence of internal structure validity for inferences made for the entire sample as well as for subgroups and across time. This evidence provided confidence in the interpretation of the results for the groups. For instance,

paired-samples *t*-tests were conducted to investigate the change in U and ES during the semester for the entire OCII class. No evidence of significant difference was observed. Since power analysis indicates that changes at the level of a small effect would be observable in this case, this result suggests that students in OCII, who already have had at least 3 other chemistry courses prior to their current course, may have already formed stable attitudes toward chemistry. In lower-level courses, attitude tends to decline over the course of the semester (*i.e.,* Chapter 5); however, in this case no evidence of significant change was observed.

As shown throughout this work, often subgroup comparisons display noticeable differences between groups. High- and low-achieving students' attitudes in this case showed significant differences in ES at both times in the semester, and U at the end of the semester with medium to large effect sizes. In each case the differences favored the high-achieving students, suggesting that higher attitude scores accompany higher performance in the course. Interestingly, both groups of students began the course with a similar U score, yet toward the end of the semester, U declined for low-achieving students, and increased for high-achieving students, widening the gap at the end of the semester. This result indicates that throughout the semester, low-achieving students found chemistry less useful than they did at the beginning, perhaps due to a realization that careers that require this course (*i.e*., medicine) might no longer be feasible. Conversely, high-achieving students' U score increases throughout the semester, indicating that these students continue to internalize the importance and utility of the discipline particularly because for many of them OCII might be the last chemistry course they will take. Additionally, it would be interesting to conduct a similar study in other courses such as first semester general and organic

chemistry when students have not yet experienced these courses and may not be confident about the utility in these new contexts.

In an effort to demonstrate validity evidence of relation to other variables, I chose to collect data with a *Perceived Competence* scale as a proxy for the IA factor that was removed from the instrument as previously discussed. Although this short instrument displayed appropriate model fit at the beginning and end of the semester, longitudinal and subgroup measurement invariance testing did not hold. Therefore comparisons with this instrument were not supported. Longitudinal measurement invariance held to the metric level indicating similar factor meaning across time. Longitudinal or subgroup comparisons were not supported; however, I was able to explore correlational relationships between the ES, U, and PC factors, as well as their relationships to the exam scores both at the beginning and end of the semester. All of the relationships were significant correlations except U-pre with Exam 1. This result indicates that even students who did poorly on exams believed chemistry was useful based on a high observed mean score for U. Therefore, it follows that at the beginning of the semester, no matter how students will perform on their test, they believe chemistry has utility in their lives. At the end of the semester the U scores followed a similar trend than the exam scores, therefore a significant correlation was observed. On the other hand, ES and PC have both strong correlations to the exam scores throughout the semester and to each other as well. Since PC was chosen as a proxy of IA due to conceptual construct similarities, it is no wonder that the correlation between PC and ES is about twice as high as the correlation between PC and U given the historical strong correlation of IA and ES. This result is evidence that the relationships between constructs are the way we expected.

A SEM was conducted for the reciprocal causation relationship between ASCI-UE and exam scores. Model A displayed the best theoretical (Pekrun, 2006; Pekrun, Maier and Elliott, 2009; Gibbons and Raker, 2018; Gibbons *et al.,* 2018) and statistical results. This model showed the small but significant relationships between ES and the subsequent exams, while the U-exam relationships were non-significant. This occurrence is explained by the fact that even students who performed poorly on the exam reported that chemistry was useful. Even though the gap in U widened at the end of the semester for high- and low-achieving students, the pattern was still that students found chemistry highly useful regardless of their grade. Therefore, these relationships were less noticeable than the relationship between ES and exam scores. This result is encouraging in the sense that students' perception of utility is more stable for students in OCII. However, ES is more closely tied to whether students perform well or not on exams, and therefore ES is less stable.

Finally, the ASCI-UE was simultaneously created in English and Spanish. Data were collected in the U.S. and in Chile in English and Spanish, respectively. Measurement invariance testing was conducted to investigate the extent to which the internal structure of the instrument held in two languages and two countries. Metric invariance was achieved, indicating that construct meaning was similar across the groups. This exciting result is the culmination of a rigorous process of instrument development across different countries and languages that resulted in an instrument that can be utilized in future cross-country investigations. Therefore, I encourage researchers and practitioners to use this instrument in a variety of settings and in both languages when their research interests align with the ASCI-UE constructs. I encourage researchers in other Spanish-speaking countries and regions as well as diverse English-speaking settings to test this instrument

in their sites and report their results to continue learning about the stability of this instrument in both languages.

Implications

The data collected with the ASCI-UE in OCII provided interesting and important insights on student attitudes toward the discipline of chemistry. The subgroup comparison results yielded interesting implications, particularly for the U factor. Interestingly, the U factor showed no evidence of significant difference between subgroups (high- and low-achievement) at the beginning of the semester, but a significant difference at the end. This result suggests the importance of explicitly teaching utility of the discipline beyond just medicine. Students who perceived chemistry as less useful at the end of the semester could have found it so because of potential change in careers based on their low exam scores. One of our purposes as instructors and researchers is to provide students a sense of utility of this subject regardless of their future career goals. The students that go through our chemistry courses, whether they will actively pursue careers in which chemistry will be an active component or not, should be, at the very least, informed citizens that can see how chemistry as a central science is useful in any realm. Therefore, encouraging students to find utility in the subject of chemistry is critical. For instance, Wang and colleagues (2020) suggest a simple classroom intervention designed to improve students' sense of utility of the subject of chemistry which also showed to improve students' exam scores.

This and other studies, including the ones discussed in this dissertation, have shown that stronger direct relationships exist between the affective measures and achievement than the cognitive measures and achievement in chemistry (*i.e.*, Rocabado *et al*., 2019; Wang *et al*., 2020). It is worthwhile for researchers and practitioners to think about the pedagogies that are used in the classroom and whether they are designed to influence affect and emotion or not. Given the evidence that affect can significantly influence achievement, a call for a greater focus on investigating emotions in chemistry was given by Flaherty (2020) and it is worth reiterating here after the evidence provided in this chapter. This focus can lead to a stronger impact on achievement for students in chemistry, which may also lead to greater retention of students in STEM fields.

In this work I have presented an additional instrument to measure attitude that includes a refined *Emotional Satisfaction* factor and a new *Utility* factor, a salient construct of significance for students in general and organic chemistry courses. This new instrument was created using *The Standards* (AERA *et al*., 2014) of measurement that prescribe the need to gather several aspects of validity evidence when developing and using instruments in research studies. This instrument does not replace the original ASCIv2, but rather it presents an additional choice for measuring attitude for instructors and researchers to use in their studies. I urge instructors and researchers to select which of the instruments serve their study design best.

Undertaking instrument development and refinement is a rigorous process which requires time and significant resources. Doing this process in two languages and in two countries was even more difficult. However, the resulting instrument is a tool for cross-country and cross-language studies. In this chapter I have presented evidence of similar construct meaning across the two

languages and countries. This encouraging result is evidence of the success of this project, even under adverse circumstances of data collection in Chile during the COVID-19 pandemic. Further data collection with the Spanish instrument under more favorable circumstances might provide the evidence to support cross-country comparisons. I encourage researchers and instructors in both countries to gather more data with this instrument and continue to conduct rigorous analyses to test the feasibility of cross-country comparisons. I also encourage researchers and practitioners of other countries to use this new instrument and gather further validity evidence in diverse settings.

Furthermore, by simultaneously creating the ASCI-UE in English and Spanish, I together with my collaborators, endeavored to create a model for future instrument development that would allow cross-country comparisons even while the instrument is administered in different languages. This model presented herein should be an example for other researchers to follow when developing instruments that may be of use across the world.

Limitations

Limitations in this study arise from a convenient sample. In the fall semester of 2019 I had access to two sections of OCII taught be the same instructor. The instrument was piloted in other courses, yet the investigation proceeded in OCII. Therefore, interpretation of the results is limited to students experiencing OCII and may be problematic to extrapolate to students in other courses. Similarly, the data collected in Chile also came from a convenient sample. These data were collected online during the spring 2020, in the middle of a global pandemic. Data collection was

difficult as instructors were moving their courses to online modality and many were unable or unwilling to provide the means to collect data with their students. The data came from students in different courses and different universities, which may be the reason why higher levels of invariance were not achieved given such a varied sample.

Another limitation of this study is that the volunteer interviewees in the U.S. were mostly female students. Only one male student volunteered to be interviewed. In the recruiting process I randomly recruited about 200 students to participate from general chemistry and organic chemistry. Only one male student volunteered, while several females participated despite recruiting similar numbers of male and female students.

Finally, using a purely quantitative approach of analysis of data collection with the new instrument is a limitation particularly with this sample of OCII students. During this course many students make a decision about their future career goals. Students who wanted to go into healthcare professions that did not obtain a high grade may be contemplating alternative careers. Capturing their attitudes and thought process during this time with qualitative data would have enriched the results and inferences made in this chapter.

References

AERA, APA and NCME, (2014), *Standards for educational and psychological testing*, Washington, DC: American Psychological Association.

Ajzen L. and Fishbein M., (2000), *Attitudes and the attitude-behavior relation: Reasoned and automatic processes*. In European review of social psychology (Vol. 11) Stroebe W. and Hewstone M. (Eds.), Chichester, England: Wiley, pp. 1-33.

Arjoon J. A., Xu X. and Lewis J. E., (2013), Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence, *J. Chem. Educ.*, **90**, 536–545. DOI:10.1021/ed3002013

Bandura A., (1977), Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.,* **84**(2), 191-215. DOI: 10.1037/0033-295X.84.2.191

Bauer C. F., (2008), Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts, *J. Chem. Educ.*, **85**, 1440−1445. DOI:10.1021/ed085p1440

Brandriet A. R., Ward R. M. and Bretz S. L., (2013), Modeling meaningful learning in chemistry using structural equation modeling, *Chem. Educ. Res. Pract.*, **14**, 421-430. DOI:10.1039/C3RP00043E

Brandriet A. R., Xu X., Bretz S. L. and Lewis J. E., (2011), Diagnosing changes in attitude in first-year college chemistry students with a shortened version of Bauer's semantic differential, *Chem. Educ. Res. Pract*., **12**, 271-278. DOI:10.1039/C1RP90032C

Bulmer M. G., (1979), *Principles of Statistics*. New York: Dover.

Chan J. Y. K. and Bauer C. F., (2014), Identifying at-risk students in general chemistry via cluster analysis of affective characteristics, *J. Chem. Educ.*, **91**, 1417-1425. DOI:10.1021/ed500170x

Chan J. K. and Bauer C. F., (2016), Learning and studying strategies used by general chemistry students with different affective characteristics, *Chem. Educ. Res. Pract*., **17**, 675-684. DOI:10.1039/C5RP00205B

Chen F. F., (2007), Sensitivity of goodness of fit indexes to lack of measurement invariance, *Struct. Equ. Modeling*, **14**(3), 464-504. DOI: 10.1080/10705510701301834

Cheng-Hsien L., (2016), Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares, *Behav. Res.,* **48,** 936-949. DOI:10.3758/s13428-015-0619-7

Cohen J., (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ.

Cortina J. M., (1999), What is coefficient alpha? An examination of theory and applications, *J. App. Psychol.,* **78,** 98-104. DOI: 10.1037/0021-9010.78.1.98

Cracolice M. S. and Busby B. D., (2015), Preparation for College General Chemistry: More than just a matter of content knowledge acquisition, *J. Chem. Educ.,* **92**, 1790-1797. DOI:10.1021/acs.jchemed.5b00146

Cronbach L. J., (1951), Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**, 297-334. DOI: 10.1007/BF02310555

Curtis B. and Curtis C., (2017), I-Depth Interviewing – The Interactive Base. In *Social Research: A Practical Introduction,* SAGE Publications, Inc.

Damasio A. R., (1994), Descartes' error: Emotion, rationality and the human brain, Putnam 352: New York, NY

Damo K. L. and Prudente M. S., (2019), Investigating students' attitude and achievement in organic chemistry using interactive application, *Assoc. Computing Machinery,* 36-41, Tokyo, Japan. DOI:10.1145/3306500.3306562

Flaherty A. A., (2020), A review of affective chemistry education research and its implication for future research, *Chem. Educ. Res. Pract.,* **21,** 698-713. DOI: 10.1039/C9RP00200F

Gibbons R. E. and Raker J. R., (2018), Self-beliefs in organic chemistry: Evaluation of a reciprocal causation, cross-lagged model*, J. Res. Sci. Teach.,* **56**(5), 598-615. DOI:10.1002/tea.21515
Gibbons R. E., Xu X., Villafañe S. A. and Raker J. R., (2018), Testing a reciprocal causation model between anxiety, enjoyment and academic performance in postsecondary organic chemistry, *Educ. Psychol.* **38**(6), 838-856. DOI:10.1080/01443410.2018.1447649
Guba E. G. and Lincoln Y. S., (1983), Epistemological and methodological bases of naturalistic inquiry, *Educ. Comm. Tech. J.,* **4**(30), 311-333. ISSN 0148-5806
Harter S., (1982), Perceived competence scale for children, *Child Development, 53*(1), 87-97. https://www.jstor.org/stable/1129640
Hu L. T. and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Struct. Equ. Modeling,* **6**(1), 283-292. DOI: 10.1080/10705519909540118
Kenny D. A., Kaniskan B. and McCoach D. B., (2015), The performance of RMSEA in models with small degrees of freedom, *Sociol. Methods Res.*, **44**(3), 486-507. DOI:10.1177/0049124114543236
Kahveci A., (2015), Assessing High School Students' Attitudes Toward Chemistry with a Shortened Semantic Differential, *Chem. Educ. Res. Pract*., **16**, 283−292. DOI: 10.1039/C4RP00186A
Kline R. B., (2015), *Principles and Practice of Structural Equation Modeling*, 3rd ed., Guilford Press: New York.
Komperda R., Pentecost T. C. and Barbera J., (2018), Moving beyond alpha: A primer on alternative sources of single-administrations reliability evidence for quantitative chemistry education research, *J. Chem. Educ*., **95**, 1477-1491. DOI: 10.1021/acs.jchemed.8b00220
Krech D., Crutchfield R. and Ballachey E., (1962), *Individual in society,* New York: McGraw-Hill.
Messick S., (1980), Test validity and ethics of assessment, *Am. Psychol.,* **35**(11), 1012-1027. DOI: 10.1037/0003-066X.35.11.1012
Montes L. H., Ferreira R. A. and Rodriguez C., (2018), Explaining Secondary School Students' Attitudes Towards Chemistry in Chile, *Chem. Educ. Res. Pract*., **19**(2), 533−542. DOI:10.1039/C8RP00003D
Mooring S. R., Mitchell C. E. and Burrows, N. L., (2016), Evaluation of a flipped, large enrollment organic chemistry course on student attitude and achievement, *J. Chem. Educ.*, **93**, 1972−1883. DOI:10.1021/acs.jchemed.6b00367
Muthén L. K. and Muthén B. O., (1998-2007), *Mplus User's Guide*, 5th ed., Muthén & Muthén: Los Angeles, CA.
Nenning H. T., Idarraga K. L., Salzer L. D., Blaske-Rechek A. and Theisen R. M., (2019), Comparison of student attitudes and performance in an online and face-to-face inorganic chemistry course, *Chem. Educ. Res. Pract*., **21,** 168-177. DOI:10.1039/C9RP00112C
Pekrun R., Maier M. A., Elliot A. J., (2009), Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance, *J. Educ. Psychol*., **101**(1), 115−135. DOI:10.1037/a0013383
Pekrun R.. (2006), The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice, *Educ. Psychol. Rev.,* **18,** 315–341. DOI:10.1007/s10648-006-9029-9

Pintrich P. R. and Schrauben B., (1992), Students' Motivational Beliefs and Their Cognitive Engagement in the Classroom Academic Tasks, In *Student Perceptions in the Classroom,* Schunk D. H. and Meece J. L. (Eds), Routledge: New York, NY

Rocabado G. A., Kilpatrick N. A., Mooring S. R., and Lewis J. E., (2019), Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course, *J. Chem. Educ.,* **96**(11), 2371-2382. DOI:10.1021/acs.jchemed.9b00516

Rocabado G. A., Komperda R., Lewis J. E. and Barbera J., (2020), Addressing diversity and social inclusion through groups comparisons: A primer on measurement invariance testing, *Chem. Educ. Res. Pract.,* **21,** 969-988. DOI:10.1039/D0RP00025F

Rosenberg J. and Hovland I., (1960), *Cognitive, affective, and behavioral components of attitudes,* In Attitude rganization and change: An analysis of consistency among attitude components, Rosenberg M. J. et al., (Eds), New Haven, CT: Yale University Press, pp. 1-14.

Ryan R. M. and Deci E. L., (2017), *Self-Determination Theory: Basic psychological needs in motivation, development, and wellness,* The Guilford Press, New York, NY.

Sass D., (2011), Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework, *J. Psychoeduc. Assess.*. **29**(4), 347-363. DOI:10.1177/0734282911406661

Sen S., Yilmaz A. and Temel S., (2016), Adaptation of the Attitude toward the Subject of Chemistry Inventory (ASCI) into Turkish. *J. Educ. Training Stud,* **4**(8), 27-33. ISSN-2324-805X

Stanich C. A., Pelch M. A., Theobald E. J. and Freeman S., (2018), A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women, *Chem. Educ. Res. Pract*., **19**, 846-866. DOI:10.1039/C8RP00044A

Underwood S. M., Reyes-Gastelum D. and Cooper M. M., (2016), When do students recognize relationships between molecular structure and properties? A longitudinal comparison of the impact of traditional and transformed curricula, *Chem. Educ. Res. Pract.*, **17**, 365-380. DOI:10.1039/C5RP00217F

Villafañe S. M. and Lewis J. E., (2016), Exploring a measure of science attitude for different groups of students enrolled in introductory college chemistry, *Chem. Educ. Res. Pract*., **17,** 731−742. DOI:10.1039/C5RP00185D

Villafañe S. M., Xu, X. and Raker J. R., (2016), Self-efficacy and academic performance in first-semester organic chemistry: Testing a model of reciprocal causation, *Chem. Educ. Res. Pract.,* **17**, 973-984. DOI: 10.1039/C6RP00119J

Vishnumolakala V. R., Qureshi S. S., Treagust D. F., Mocerin, M., Southam D. S. and Ojeil J., (2018), Longitudinal impact of process-oriented guided inquiry learning on the attitudes, self-efficacy and experiences of pre-medical chemistry students. *QScience Connect,* **1**, 1-12. DOI:10.5339/connect.2018.1

Vishnumolakala V. R., Southam D. C., Treagust D. F., Mocerino M. and Qureshi S. (2017), Students' attitudes, self-efficacy and experiences in a modified process-oriented guided inquiry learning undergraduate chemistry classroom, *Chem. Educ. Res. Pract*., **18**, 340-352. DOI:10.1039/C6RP00233A

Wang Y., Rocabado G. A., Lewis J. E. and Lewis S. E., (2020), Prompts to Promote Success: Evaluating Utility Value and Growth Mindset Interventions on General Chemistry Students' Affect and Academic Performance, J. Chem. Educ., (In Review).

Wigfield, A., Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemp. Educ. Psychol., 25*(1)*,* 68-81. DOI: 10.1006/ceps.1999.1015

Wilkinson S., Joffe H. and Yardley L., (2004), Qualitative data collection: interviews and focus groups. SAGE Publications.

Williams G. C. and Deci E. L., (1996), Internalization of biopsychosocial values by medical students: A test of self-determination theory, *J. Pers. Soc. Psychol.*, **70**, 767-779. DOI: 10.1037/0022-3514.70.4.767

Willis G. B., (1999), Cognitive Interviewing: A "How To" Guide. *Meeting of the American Statistical Association,* Research Triangle Institute.

Xu X., Southam D. and Lewis J. E., (2012), Attitude Towards the Subject of Chemistry in Australia: An ALIUS and POGIL Collaboration to Promote Cross-National Comparisons, *Aus. J. Educ. Chem.*, **72**, 32−36. ISSN-14459698

Xu X., Alhoosani K., Southam D. and Lewis J. E., (2015), *Gathering Psychometric Evidence for ASCIv2 to Support Cross-Cultural Attitudinal Studies for College Chemistry Programs*. In Affective Dimensions in Chemistry, Springer-Verlag: Berlin, pp. 177−194.

Xu X. and Lewis J., (2011), Refinement of a Chemistry Attitude Measure for College Students, *J. Chem. Educ.,* **88**, 561-568. DOI:10.1021/ed900071q

Xu X., Villafañe, S. M. and Lewis J. E., (2013), College students' attitudes toward chemistry, conceptual knowledge and achievement: Structural equation model analysis. *Chem. Educ. Res. Pract*., **14**, 188−200. DOI:10.1039/C3RP20170H

CHAPTER 7:

CONCLUSION

Women of Color deserve the spotlight in our research and in our classrooms. The scientific community has ignored Women of Color for too long, attempting to address diversity, inclusion, and equity issues only through gender *or* race (Ong et al., 2011). But Women of Color experience a compounded marginalization due to intersectional, disenfranchised identities (Crenshaw, 1989; Litzler, Samuelson and Lorah, 2014; Ireland et al., 2018). Consequently, unwelcoming STEM spaces can become difficult to navigate, and many students leave (Seymour and Hunter, 2019) without fulfilling President Obama's (2010) mandate to diversify STEM. Investigating the perceptions and experiences that Women of Color, and other subgroups with marginalized intersectional identity backgrounds have in our classrooms, can help the field of CER understand how to improve the curriculum and design pedagogies and interventions that are more inclusive to these diverse groups. This simple commitment, when done conscientiously, safeguarding against the propagation of inequities (García *et al.*, 2018; Gillborn *et al.,* 2018), can help improve student attitudes, which in turn may inspire more Women of Color to succeed and stay in their chosen STEM career paths. The following sections contain a summary of the results of the work I have done to address the gap in knowledge about students' attitudes in organic chemistry classrooms, paying particular attention to Women of Color. Following the summary of results, I present implications for practitioners, researchers, and policy makers drawn from this work.

Summary of Results

In chapter 3 (Rocabado *et al.,* 2019) we investigated whether the positive gains in attitude measured by the ASCIv3 and achievement (test scores) observed in an organic chemistry flipped course (Mooring *et al*., 2016) extended to the Black female students in this course. The results indicated that the upward trend of attitude and achievement observed for the class was also perceived for the Black female students in the course. However, it was also noted that both attitude and achievement scores began and ended lower for the Black female students. Thus, it was concluded that the flipped classroom pedagogy was a positive experience for all students, even though this pedagogy didn't close the gap that existed from the beginning between Black female students and all their peers. Furthermore, it was demonstrated that attitude and achievement displayed a reciprocal-causation relationship, particularly with *emotional satisfaction* influencing subsequent test scores with a small but significant effect. Thus it was demonstrated that attitude throughout the semester was a significant predictor of the final test score even when taking into account the first exam score.

In chapter 4 (Rocabado *et al.*, 2020), we provided a primer on measurement invariance testing due to its underuse in the field of CER. We recognized that as the field moves toward greater diversity and inclusion initiatives, group comparisons in research and teaching would increase, and our goal was to provide an overview of a method that can be utilized to give researchers check points in which to reflect the feasibility of subgroup comparisons. We provided a step-by-step tutorial as well as software code for the interested readers to follow. We culminated this manuscript with a summary table for easy access with which researchers and practitioners, as

well as journal editors and reviewers can guide their work when reviewing or conducting subgroup comparisons.

In chapter 5, we found similar evidence than in chapter 3, that Women of Color, in this case Hispanic female students in an organic chemistry course, display less positive attitudes than others (White female students in this case). One of the notable differences in this study was that there was no intervention in this course and attitude dropped over the semester. A meta-analysis of attitude change over one semester aided the investigation, concluding that both *intellectual accessibility* (IA) and *emotional satisfaction* (ES) are malleable factors. However, certain pedagogical interventions, such as a flipped classroom or a POGIL classroom, might make more headway in positively impacting IA and ES, although ES seems to be more resistant to change within the time frame of one course. Additionally, at a pivotal point in our data analysis, we became aware of our deficit mindset when we were about to conclude that Hispanic female students have less positive attitudes than their White female peers and this result might lead to lower retention rates and other negative outcomes. However, our awareness of deficit mindset led us to investigate further and ask additional questions to find that, in terms of retention to the next course, given success in first semester organic chemistry we found no evidence of difference between these two groups. This result was encouraging because it provided evidence of the use of a persistence asset, since our expectations through our deficit mindset lens was that of lower retention for Hispanic female students given their less positive attitude.

Finally, in chapter 6 (previously unpublished) we demonstrated the methods used for instrument development in two languages (English and Spanish) and two countries (U.S. and

Chile), such as cognitive interviews and expert panel review (Arjoon *et al*., 2013; AERA *et al*., 2014). We adapted the ASCIv2 to contain a refined *Emotional Satisfaction* scale, and a new *Utility* scale. We demonstrated that the internal structure holds in both contexts through CFA and measurement invariance testing that held to metric invariance. This result indicated that the meaning of the factors is similar across groups (Gregorich, 2006; Rocabado *et al.,* 2020), which is a promising result given the different contexts of these groups. Furthermore, scalar invariance was not reached, potentially due to a 'ceiling effect' for some of the items found in the data for the students in Chile and also due to the unusual data collection approaches necessary because of the Covid-19 pandemic. With this study we concluded that it is possible to adapt and/or create an instrument based on cognitive interviews in two countries and two languages. We gathered various aspects of validity evidence along the way culminating in SEM analyses of attitude-achievement relationship for two subgroups of students, namely high- and low-achieving students. The reciprocal causation model indicated that ES has a small but significant positive relationship with the subsequent exam scores; however, the U-exam relationship was non-significant. This result suggests that students' sense of utility of the discipline is similar at the beginning and at the end of the semester in OCII and not significantly influenced by academic performance throughout the course. However, a closer analysis between high- and low-achieving groups of students, a drop in U was observed for low-achievers and a positive gain observed for the high-achievers.

Implications for Researchers

As indicated by Fazio (1986), attitudes are formed when a person is exposed to stimuli that can spur evaluations of an attitude object. The continual exposure to these stimuli and attitude object can lead to stable attitudes toward the object, yet these attitudes can change over time when influenced in appropriate ways (Ajzen and Sexton, 1999; Reid, 2006). In chapters 3, 5 and 6 I explored changes in attitude over a semester. Overall, it was concluded that certain pedagogies were successful in producing positive changes in attitude (*i.e.*, Mooring *et al.*, 2016) including for the Black female students in the class (Rocabado *et al*., 2019). However, through a meta-analysis of longitudinal studies that utilized the ASCIv2, it was found that IA may be positively impacted by certain types of interventions (i.e., POGIL) during a semester, yet ES is less disposed to change (Chapter 5). These results imply that the investigation of attitude is complex and it takes time. Researchers may wish to investigate attitude longitudinally during a semester; however, they may find that longer investigations may provide greater insight into the changes that are possible when students go through our classrooms. Longer investigations are particularly relevant for courses such as general or organic chemistry that are typically taught over two semesters and provide a longer time of exposure to the attitude object.

Another relevant implication that emerges from this work is the importance of disaggregating data for subgroup investigations that may be of interest in our research. More notably, this work emphasizes the importance of investigating intersectional identities when possible because of the compounding effect of certain intersectional marginalized identities such as gender and race (*i.e.*, Women of Color) that prove to have a negative effect on attitude,

achievement, and retention in science (Catsambis, 1995). To my knowledge, no other work in CER has undertaken the study of attitude toward chemistry for Women of Color in particular, thus the critical need to continue this work across the U.S. and elsewhere. As researchers learn more about women of different racial and ethnic backgrounds and their particular trajectories are brought to light, it may be concluded that they experience our classrooms differently and the curriculum or interventions implemented may not be appropriate nor conducive to their retention or success. However, the investigation of counterstories and asset use (Yosso, 2005; Ong, Smith and Ko, 2018; Gallard Martínez *et al*., 2019) for these groups may elucidate alternate ways to support these students' needs and strengths in order to increase their participation and retention in science courses and majors. I encourage researchers to use this approach when appropriate to purposefully address President Obama's 2010 mandate to diversify STEM.

When utilizing quantitative methods, it is imperative to safeguard against possible threats to the validity of the inferences that are drawn because numbers are trusted blindly. Approaching research methods with a mindset that 'numbers are not neutral' and that researchers nor research are objective or without bias (García *et al*., 2018; Gillborn *et al*., 2018) is a good starting point. Furthermore, checking for evidence and challenging deficit mindset in our research practice moves toward a more inclusive and equitable approach to our studies (Gorski, 2011). In order to ensure our best efforts in our research to serve diversity, inclusion, and equity initiatives, researchers should carefully choose appropriate methods of investigation. For instance, in this work I have promoted the use of measurement invariance testing (chapters 3-6) when conducting group comparisons. When sample size is permissible, this method provides evidence of the feasibility of group comparisons and multiple check points to reflect on potential threats to the validity of the

inferences we could make with our data (Rocabado *et al.,* 2020). I encourage researchers to make use of the software code and step-by-step tutorial we have provided in Rocabado *et al.*, (2020), or other appropriate methods that provide ample opportunities to check for threats that may result from trusting numbers blindly without examining potential biases. Additionally, as demonstrated throughout this work, when adhering to the tenets of QuantCrit and purposefully evaluating the quantitative data that is available, we can center Women of Color in their environments. Taking steps to consider context can demonstrate a commitment to achieve greater inclusion and equity in our field.

Similarly, researchers should not only make sure that their methods are appropriate for their interests and data available, but also that the instruments they use are appropriate for the different groups they are investigating. As demonstrated in chapters 3-6 we utilized measurement invariance testing to investigate the feasibility of comparisons through the investigation of the stability of the internal structure across the groups we compared. Furthermore, in chapter 6 we exhibited the various ways in which we adapted the ASCIv2 and created the ASCI-UE in two languages and used it in U.S. and in Chile. We followed the recommendations delineated in the *Standards for Education and Psychological Testing* (AERA *et al.,* 2014). Researchers should scrutinize the instruments they use and find several aspects of validity evidence gathered when instruments were developed and used. Researchers should also continue to gather validity evidence with the instruments they choose to use and conduct rigorous tests that provide as much evidence as possible that their inferences are appropriate for their data collected.

Implications for Practitioners

College chemistry classrooms may be the only opportunities for many students to experience the field of chemistry. Hence, our classes are where students form their attitude toward chemistry that will shape their future attitudes and behaviors toward the field of chemistry throughout their lives. It becomes critical that within our classrooms these students are exposed to practices that encourage positive attitudes and foster diversity, inclusion, and equity. A few recommendations emerge from the work presented herein.

First, I have shown in chapters 3 and 5 that implementing certain pedagogies, such as a flipped classroom, may positively impact attitude for all students (Mooring *et al*., 2016). Practitioners should investigate and implement pedagogies that have shown to promote positive attitudes and evaluate the success of these pedagogies both from empirical studies and in their own classrooms. Particularly in chapter 5, a meta-analysis of longitudinal studies that utilized the ASCIv2 showed that active learning techniques such as POGIL and flipped classrooms promote greater positive attitude gains than other interventions. In addition, practitioners should investigate the impact these pedagogical interventions have with diverse subgroups of students. For instance, in Rocabado *et al*., (2019) we investigated whether positive gains in attitude for the entire class extended to the Black female students. I encourage practitioners to examine the impact of their interventions on subgroups of students, particularly Women of Color, who experience the 'double bind' (Ong et al., 2011) and whose experiences may be different from others in the class (chapter 5).

Second, I encourage practitioners to utilize instruments, such as the ASCIv2 (Xu and Lewis, 2011) or the ASCI-UE (chapter 6) developed with rigorous methods and tested with diverse populations. Moreover, I encourage practitioners to conduct rigorous analyses with these instruments to continue to gather validity evidence as demonstrated throughout this work and safeguard against possible threats to the validity of the inferences made with their results. If analyses such as CFA or measurement invariance testing are not possible due to the demand of large sample sizes, practitioners should search for alternate ways to gather validity evidence, such as correlational analyses to confirm internal structure and other forms of validity evidence (see examples in Rocabado et al., 2020). Alternatively, practitioners may choose to utilize qualitative methods such as cognitive interviews (Willis, 1999).

Third, I encourage practitioners to evaluate their classroom pedagogies and embed in their practice ways to provide emotional support for their students. In chapters 3 and 6 I have demonstrated the reciprocal relationship between attitude and achievement in a semester of Organic Chemistry. It was clear that one of the significant relationships was that of *Emotional Satisfaction* and the subsequent exam scores. Yet, most of our pedagogies and classroom interventions are solely designed to support students' cognition. Given these results, practitioners should labor to find ways to support both attitude domains in their classrooms to influence a better outcome, particularly for students of marginalized populations or for students at risk.

Finally, the study of attitude in chemistry courses is of relevance because this construct has been shown to positively relate to metrics of achievement (Brandriet, Ward, and Bretz, 2013; Xu, Villafañe, and Lewis, 2013; Villafañe and Lewis, 2016; Rocabado *et al*., 2019) and retention

(Halpern *et al.,* 2007). However, often it is automatically determined that students who have less positive attitude will also perform worse on exams and in the course, and will not advance to the next course. Although there is empirical evidence of the positive relationship between attitude and achievement included in this dissertation (Rocabado *et al*., 2019; Chapter 6), the mindset that this relationship is somehow 'fate,' is evidence of a deficit mindset, which indicates that students who "lack" certain traits (*i.e*., more positive attitudes) don't have what it takes to succeed (Gorski, 2011). This mindset also sentences students to be the ones who leave STEM courses and programs. However, throughout this work, particularly in chapter 5, I recount my experience challenging a deficit mindset as I looked for ways in which Hispanic female students displayed the use of the asset of persistence (Rodriguez, Cunningham and Jordan, 2019). Practitioners should also challenge a deficit mindset in their practice and in their investigations, and engage in active promotion of asset use among their students, particularly URGs. I believe this practice can encourage marginalized groups of students (*i.e*., Women of Color) to be empowered to use their strengths in the production of their own solutions to their challenges (Myende, 2015).

Implications for Policy

One of the reasons I chose to investigate attitude toward chemistry is because of its significance in influencing not only achievement and retention (see Halpern *et al*., 2007; Rocabado *et al.,* 2019), but also future behavioral intentions (Ajzen and Fishbein, 2000). Thus, this construct is a relevant topic of investigation in CER, and in all STEM education fields. Although investigating attitudes can be done both with qualitative and quantitative methods, often the field

of CER has utilized quantitative methods (*i.e.,* Bauer, 2008; Xu, Villafañe and Lewis, 2013, etc.) including the work in this dissertation (chapters 3-6). However, throughout this work I have also demonstrated that quantitative studies must be conducted in a responsible manner. An important implication relevant to policy makers that is emphasized throughout this dissertation is that 'numbers are not neutral' (García *et al*., 2018; Gillborn *et al*., 2018). This statement indicates that the 'numbers' that are utilized to make decisions in educational settings which affect students, teachers, and researchers have to be carefully scrutinized to safeguard against propagating systemic biases and social injustices. A large focus of the educational system is 'closing the gaps,' which is by definition a group comparison. In this work we have demonstrated one method (measurement invariance testing) that can be employed to check whether group comparisons are feasible, providing several check points to scrutinize our approach (Rocabado *et al*., 2020). Moreover, often the differences between groups and the need to 'close the gaps' are confused with group deficiencies (Gorski, 2011). Therefore, challenging a deficit mindset and carefully scrutinizing the way in which we use the quantitative results we obtain from our investigations in research, practice, and in policy making is vital to all diversity, inclusion, and equity initiatives in CER.

References

AERA, APA and NCME, (2014), *Standards for educational and psychological testing*, Washington, DC: American Psychological Association.

Ajzen L. and Fishbein M., (2000), *Attitudes and the attitude-behavior relation: Reasoned and automatic processes*. In European review of social psychology (Vol. 11) Stroebe W. and Hewstone M. (Eds.), Chichester, England: Wiley, pp. 1-33.

Ajzen L., and Sexton J., (1999), *Depth of processing, belief congruence, and attitude-behavior correspondence*. In Dual-process theories in social psychology, Chaiken S. and Trope Y. (Eds.), New York: Guilford, pp. 117-140.

Arjoon J. A., Xu X. and Lewis J. E., (2013), Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence, *J. Chem. Educ.*, **90**, 536–545. DOI:10.1021/ed3002013

Bauer C. F., (2008), Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts, *J. Chem. Educ.*, **85**, 1440−1445. DOI:10.1021/ed085p1440

Brandriet A. R., Ward R. M. and Bretz S. L., (2013), Modeling meaningful learning in chemistry using structural equation modeling, *Chem. Educ. Res. Pract.*, **14**, 421-430. DOI:10.1039/C3RP00043E

Catsambis S., (1995), Gender, Race, Ethnicity, and Science Education in the Middle Grades, *J. Res. Sci. Teach.*, **32**(2), 243−257. DOI:10.1002/tea.3660320305

Crenshaw K., (1989), Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, *University of Chicago Legal Forum*, 139−168. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Fazio R. H., (1986), *How do attitudes guide behavior?* In The handbook of motivation and cognition: Foundations of social behavior, Sorrentino R. M. and Higgins E. T. (Eds.), New York: Guilford.

Gallard Martinez A. J., Pitts W., Ramos de Robles S. L., Milton Brkich K. L., Flores Bustos B. and Claeys L., (2019), Discerning contextual complexities in STEM career pathways: insights from successful Latinas, *Cult. Stud. Sci. Educ.,* **14**, 1079-1103. DOI:10.1007/s11422-018-9900-2

García N. M., López N. and Vélez V. N., (2018), QuantCrit: Rectifying quantitative methods through critical race theory, *Race Ethn. Educ.,* **21**(2), 149-157. DOI:10.1080/13613324.2017.1377675

Gillborn D., Warmington P. and Demack S., (2018), QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics, *Race Ethn. Educ.,* **21**(2), 158-179. DOI:10.1080/13613324.2017.1377417

Gorski P. C., (2011), Unlearning deficit ideology and the scornful gaze: Thoughts on authenticating the class discourse in education, *Counterpoints,* **402**, 152-173. https://www-jstor-org.ezproxy.lib.usf.edu/stable/42981081

Gregorich S. E., (2006), Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework, *Med Care,* **44** (11 Suppl 3), S78-S94. DOI:10.1097/01.mlr.0000245454.12228.8f

Halpern D. F., Benbow C. P., Geary D. C., Gur R., Hyde, J. S. and Gernsbacher M. A. (2007), The science of sex differences in science and mathematics, *Psychol. Sci. Pub. Int.,* **8**, 1–51. DOI:10.1111/j.1529-1006.2007.00032.x

Ireland D. T., Freeman K. E., Winston-Proctor C. E., DeLaine K. D., McDonald Lowe S. and Woodson K. M., (2018), (Un)hidden Figures: A Synthesis of Research Examining the Intersectional Experiences of Black Women and Girls in STEM, *Rev. Res. Educ.*, **42**, 226−254. DOI: 10.3102/0091732X18759072

Litzler E., Samuelson C. C. and Lorah J. A., (2014), Breaking it Down: Engineering Student STEM Confidence at the Intersection of Race/ Ethnicity and Gender, *Res. High. Educ*., **55**, 810−832. DOI:10.1007/s11162-014-9333-z

Mooring S. R., Mitchell C. E. and Burrows, N. L., (2016), Evaluation of a flipped, large enrollment organic chemistry course on student attitude and achievement, *J. Chem. Educ.*, **93**, 1972−1883. DOI:10.1021/acs.jchemed.6b00367

Myende P. E., (2015), Tapping in the asset-based approach to improve academic performance in rural schools, *J. Hum. Ecol,* **50**(1), 31-42. DOI:10.1080/09709274.2015.11906857

Obama B., (2010), Science, technology, engineering and math: Education for global leadership, *Retrieved on* July 8, 2020. https://www.ed.gov/sites/default/files/stem-overview.pdf

Ong M., Smith J. M. and Ko L. T., (2018), Counterspaces for Women of Color in STEM higher education: Marginal and central spaces for persistence and success, *J. Res. Sci. Teach.,* **55**(2), 206-245. DOI:10.1002/tea.21417

Ong M., Wright C., Espinosa L. L. and Orfield G., (2011), Inside the Double Bind: A Synthesis of Empirical Research on Undergraduate and Graduate Women of Color in Science, Technology, Engineering, and Mathematics, *Harvard Educ. Rev.*, **81**(2), 172− 208. DOI:10.17763/haer.81.2.t022245n7x4752v2

Reid N., (2006), Thoughts on attitude measurement, *Res. Sci.Technol. Educ.*, **24**(1), 3-27, DOI:10.1080/02635140500485332

Rocabado G. A., Kilpatrick N. A., Mooring S. R., and Lewis J. E., (2019), Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course, *J. Chem. Educ.,* **96**(11), 2371-2382. DOI:10.1021/acs.jchemed.9b00516

Rocabado G. A., Komperda R., Lewis J. E. and Barbera J., (2020), Addressing diversity and social inclusion through groups comparisons: A primer on measurement invariance testing, *Chem. Educ. Res. Pract.,* **21,** 969-988. DOI:10.1039/D0RP00025F

Rodríguez S., Cunningham K. and Jordan A., (2019), STEM identity development for Latinas: The role of self- and outside recognition, *J. Hispan. High. Educ.,* **18**(3), 254-272. DOI:10.1177/1538192717739958

Seymour E. and Hunter A-B., (2019), *Talking About Leaving Revisited: Persistence, relocation, and loss in undergraduate STEM education*, Springer Nature Switzerland.

Villafañe S. M. and Lewis J. E., (2016), Exploring a measure of science attitude for different groups of students enrolled in introductory college chemistry, *Chem. Educ. Res. Pract*., **17,** 731−742. DOI:10.1039/C5RP00185D

Willis G. B., (1999), Cognitive interviewing: A "how to" guide, In *Meeting of the American Statistical Association,* Research Triangle Institute.

Xu X. and Lewis J., (2011), Refinement of a Chemistry Attitude Measure for College Students, *J. Chem. Educ.,* **88**, 561-568. DOI:10.1021/ed900071q

Xu X., Villafañe, S. M. and Lewis J. E., (2013), College students' attitudes toward chemistry, conceptual knowledge and achievement: Structural equation model analysis. *Chem. Educ. Res. Pract*., **14**, 188−200. DOI:10.1039/C3RP20170H

Yosso T., (2005), Whose culture has capital? A critical race theory discussion of community cultural wealth, *Race Ethn. Educ.,* **8**(1), 69-91. DOI:10.1080/1361332052000341006

APPENDIX A

COMMONLY USED ABBREVIATIONS

ASCI	Attitude toward the Subject of Chemistry Inventory
ASCIv2	Attitude toward the Subject of Chemistry Inventory version 2
ASCI-UE	Attitude toward the Subject of Chemistry Inventory – Utility and Emotional Satisfaction
CER	Chemistry Education Research
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit index
ES	Emotional Satisfaction
IA	Intellectual Accessibility
OCI/OCII	Organic Chemistry I/Organic Chemistry II
PC	Perceived Competence
RMSEA	Root Mean Square of Approximation
SRMR	Standardized Root Mean Square Residual
STEM	Science, Technology, Engineering, and Math
U	Utility
URG	Underrepresented Group
URM	Underrepresented Minority

APPENDIX B

PUBLISHER PERMISSIONS DOCUMENTATION

B.1. Chapter 3 (Journal of Chemical Education)

B.2. Chapter 4 (Chemistry Education Research and Practice)

Addressing diversity and inclusion through group comparisons: a primer on measurement invariance testing

G. A. Rocabado, R. Komperda, J. E. Lewis and J. Barbera, *Chem. Educ. Res. Pract.*, 2020, **21**, 969
DOI: 10.1039/D0RP00025F

If you are the author of this article you do not need to formally request permission to reproduce figures, diagrams etc. contained in this article in third party publications or in a thesis or dissertation provided that the correct acknowledgement is given with the reproduced material.

Reproduced material should be attributed as follows:

- For reproduction of material from NJC:
  [Original citation] - Reproduced by permission of The Royal Society of Chemistry (RSC) on behalf of the Centre National de la Recherche Scientifique (CNRS) and the RSC
- For reproduction of material from PCCP:
  [Original citation] - Reproduced by permission of the PCCP Owner Societies
- For reproduction of material from PPS:
  [Original citation] - Reproduced by permission of The Royal Society of Chemistry (RSC) on behalf of the European Society for Photobiology, the European Photochemistry Association, and RSC
- For reproduction of material from all other RSC journals:
  [Original citation] - Reproduced by permission of The Royal Society of Chemistry

If you are the author of this article you still need to obtain permission to reproduce the whole article in a third party publication with the exception of reproduction of the whole article in a thesis or dissertation.

Information about reproducing material from RSC articles with different licenses is available on our Permission Requests page.

APPENDIX C:

SUPPORTING INFORMATION

Chapters 3 and 4 were previously published and electronic supplementary information (ESI) documents were included with each publication. Each of those additional documents are included in this appendix. Additional information for chapters 5 and 6 are also included in this Appendix.

C.1. Chapter 3: Electronic Supplementary Information

Attitude toward the Subject of Chemistry Inventory version 3 and 2 presented in figures S3.1a and S3.1b respectively. Version 2 is the original adaptation by Xu and Lewis in 2011. Version 3 was developed to test whether item order played a role in the factor structure. Version 3 is the one utilized in this study.

ASCIv3

A list of opposing words appears below. Rate how well these words describe your feelings about chemistry. Think carefully and try not to include your feelings toward the chemistry teachers or chemistry courses. For each line, choose a position between the two words that describes exactly how you feel. The middle position is if you are undecided or have no feelings related to the terms on that line.

		Easy					**Hard**
1.	Chemistry is…	1	2	3	4	5	6	7
		Chaotic					**Organized**
2.	Chemistry is…	1	2	3	4	5	6	7
		Confusing					**Clear**
3.	Chemistry is…	1	2	3	4	5	6	7
		Comfortable					**Uncomfortable**
4.	Chemistry is…	1	2	3	4	5	6	7
		Satisfying					**Frustrating**
5.	Chemistry is…	1	2	3	4	5	6	7
		Challenging					**Not challenging**
6.	Chemistry is…	1	2	3	4	5	6	7
		Pleasant					**Unpleasant**
7.	Chemistry is…	1	2	3	4	5	6	7
		Complicated					**Simple**
8.	Chemistry is…	1	2	3	4	5	6	7

Figure S3.1a: Attitude toward the Subject of Chemistry Inventory version 3 (ASCIv3). This is the instrument used for the present study. Items 2 and 8 switch places from the original version of the instrument (ASCIv2 shown in Figure S1b).

ASCIv2

A list of opposing words appears below. Rate how well these words describe your feelings about chemistry. Think carefully and try not to include your feelings toward the chemistry teachers or chemistry courses. For each line, choose a position between the two words that describes exactly how you feel. The middle position is if you are undecided or have no feelings related to the terms on that line.

		Easy					**Hard**
1.	Chemistry is…	1	2	3	4	5	6	7
		Complicated					**Simple**
2.	Chemistry is…	1	2	3	4	5	6	7
		Confusing					**Clear**
3.	Chemistry is…	1	2	3	4	5	6	7
		Comfortable					**Uncomfortable**
4.	Chemistry is…	1	2	3	4	5	6	7
		Satisfying					**Frustrating**
5.	Chemistry is…	1	2	3	4	5	6	7
		Challenging					**Not challenging**
6.	Chemistry is…	1	2	3	4	5	6	7
		Pleasant					**Unpleasant**
7.	Chemistry is…	1	2	3	4	5	6	7
		Chaotic					**Organized**
8.	Chemistry is…	1	2	3	4	5	6	7

Figure S3.1b: Attitude toward the Subject of Chemistry Inventory version 2 (ASCIv2). This is the original instrument developed by Xu & Lewis in 2011 as an adaptation of the original created by Bauer in 2008.

Demographic and Missing Data Analysis

Table S3.1 displays the demographic breakdown and proportions of missing data in each category for this study. The first column of Table S3.1 shows that there were only nine missing

item responses on the pre-test. The second column indicates that all students who took the post-test answered every item. The last two columns in Table S3.1 document that, out of 395 students, 46 students (12%) responded to the post-test but not to the pre-test, and 51 students (13%) responded to the pre-test but not to the post-test. Among these students, Black males had the lowest response rate (> 25% missing data on all values for the post-test). Any desired future comparisons to this specific group should be made with caution, due to the relatively high proportion of missing data that might suggest this group is not being well represented in this sample. This pattern of missing data influenced the decision to compare Black female students to all other students rather than attempt to compare all demographic subgroups in the sample.

Table S3.1: Demographic Table with Missing Data Analysis for Organic Chemistry I

			Pre	Post	Pre	Post
Race	**Gender**	**Sample size**	**# of missing item responses[a]**	**# of missing item responses[b]**	**# of cases with missing data on all values[c]**	**# of cases with missing data on all values[d]**
Black	Female	125	1 (0.1%)	0 (0.0%)	13 (10.0%)	12 (9.6%)
	Male	39	6 (1.9%)	0 (0.0%)	7 (18.0%)	10 (26.0%)
White	Female	48	1 (0.3%)	0 (0.0%)	6 (13.0%)	6 (13.0%)
	Male	32	0 (0.0%)	0 (0.0%)	2 (6.3%)	6 (19.0%)
Asian	Female	64	0 (0.0%)	0 (0.0%)	9 (14.0%)	5 (7.8%)
	Male	43	0 (0.0%)	0 (0.0%)	6 (14.0%)	8 (19.0%)
Other	Female	27	1 (0.5%)	0 (0.0%)	1 (3.7%)	2 (7.4%)
	Male	17	0 (0.0%)	0 (0.0%)	2 (12.0%)	2 (12.0%)
Total		395	9 (0.3%)	0 (0.0%)	46 (11.6%)	51 (12.9%)

[a]These values indicate the number of students within the demographic category who responded to some of the items but not others in the pre-test. [b]There were no missing item responses in the post-test. [c,d]These values indicate the number students within the demographic category did not respond to any of the items.

Table S3.2 displays missing data for the pre-test by item. All 395 students who responded to the pre-test answered the first item. The highest frequency of missing data was for the final item. Still, only four of the 395 students (1%) did not respond to that item.

Table S3.2: Proportion of Missing Values for Pre-test

Organic Chemistry I *n* = 395		
	# missing values	**Proportion[a]**
Item 1	0	0.000
Item 2	1	0.003
Item 3	2	0.005
Item 4	2	0.005
Item 5	2	0.005
Item 6	2	0.005
Item 7	2	0.005
Item 8	4	0.010
Total	15	0.005

[a]Proportion of missing values calculated by dividing number of missing values by *n* (395).

MANOVA

The study done by Mooring and colleagues in 2016 utilizes MANOVA to investigate attitude gains in a flipped classroom compared to a traditional classroom over the course of the first semester of organic chemistry, finding that the flipped classroom was associated with gains in both *intellectual accessibility* (IA) and *emotional satisfaction* (ES) for the overall sample of 297 students with complete pre- and post-test data. Whether the gains extended to the Black female students within the sample was not investigated in that study. Table S3.3 contains the disaggregated raw gain scores (post-pre) for Black female and all other students. For both IA and ES, Black female students have a slight decrease in score from pre- to post-test in the traditional classroom, but an increase in the flipped classroom. This pattern is similar to that for all other students.

Table S3.3: Descriptive Statistics for ES- and IA-Gains, Black Female and All Other Students in Organic Chemistry I Flipped and Traditional Classrooms

				Pre-test		Post-test			
			N	Mean	S.D.	Mean	S.D	Gain Mean	Gain S.D
IA	All Other	Traditional	103	2.90	1.015	3.02	1.122	0.12	1.258
		Flipped	94	2.97	1.073	3.52	1.071	0.55	1.186
	Black Female	Traditional	57	2.73	1.107	2.55	1.211	-0.18	1.083
		Flipped	43	2.71	0.945	3.34	1.287	0.63	1.276
ES	All Other	Traditional	103	4.00	1.085	3.99	1.321	-0.01	1.412
		Flipped	94	4.09	1.132	4.39	1.237	0.30	1.431
	Black Female	Traditional	57	3.67	0.993	3.51	1.290	-0.16	1.270
		Flipped	43	3.71	1.114	4.25	1.224	0.54	1.231

N = Sample size. S.D. = Standard Deviation.

Tables S3.4 and S3.5 display the results of MANOVA tests documenting that there is no evidence of a significant difference in IA and ES gain scores for Black female students in the flipped and traditional classroom as compared to those for all other students. These positive results for Black female students highlighted the importance of conducting measurement invariance testing to be certain the results would hold.

Table S3.4: MANOVA (Attitude towards the Subject of Chemistry Inventory version 3 Gain Scores, Black Female students and all other students in Organic Chemistry Flipped Classroom)

Test of Between-Subjects Effects

Source	Dependent Variable	Type III Sums of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Square
Corrected Model	IA-gain	0.193	1	0.193	0.131	0.718	.001
	ES-gain	1.702	1	1.702	0.904	0.343	.007
Intercept	IA-gain	40.729	1	40.729	27.613	0.000	.176
	ES-gain	20.879	1	20.879	11.089	0.001	.076
Black female All other	IA-gain	0.193	1	0.193	0.131	0.718	.001
	ES-gain	1.702	1	1.702	0.904	0.343	.007
Error	IA-gain	199.125	135	1.475			
	ES-gain	254.189	135	1.883			
Total	IA-gain	244.206	137				
	ES-gain	275.250	137				
Corrected Total	IA-gain	199.318	136				
	ES-gain	255.891	136				

Table S3.5: MANOVA (Attitude towards the Subject of Chemistry Inventory version 3 Gain Scores, Black Female students and all other students in Organic Chemistry Traditional Classroom)

Test of Between-Subjects Effects

Source	Dependent Variable	Type III Sums of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Square
Corrected Model	IA-gain	3.190	1	3.190	2.220	0.138	.014
	ES-gain	0.708	1	0.708	0.381	0.538	.002
Intercept	IA-gain	0.082	1	0.082	0.057	0.812	.000
	ES-gain	1.037	1	1.037	0.558	0.456	.004
Black female – All other	IA-gain	3.190	1	3.190	2.220	0.138	.014
	ES-gain	0.708	1	0.708	0.381	0.538	.002
Error	IA-gain	227.004	158	1.437			
	ES-gain	293.697	158	1.859			
Total	IA-gain	230.250	160				
	ES-gain	295.063	160				
Corrected Total	IA-gain	230.194	159				
	ES-gain	294.406	159				

Confirmatory Factor Analysis

Confirmatory factor analysis was performed to determine whether the factor structure commonly used for ASCIv2 would hold for ASCIv3. Correlated errors were necessary for good model fit. The first correlated error term added to the model was for Items 2 (Chaotic-Organized) and 3 (Confusing-Clear). Although these items belong to different factors, Item 3 has been shown to demonstrate conflicting loading patterns when the order of items switches from ASCIv2 to ASCIv3 (Xu, 2010). This order alteration appears to influence Item 3 to elicit a response that is related to the *emotional satisfaction* construct. In ASCIv3, Item 2 is modeled under the *emotional satisfaction* factor, and it seems to affect Item 3 to either cross-load on both factors or highly correlate residual variances with items in the *emotional satisfaction* factor (Xu, 2010). The second correlated error term was for Items 6 (Challenging-Not challenging) and 8 (Complicated-Simple). Item 6 has been shown to perform idiosyncratically with diverse populations and has shown

conflicting loading patterns (Xu, *et al*., 2015; Montes, *et al*., 2018). Each of these correlated error terms was added one at a time for both groups, and each showed improved model fit. The model with both terms reaches the level of acceptable model fit. The same process was adopted for OC1 post-test data, resulting in correlated errors added between the same item pairs as for the pre-test. Figure S3.2 shows the final models (with both correlated error terms) for pre- and post- data for Black female students and all other students.

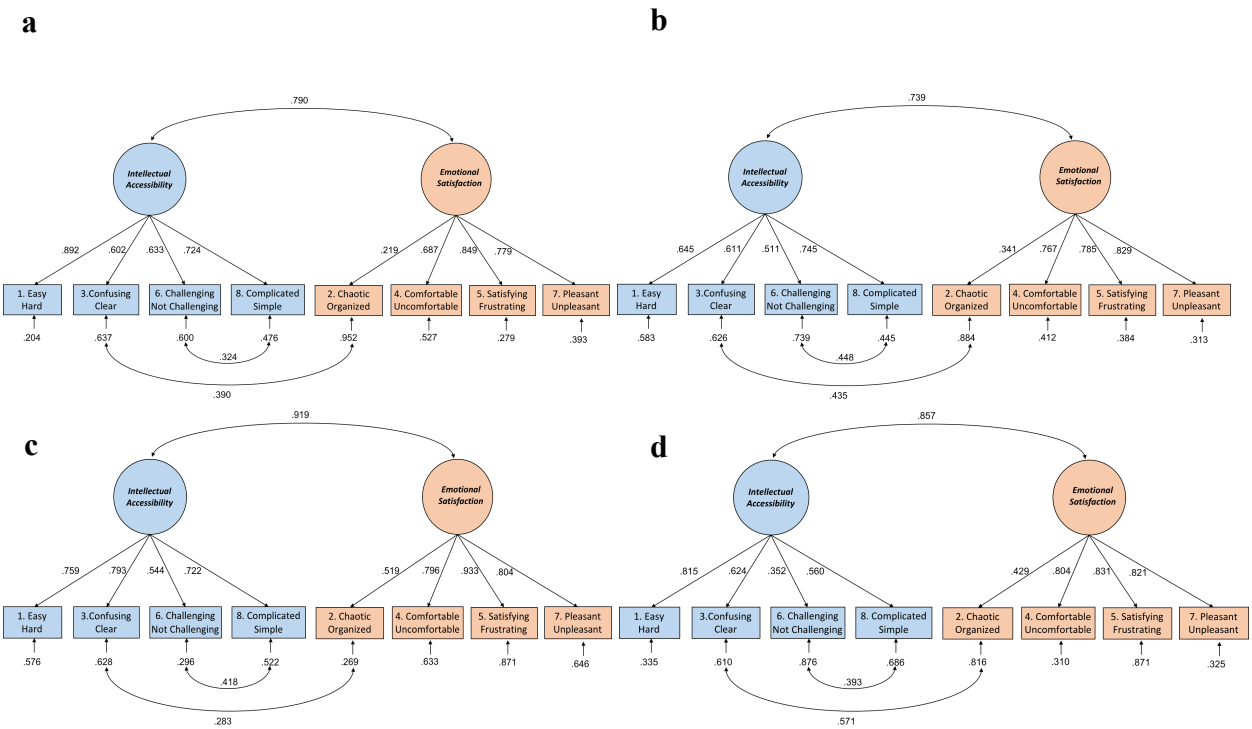


Figure S3.2: a) ASCIv3 CFA for pre-test for Black female students in Organic Chemistry I. b) ASCIv3 CFA for pre-test for all other students in Organic Chemistry I. c) ASCIv3 CFA for post-test for Black female students in Organic Chemistry I. d) ASCIv3 CFA for post-test for all other students in Organic Chemistry I. All loadings, variances and covariances are significant to the .05 level.

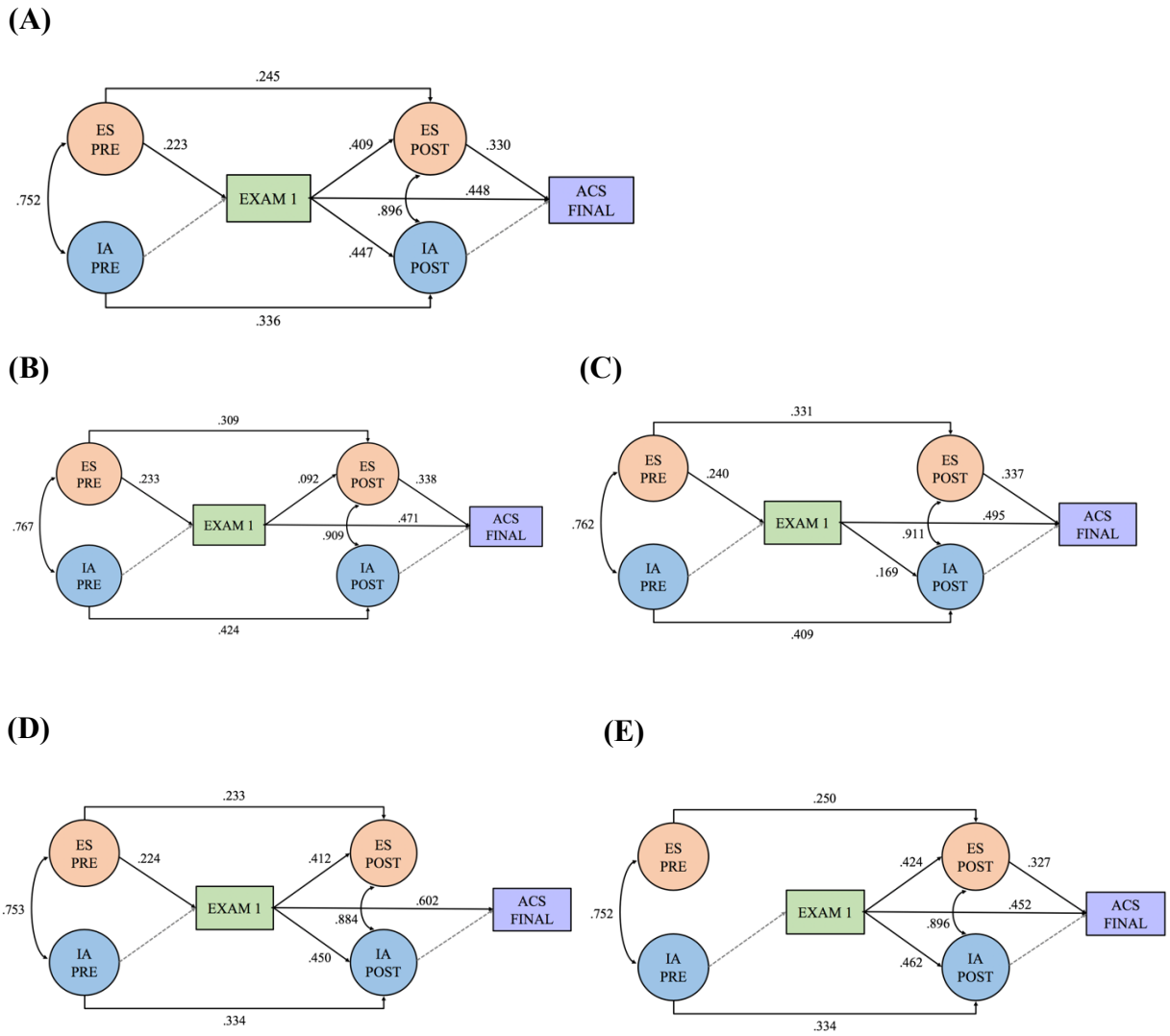


Figure S3.3: Alternative models evaluated for students in the Flipped Classroom in OC1. (A) Theorized model. (B) Model without Exam 1 and IA Post relationship. (C) Model without Exam 1 and ES Post relationship. (D) Model without ES Post and ACS Final relationship. (E) Model without ES Pre and Exam 1 relationship.

Relationship to Other Variables

The main body of the manuscript discusses a model of the relationship between intellectual accessibility, emotional satisfaction, and exam performance. This model was derived from a model

building strategy in which multiple theoretically reasonable models were tested. Figure S3.3 displays Models A-E, which were the theorized models based on other reported models for this instrument (Xu, et al., 2013). Model A is the full reciprocal causation model theorized for this data (Gibbons, et al., 2018; Pekrun, 2006). In this model, all paths are significant at the .05 level, except for the paths from IA Pre and IA Post to Exam 1 and ACS Final, respectively. This model indicated acceptable model fit (Hu & Bentler, 1999). Models B-E are nested models that removed one path from the full model (A) and tested model fit. All models showed normal estimation and convergence; however, model A displays the best statistical and theoretical fit.

Table 3.6: Model Fit for Structural Equation Models A-E

	χ^2	*df*	*p*	CFI	SRMR	RMSEA
Model A[a]	191.758	123	0.0001	.927	.069	.064
Model B[b]	215.652	124	< .0001	.903	.093	.073
Model C[b]	211.569	124	< .0001	.907	.089	.072
Model D[b]	206.280	124	< .0001	.913	.080	.070
Model E[b]	197.253	124	< .0001	.922	.087	.066

[a]Model A shows best fit. [b]Models B through E show normal estimation; however model fit is worse than Model A.

Reliability

The main body of the manuscript argues that omega is the appropriate internal consistency reliability coefficient for non-tau-equivalent scales, and provides omega values for the relevant data collections demonstrating acceptable internal consistency. Because some readers may be more familiar with Cronbach's alpha, we provide the corresponding Cronbach's alpha values in Table S3.7. The conclusion regarding acceptable internal consistency is not substantively different

with the alternative values. Additionally, the model we test in this study is a multidimensional (2 factors) model with correlated factors. The omega coefficient we report assumes a unidimensional model (1 factor), or at least multiple factors that are not correlated. Thus, we compare our reliability results utilizing yet another reliability measure for multidimensional models with correlated factors suggested by Cho (2016).

Table 3.: Reliability Coefficients Calculated for Pre- and Post-tests of ASCIv3 for Black Female and All Other Students

		Cronbach's Alpha		Correlated Factors	
		IA	**ES**	**IA**	**ES**
Pre	**Black Female**	.812	.728	.809	.749
	All Other	.759	.778	.725	.790
Post	**Black Female**	.830	.848	.797	.794
	All Other	.732	.811	.688	.782

IA=*Intellectual Accessibility*. ES=*Emotional Satisfaction*

Measurement Invariance Testing

Measurement invariance testing is performed to ensure that the internal structure of an instrument is consistent for different groups. If configural, metric, and scalar models hold for both groups with acceptable model fit indices, then comparisons of scores can be made between those groups with the assurance that differences observed are not likely to be an artifact of the instrument. Tables S3.8 and S3.9 display evidence that the two-factor structure of this instrument is consistent at the beginning and end of the semester for students in the traditional and flipped classrooms in Organic Chemistry I.

Table S3.8: Measurement Invariance Testing for Traditional and Flipped Classrooms at the Beginning of the Semester

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	65.451	34	.0009	.955	.055	-	-	-	-	-
Metric vs. Configural	69.117	40	.0029	.959	.064	3.666	6	.7218	.005	.009
Scalar vs. Metric	80.518	46	.0012	.951	.067	11.401	6	.0767	.008	.003

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Traditional classroom (n = 201) and Flipped classroom (n = 194). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. And the scalar model adds the constraint of equal intercepts for both groups. Each constraint is added one at a time. *df* = degrees of freedom.

Table S3.9: Measurement Invariance Testing for Traditional and Flipped Classrooms at the End of the Semester

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	86.770	34	< .0001	.945	.050	-	-	-	-	-
Metric vs. Configural	91.248	40	< .0001	.947	.057	4.478	6	.6123	.002	.007
Scalar vs. Metric	93.564	46	< .0001	.951	.057	2.316	6	.8885	.004	.000

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Traditional classroom (n = 201) and Flipped classroom (n = 194). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. And the scalar model adds the constraint of equal intercepts for both groups. Each constraint is added one at a time. *df* = degrees of freedom.

After obtaining evidence through measurement invariance testing that factor scores can be compared between different groups, the scalar model is used in the longitudinal factor score comparison for Black female students and all other students regardless of whether they are in the traditional or flipped classroom for the ASCIv3. We observe that there is a significant difference only in IA for all other students, who display higher scores at the end of the semester.

Table S10: Longitudinal Latent Factor Score Comparison for Black Female Students and All Other Students

	Black Female			All Other		
	Pre[a]	Post[b]	*p*	Pre[a]	Post[b]	*p*
Intellectual Accessibility	0.000	0.219	0.104	0.000	0.391	< .0001
Emotional Satisfaction	0.000	0.099	0.128	0.000	0.153	0.140

[a]Reference group with latent mean score of zero. [b]Latent factor score calculated as a deviation from the reference group.

Table S3.11 and Table S3.12 display the longitudinal measurement invariance testing model fit indices that pertain to Black female students (Table S3.12) and all other students (Table S3.11). These results show no evidence of significant difference between the models, suggesting that longitudinal comparisons within the groups are supported.

Table S3.11: Longitudinal Measurement Invariance Testing for All Other Students

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	175.253	94	< .0001	.933	.064	-	-	-	-	-
Metric vs. Configural	183.094	100	< .0001	.931	.068	7.841	6	.250	.002	.004
Scalar vs. Metric	192.842	106	< .0001	.928	.068	9.748	6	.136	.003	.000

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Pre-test and Post-test for All Other students (n = 270). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. And the scalar model adds the constraint of equal intercepts for both groups. Each constraint is added one at a time. *df* = degrees of freedom.

Table S3.12: Longitudinal Measurement Invariance Testing for Black Female Students

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	170.347	94	< .0001	.905	.068	-	-	-	-	-
Metric vs. Configural	180.587	100	< .0001	.900	.085	10.240	6	.115	.005	.017
Scalar vs. Metric	191.783	106	< .0001	.893	.084	11.196	6	.083	.007	.001

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Pre-test and Post-test for Black female students (n = 125). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. And the scalar model adds the constraint of equal intercepts for both groups. Each constraint is added one at a time. *df* = degrees of freedom.

C.2. Chapter 4: Electronic Supplementary Information

The purpose of the electronic supplementary information (ESI) is to provide readers with the data and code necessary to reproduce the examples from the main body of the paper as well as to provide a template for conducting invariance testing on a simulated data set that can be modified for those interested in conducting invariance testing on their own data. The code in the ESI is primarily written for the R statistical computing language, though Mplus code is also included for conducting invariance testing. The code in the ESI is also available through GitHub (https://github.com/RegisBK/Invariance_CERP) as this provides an easier way to download and use the code rather than cutting and pasting from this document. All analyses were conducted with R version 3.6.1 (R Core Team, 2019) and Mplus version 8.2.

This document assumes a basic understanding of how to work with R and/or Mplus. Users less familiar with these programs are encouraged to consult any of the resources available describing the use of these programs (Hirschfeld and Von Brachel, 2014; Komperda, 2017; Muthén and Muthén, 2017; Rosseel, 2020). Unless otherwise noted, the code provided here is intended to be entered directly into the software and is written in a different font to distinguish it from explanatory text.

Simulation and Visualization of Data in R

Simulation of Identical Group Data

The data used for the examples in the main article are simulated data created in R to follow the structure of the fictional Perceived Relevance of Chemistry Questionnaire (PRCQ). The PRCQ is conceptualized as containing three fictitious subconstructs: Importance of Chemistry (IC),

Connectedness of Chemistry (CC), and Applications of Chemistry (AC). Additionally, the fictitious PRCQ is designed to be a 12-item instrument, where there are four items designed to measure each of the three subconstructs. To simulate this data in R first requires the installation and loading of the package `simstandard` (Schneider, 2019) which requires other dependent packages such as `dplyr` (Wickham *et al.*, 2019) to be installed as well.

```
install.packages("simstandard")
library(simstandard)
```

Syntax from the `lavaan` factor analysis package (Rosseel, 2012) is used to specify a three-factor model with four items associated with each factor. For this model, named `PRCQ`, items 1–4 are associated with the IC factor, 5–8 with the CC factor, and 9–12 with the AC factor. All items are assigned to have the same strength of association with their respective factors, a standardized value of 0.8. This value was chosen as it is relatively strong but not perfect association. In addition, each factor was simulated as having a weak association with the other factors. IC and CC have an association of 0.3, IC and AC have an association of 0.2 and CC and AC have an association of 0.1.

```
PRCQ<-'
      IC =~ 0.8*I1 + 0.8*I2 + 0.8*I3 + 0.8*I4
      CC =~ 0.8*I5 + 0.8*I6 + 0.8*I7 + 0.8*I8
      AC =~ 0.8*I9 + 0.8*I10 + 0.8*I11 + 0.8*I12

      IC  =~  0.3*CC
      IC  =~  0.2*AC
      CC  =~  0.1*AC
'
```

Now, observed data that follow the relations described by the model can be simulated. The `set.seed()` function is used to ensure reproducibility across uses by simulating the same

pseudorandom data each time the code is run. Following the example from the main text, data are simulated separately for 1000 fictional students in the STEM majors group and for 1000 students in the non-STEM majors group. A column named `group` is added to distinguish the data from each group and the two datasets are combined to form the new dataset named `combined`.

```
set.seed(1234)
STEM <- sim_standardized(PRCQ, n = 1000, observed = T, latent = F,
errors = F)
nonSTEM <- sim_standardized(PRCQ, n = 1000, observed = T, latent = F,
errors = F)

STEM$group<-"STEM"
nonSTEM$group<-"nonSTEM"

combined<-rbind(STEM, nonSTEM)
```

The data generated with `sim_standardized()` are standardized meaning they have an average value of 0 and standard deviation of 1 as well as a normal distribution. Descriptive statistics for the complete dataset and for each group within the dataset can be generated using the `describe()` and `describeBy()` functions in the `psych` package (Revelle, 2018) and are shown in Figure ESI4.1 and ESI4.2. Note that statistics are not generated for the `group` variable as it is a character, not a number.

```
library(psych)

describe(combined)

describeBy(combined, group="group")
```

```
> describe(combined)
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 2000 -0.01 0.98  -0.01   -0.01 0.97 -3.26 3.22  6.48  0.03    -0.02 0.02
I2        2 2000  0.00 1.00  -0.06    0.00 1.01 -3.44 3.44  6.88  0.02    -0.05 0.02
I3        3 2000 -0.03 0.99  -0.02   -0.03 0.98 -3.22 3.07  6.28  0.01    -0.05 0.02
I4        4 2000 -0.03 1.01   0.00   -0.03 1.02 -3.06 3.47  6.53  0.03    -0.12 0.02
I5        5 2000 -0.02 0.98  -0.01   -0.01 0.97 -3.13 3.39  6.52 -0.08    -0.06 0.02
I6        6 2000 -0.02 0.99  -0.01   -0.02 1.00 -3.53 3.36  6.89 -0.03     0.04 0.02
I7        7 2000  0.00 0.99   0.01    0.00 1.01 -3.03 4.22  7.24 -0.01     0.04 0.02
I8        8 2000  0.00 1.00   0.00    0.01 1.01 -3.63 3.03  6.66 -0.06    -0.03 0.02
I9        9 2000  0.01 0.97   0.03    0.01 0.97 -3.05 3.61  6.65  0.01     0.02 0.02
I10      10 2000  0.03 1.00   0.02    0.02 1.00 -2.95 3.40  6.35  0.11    -0.07 0.02
I11      11 2000  0.02 0.98   0.05    0.01 0.95 -2.88 3.48  6.36  0.04     0.00 0.02
I12      12 2000  0.01 0.98   0.01    0.00 0.98 -3.58 4.11  7.69  0.15     0.11 0.02
group*   13 2000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
```

Figure ESI4.1. Output from the `describe()` function using the dataset named `combined`.

```
> describeBy(combined, group="group")

 Descriptive statistics by group
group: nonSTEM
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 1000 -0.01 1.00  -0.04   -0.01 0.99 -3.26 3.15  6.42  0.06    -0.06 0.03
I2        2 1000 -0.02 1.01  -0.08   -0.02 1.00 -3.44 2.84  6.28 -0.03     0.02 0.03
I3        3 1000 -0.03 0.98  -0.01   -0.03 0.99 -2.90 3.07  5.97  0.01    -0.06 0.03
I4        4 1000 -0.05 1.02  -0.03   -0.05 1.06 -3.06 3.47  6.53  0.05    -0.17 0.03
I5        5 1000 -0.02 0.98  -0.01   -0.01 0.97 -3.13 3.39  6.52 -0.13     0.09 0.03
I6        6 1000 -0.05 1.01  -0.02   -0.04 1.03 -3.53 3.19  6.73 -0.12    -0.06 0.03
I7        7 1000 -0.04 1.01  -0.02   -0.03 1.04 -3.03 3.15  6.17 -0.05    -0.07 0.03
I8        8 1000 -0.01 1.02   0.00    0.00 1.00 -3.63 2.98  6.61 -0.08     0.12 0.03
I9        9 1000  0.04 0.97   0.06    0.04 0.97 -3.05 3.61  6.65  0.05     0.13 0.03
I10      10 1000  0.07 0.99   0.05    0.06 1.01 -2.74 3.08  5.82  0.12    -0.16 0.03
I11      11 1000  0.04 0.98   0.06    0.03 1.00 -2.66 3.35  6.00  0.10    -0.17 0.03
I12      12 1000  0.05 0.97   0.04    0.04 0.98 -3.58 4.11  7.69  0.15     0.35 0.03
group*   13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
------------------------------------------------------------------------------------
group: STEM
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 1000  0.00 0.97   0.01    0.00 0.92 -3.02 3.22  6.23  0.00     0.01 0.03
I2        2 1000  0.01 0.99  -0.05    0.01 1.04 -2.74 3.44  6.18  0.07    -0.15 0.03
I3        3 1000 -0.03 1.00  -0.04   -0.03 0.99 -3.22 3.06  6.28  0.02    -0.06 0.03
I4        4 1000  0.00 1.00   0.00   -0.01 0.98 -3.05 3.21  6.26  0.02    -0.07 0.03
I5        5 1000 -0.01 0.98  -0.02   -0.01 0.99 -2.95 2.95  5.90 -0.03    -0.21 0.03
I6        6 1000  0.00 0.98  -0.01    0.00 0.96 -3.22 3.36  6.58  0.07     0.12 0.03
I7        7 1000  0.03 0.97   0.03    0.03 0.97 -2.91 4.22  7.12  0.05     0.13 0.03
I8        8 1000  0.01 0.98   0.02    0.02 1.02 -3.03 3.03  6.06 -0.03    -0.23 0.03
I9        9 1000 -0.02 0.97   0.00   -0.01 0.98 -2.79 2.89  5.68 -0.03    -0.11 0.03
I10      10 1000 -0.01 1.00  -0.03   -0.01 0.99 -2.95 3.40  6.35  0.10     0.00 0.03
I11      11 1000  0.00 0.97   0.04    0.00 0.90 -2.88 3.48  6.36 -0.02     0.15 0.03
I12      12 1000 -0.02 0.99  -0.03   -0.03 0.98 -2.67 3.14  5.81  0.15    -0.12 0.03
group*   13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
```

Figure ESI4.2. Output by group from the `describeBy()` function using the dataset named `combined`.

Additionally, the data are complete with no missing cases. These data may not be representative of the type of data obtained in chemistry education research using a non-fictional assessment instrument. For the purposes of this example, as in the main body of the text, this dataset will continue to be used. Further procedures in the ESI will demonstrate converting the data from continuous into categorical, which may better match authentic data.

Simulation of Data with Unequal Factor Loadings and Unequal Item Means

The previous section described the simulation of data for two groups using the same model in each group. To illustrate the effect of invariance at different levels, modifications were made to the data. The data are simulated to highlight specific issues that could be encountered (i.e., noninvariant loadings, noninvariant intercepts) but are unlikely to be representative of authentic data which could have numerous issues simultaneously. The model below is used to simulate data with a lower association between AC and I10 for the non-STEM majors group (changed to 0.3 instead of 0.8), as used to generate Figure 4.4 in the manuscript. This data is combined with the original STEM majors data to create the `combined.invar.load` dataset.

```
PRCQ.invar.load<-'
      IC =~ 0.8*I1 + 0.8*I2 + 0.8*I3 + 0.8*I4
      CC =~ 0.8*I5 + 0.8*I6 + 0.8*I7 + 0.8*I8
      AC =~ 0.8*I9 + 0.3*I10 + 0.8*I11 + 0.8*I12

      IC  =~  0.3*CC
      IC  =~  0.2*AC
      CC  =~  0.1*AC
'

nonSTEM.invar.load <- sim_standardized(PRCQ.invar.load, n = 1000,
observed = T, latent = F, errors = F)

nonSTEM.invar.load$group<-"nonSTEM"

combined.invar.load<-rbind(STEM, nonSTEM.invar.load)
```

To create data with a higher mean for I3 in the STEM majors group, as used to generate Figures 4.4 and 4.5 in the manuscript, a new dataset is created from the original STEM majors data and constant of 2 is added to all values for I3 in this new data. The STEM majors data is combined with the original non-STEM majors data to create a `combined.invar.mean` dataset. The `describeBy()` function can be used to confirm differences between the groups as seen in the descriptive statistcs in Figure ESI3.

```
STEM.invar.mean<-STEM
STEM.invar.mean$I3<-STEM.invar.mean$I3+2

STEM.invar.mean$group<-"STEM"

combined.invar.mean<-rbind(STEM.invar.mean, nonSTEM)

describeBy(combined.invar.mean, group="group")
```

```
> describeBy(combined.invar.mean, group="group")

 Descriptive statistics by group
group: nonSTEM
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 1000 -0.01 1.00  -0.04   -0.01 0.99 -3.26 3.15  6.42  0.06    -0.06 0.03
I2        2 1000 -0.02 1.01  -0.08   -0.02 1.00 -3.44 2.84  6.28 -0.03     0.02 0.03
I3        3 1000 -0.03 0.98  -0.01   -0.03 0.99 -2.90 3.07  5.97  0.01    -0.06 0.03
I4        4 1000 -0.05 1.02  -0.03   -0.05 1.06 -3.06 3.47  6.53  0.05    -0.17 0.03
I5        5 1000 -0.02 0.98  -0.01   -0.01 0.97 -3.13 3.39  6.52 -0.13     0.09 0.03
I6        6 1000 -0.05 1.01  -0.02   -0.04 1.03 -3.53 3.19  6.73 -0.12    -0.06 0.03
I7        7 1000 -0.04 1.01  -0.02   -0.03 1.04 -3.03 3.15  6.17 -0.05    -0.07 0.03
I8        8 1000 -0.01 1.02   0.00    0.00 1.00 -3.63 2.98  6.61 -0.08     0.12 0.03
I9        9 1000  0.04 0.97   0.06    0.04 0.97 -3.05 3.61  6.65  0.05     0.13 0.03
I10      10 1000  0.07 0.99   0.05    0.06 1.01 -2.74 3.08  5.82  0.12    -0.16 0.03
I11      11 1000  0.04 0.98   0.06    0.03 1.00 -2.66 3.35  6.00  0.10    -0.17 0.03
I12      12 1000  0.05 0.97   0.04    0.04 0.98 -3.58 4.11  7.69  0.15     0.35 0.03
group*   13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
------------------------------------------------------------------------------------
group: STEM
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 1000  0.00 0.97   0.01    0.00 0.92 -3.02 3.22  6.23  0.00     0.01 0.03
I2        2 1000  0.01 0.99  -0.05    0.01 1.04 -2.74 3.44  6.18  0.07    -0.15 0.03
I3        3 1000  1.97 1.00   1.96    1.97 0.99 -1.22 5.06  6.28  0.02    -0.06 0.03
I4        4 1000  0.00 1.00   0.00   -0.01 0.98 -3.05 3.21  6.26  0.02    -0.07 0.03
I5        5 1000 -0.01 0.98  -0.02   -0.01 0.99 -2.95 2.95  5.90 -0.03    -0.21 0.03
I6        6 1000  0.00 0.98  -0.01    0.00 0.96 -3.22 3.36  6.58  0.07     0.12 0.03
I7        7 1000  0.03 0.97   0.03    0.03 0.97 -2.91 4.22  7.12  0.05     0.13 0.03
I8        8 1000  0.01 0.98   0.02    0.02 1.02 -3.03 3.03  6.06 -0.03    -0.23 0.03
I9        9 1000 -0.02 0.97   0.00   -0.01 0.98 -2.79 2.89  5.68 -0.03    -0.11 0.03
I10      10 1000 -0.01 1.00  -0.03   -0.01 0.99 -2.95 3.40  6.35  0.10     0.00 0.03
I11      11 1000  0.00 0.97   0.04    0.00 0.90 -2.88 3.48  6.36 -0.02     0.15 0.03
I12      12 1000 -0.02 0.99  -0.03   -0.03 0.98 -2.67 3.14  5.81  0.15    -0.12 0.03
group*   13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
```

Figure ESI4.3. Output by group from the `describeBy()` function using the dataset named `combined.invar.mean` showing different means for I3 across groups.

Visualization of Data

The R code in this section can be used to generate the data visualizations (correlations and distributions) shown in Figures 4.1–4.5 of the manuscript. Correlation plots can be made with the `corrplot` package (Wei and Simko, 2017). To use the `corrplot()` function, the numeric variables are selected from the `combined` dataset and a correlation matrix is generated with the `cor()` function. Additional function arguments are used to specify that colored boxes should be

plotted (`method="color"`), the text should be in the diagonal of the matrix in black (`tl.pos="d", tl.col="black"`), only the lower diagonal of the correlation matrix should be visualized (`type="lower"`), and that grey grid lines should appear (`addgrid.col="grey"`). Specifying the size of the margins is done to make room for the plot title (`mar=c(0,0,1,0)`).

```
library(dplyr)
library(corrplot)

combined %>% select(I1:I12) %>%   cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey", mar=c(0,0,1,0))
```

Similar plots can be generated for subsets of the data by filtering the `combined` dataset using the group variable (`filter(group=="STEM")`).

```
combined %>% filter(group=="STEM") %>% select(I1:I12) %>% cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey", title="STEM Majors",
   mar=c(0,0,1,0))

combined %>% filter(group=="nonSTEM") %>% select(I1:I12) %>% cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey",title="Non-STEM Majors",
   mar=c(0,0,1,0))
```

Using the `combined.invar.load` dataset will produce Figure 3 images from the manuscript.

```
combined.invar.load %>% select(I1:I12) %>%   cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
    type="lower", addgrid.col="grey",
    title="Combined Data Varied\n Strength of Association for I10",
   mar=c(0,0,1,0))

combined.invar.load %>% filter(group=="STEM") %>% select(I1:I12) %>%
   cor() %>% corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey", title="STEM Majors",
   mar=c(0,0,1,0))

combined.invar.load %>% filter(group=="nonSTEM")%>% select(I1:I12) %>%
   cor() %>% corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey",title="Non-STEM Majors",
   mar=c(0,0,1,0))
```

The Figure 4.4 images from the manuscript are produced using the same method with the `combined.invar.mean` dataset.

```
combined.invar.mean %>% select(I1:I12) %>%   cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col = "grey",
   title="Combined Data\n Varied Mean for I3",mar=c(0,0,1,0))

combined.invar.mean %>% filter(group=="STEM") %>% select(I1:I12) %>%
  cor() %>% corrplot(., method="color", tl.pos="d", tl.col="black",
  type="lower", addgrid.col = "grey", title="STEM Majors",
  mar=c(0,0,1,0))

combined.invar.mean %>% filter(group=="nonSTEM") %>% select(I1:I12)
   %>%  cor() %>% corrplot(., method="color", tl.pos="d",
   tl.col="black", type="lower", addgrid.col = "grey",
   title="Non-STEM Majors", mar=c(0,0,1,0))
```

In order to generate the boxplot Figure 4.5 of the manuscript the package `reshape2` (Wickham, 2007) is needed to restructure the dataset and the package `ggplot2` (Wickham, 2016) is used to create the plot. First, the STEM and non-STEM groups are given more descriptive names since those will appear in the figure legend. The groups are also ordered as with STEM Majors first since the default setting would put the groups in alphabetical order.

```
library(ggplot2)
library(reshape2)

combined.invar.mean$group<-ifelse(combined.invar.mean$group=="STEM",
                                  "STEM Majors", "Non-STEM Majors")
combined.invar.mean$group<-ordered(combined.invar.mean$group,
                             levels=c("STEM Majors", "Non-STEM Majors"))
```

Next, the `melt()` function is used to create a long-format dataset where each group, variable (Item), and value occupies a single column. This long format is necessary for plotting using the function `ggplot()` with `geom_boxplot()`. In this boxplot the x-axis is the group and the y-axis is the value for each variable (`x=group, y=value, fill=group`). Faceting by

variable (`facet_grid(.~variable)`) plots each item separately, yet within a single plot. The remainder of the code provides graphical parameters.

```
melt.mean<-combined.invar.mean %>%
                          select(I1:I12, group) %>% melt(id="group")
melt.mean$group<-melt.mean$group %>% as.factor()

ggplot(melt.mean, aes(x=group, y=value, fill=group))+
  geom_boxplot() + facet_grid(.~variable) + theme_bw() +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(),
        axis.ticks.x=element_blank(), axis.title.y=element_blank(),
        legend.position="bottom") +
  scale_fill_discrete(name="Group")
```

Conducting Invariance Testing

This section provides an overview of how to conduct measurement invariance testing using two popular software platforms, R and Mplus. Results obtained from both pieces of software will be similar, so the selection of software depends on the preferences of the researcher. In addition to R and Mplus there are other tools available for conducting measurement invariance testing, including SAS, LISREL, EQS, or the AMOS add-in for SPSS. A helpful comparison of software for structural equation modeling with multiple groups can be found in Narayana (2012) and Byrne (2004) provides a guide to AMOS.

Before introducing the specific steps to take within R and Mplus, it is worthwhile to note the default settings of both software packages. Within R, the package `lavaan` is generally used for factor analyses and in this package the default way to provide scale to the factor is to fix the value of the first item loading to one. In Mplus, the factor is given scale by setting its variance to one. Both methods are acceptable ways of identifying the model and will give equivalent results.

However, each of these methods has different implications in the context of measurement invariance testing with multiple groups.

The method of setting the factor variance to one (as in Mplus) in both groups is generally not recommended for multigroup measurement invariance testing as it implies that the latent variable has the same variance in both groups. This is described as homogeneity of variance for the latent variables. Though conceptually similar to the test for homogeneity of variance used in *t*-tests and ANOVAs, in a latent framework this is an untestable assumption (Hancock *et al.*, 2009, 168).

In the first method, used within `lavaan`, setting an item loading to one, the default is to use the first item on the scale. When the first item on the scale is set to be one for both groups the rest of the series of structural equations will be solved assuming this item has the same loading value in both groups. Yet, there is no way to know for certain if that assumption is true or if there are other scale items that would have been better to set equivalent. This seemingly inconsequential decision can have major implications for interpretation of results and researchers are advised to think carefully about which item may be best to set equal across groups based on either theoretical or observable grounds (Bontempo and Hofer, 2007; Hancock *et al.*, 2009).

Invariance Testing with R – Continuous Data

Within the R software, the package `lavaan`, previously used to generate the simulated data, can be used to test confirmatory factor (CFA) models as well as structural equation models (SEM). The function for performing CFA, `cfa()` contains built-in arguments to set various model

parameters equal for invariance testing (Hirschfeld and Von Brachel, 2014), making invariance testing a relatively simple process. In this section, the steps for measurement invariance testing will follow those in the main article using the `combined.invar.load` dataset to generate the fit index data from Table 4.1 in the manuscript. The general process for invariance testing within R is that of building up from the least constrained model (i.e., configural invariance) to the most constrained model (i.e., conservative invariance). Identical steps can be followed for the other datasets and fit indices resulting from these tests are provided later sections.

Step 0: Establishing Baseline Model

Following the steps outline in the manuscript, the baseline model is tested for each group separately. The model is specified in the same manner as was used to generate the simulated data with the main difference being that values for the loadings and associations between factors are not assigned but will be estimated by the software from the data. This model is named `model.test` to distinguish it from the model used to simulate the data.

```
library(lavaan)

model.test<-'
      IC =~ I1 + I2 + I3 + I4
      CC =~ I5 + I6 + I7 + I8
      AC =~ I9 + I10 + I11 + I12
      '
```

The function `cfa()` is now used to examine how well the data fit the proposed model. The maximum likelihood (ML) estimator is used as the data are continuous and normally distributed and are therefore appropriate for the ML estimator. Additionally, this follows the steps in the main article and aligns with the estimator used to determine the suggested fit index cut off values (Hu and Bentler, 1999). In situations where the data are known to be nonnormally

distributed the robust maximum likelihood estimator (MLR) is more appropriate and can be specified with the command `estimator="MLR"`. The results from ML and MLR are equivalent if the data are normal, and interested readers can confirm this for themselves since `lavaan` prints the output of both ML and MLR simultaneously when MLR is used. Later sections of this ESI will describe how to modify the code to accommodate categorical data. Finally, specify that the mean structure (intercepts) should be explicitly shown.

```
STEM.step0<-cfa(data=combined.invar.mean %>% filter(group=="STEM
                Majors"), model=model.test, estimator="ML",
                meanstructure=TRUE)
```

The `summary()` function provides a convenient way to view the fit statistics and model parameters from the model that was just fit to the STEM majors data.

```
summary(STEM.step0, standardized=TRUE, fit.measures=TRUE)
```

Though the output provided by `summary()` is extensive the key fit indices are indicated by boxes in Figure ESI4.4. Note that the fit indices match Table 4.1 in the manuscript and show essentially perfect fit: CFI > 0.95; SRMR < 0.08; RMSEA < 0.06 (Hu and Bentler, 1999).

```
> STEM.step0<-cfa(data = combined.invar.mean %>% filter(group=="STEM
Majors"), model = model.test, estimator="ML", meanstructure=TRUE)
> summary(STEM.step0, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 27 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         39

  Number of observations                          1000

Model Test User Model:

  Test statistic                                65.438
  Degrees of freedom                                51
  P-value (Chi-square)                           0.084

Model Test Baseline Model:

  Test statistic                              6052.309
  Degrees of freedom                                66
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.998
  Tucker-Lewis Index (TLI)                       0.997

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -13835.349
  Loglikelihood unrestricted model (H1)     -13802.630

  Akaike (AIC)                               27748.698
  Bayesian (BIC)                             27940.100
  Sample-size adjusted Bayesian (BIC)        27816.234

Root Mean Square Error of Approximation:

  RMSEA                                          0.017
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.028
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.021
```

Figure ESI4.4. Summary output for testing baseline model (Step 0) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

The same code can be executed using the non-STEM majors data and nearly identical fit is achieved (Figure ESI4.5).

```
nonSTEM.step0<-cfa(data=combined.invar.mean %>% filter(group=="Non-
                   STEM Majors"), model=model.test,
                   estimator="ML", meanstructure=TRUE)

summary(nonSTEM.step0, standardized=TRUE, fit.measures=TRUE)
```

```
> nonSTEM.step0<-cfa(data = combined.invar.mean %>% filter(group=="Non-
STEM Majors"), model = model.test, estimator="ML", meanstructure=TRUE)
> summary(nonSTEM.step0, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 30 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         39

  Number of observations                          1000

Model Test User Model:

  Test statistic                                51.931
  Degrees of freedom                                51
  P-value (Chi-square)                           0.437

Model Test Baseline Model:

  Test statistic                              6015.854
  Degrees of freedom                                66
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -13981.961
  Loglikelihood unrestricted model (H1)     -13955.996

  Akaike (AIC)                               28041.922
  Bayesian (BIC)                             28233.325
  Sample-size adjusted Bayesian (BIC)        28109.459

Root Mean Square Error of Approximation:

  RMSEA                                          0.004
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.021
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.016
```

Figure ESI4.5. R summary output for testing baseline model (Step 0) with unmodified non-STEM majors data highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

Looking through the rest of the `summary()` output gives the values for the model parameters. The column `Std.all` is most typically reported when standardized model parameters are given. For both groups, these model parameters (Figures ESI4.6 & ESI4.7) match those used to simulate the data (loadings of 0.80 as well as associations between the three factors of approximately 0.3, 0.2, and 0.1). Examining the values of the intercept terms in both groups

shows that in the STEM majors group (Figure ESI4.6) the intercept for I3 is larger than in the non-STEM majors group by a value of 2, as specified in the model used to simulate the data.

Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~						
I1	1.000				0.761	0.787
I2	1.030	0.040	25.575	0.000	0.784	0.793
I3	1.053	0.040	26.167	0.000	0.802	0.802
I4	1.075	0.041	26.429	0.000	0.818	0.815
CC =~						
I5	1.000				0.774	0.793
I6	1.017	0.039	26.278	0.000	0.788	0.805
I7	1.001	0.039	25.971	0.000	0.775	0.796
I8	1.011	0.039	26.213	0.000	0.783	0.801
AC =~						
I9	1.000				0.765	0.787
I10	1.061	0.041	25.790	0.000	0.812	0.810
I11	0.993	0.039	25.222	0.000	0.760	0.781
I12	1.027	0.041	25.268	0.000	0.786	0.792

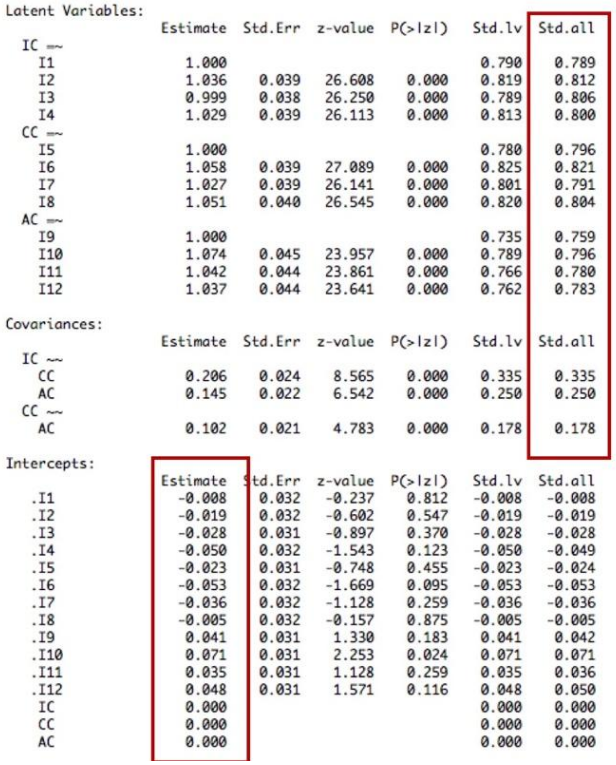
Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.174	0.023	7.666	0.000	0.295	0.295
AC	0.119	0.022	5.453	0.000	0.205	0.205
CC ~~						
AC	0.087	0.022	3.965	0.000	0.146	0.146

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.004	0.031	-0.116	0.908	-0.004	-0.004
.I2	0.012	0.031	0.391	0.696	0.012	0.012
.I3	1.974	0.032	62.479	0.000	1.974	1.976
.I4	-0.003	0.032	-0.082	0.935	-0.003	-0.003
.I5	-0.009	0.031	-0.304	0.761	-0.009	-0.010
.I6	0.004	0.031	0.142	0.887	0.004	0.004
.I7	0.033	0.031	1.077	0.281	0.033	0.034
.I8	0.013	0.031	0.431	0.666	0.013	0.014
.I9	-0.018	0.031	-0.571	0.568	-0.018	-0.018
.I10	-0.006	0.032	-0.191	0.849	-0.006	-0.006
.I11	-0.001	0.031	-0.026	0.979	-0.001	-0.001
.I12	-0.021	0.031	-0.679	0.497	-0.021	-0.021
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Figure ESI4.6. R summary output for testing baseline model (Step 0) with unchanged STEM majors data highlighting standardized model parameters and intercepts.



Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~						
I1	1.000				0.790	0.789
I2	1.036	0.039	26.608	0.000	0.819	0.812
I3	0.999	0.038	26.250	0.000	0.789	0.806
I4	1.029	0.039	26.113	0.000	0.813	0.800
CC =~						
I5	1.000				0.780	0.796
I6	1.058	0.039	27.089	0.000	0.825	0.821
I7	1.027	0.039	26.141	0.000	0.801	0.791
I8	1.051	0.040	26.545	0.000	0.820	0.804
AC =~						
I9	1.000				0.735	0.759
I10	1.074	0.045	23.957	0.000	0.789	0.796
I11	1.042	0.044	23.861	0.000	0.766	0.780
I12	1.037	0.044	23.641	0.000	0.762	0.783

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.206	0.024	8.565	0.000	0.335	0.335
AC	0.145	0.022	6.542	0.000	0.250	0.250
CC ~~						
AC	0.102	0.021	4.783	0.000	0.178	0.178

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.008	0.032	-0.237	0.812	-0.008	-0.008
.I2	-0.019	0.032	-0.602	0.547	-0.019	-0.019
.I3	-0.028	0.031	-0.897	0.370	-0.028	-0.028
.I4	-0.050	0.032	-1.543	0.123	-0.050	-0.049
.I5	-0.023	0.031	-0.748	0.455	-0.023	-0.024
.I6	-0.053	0.032	-1.669	0.095	-0.053	-0.053
.I7	-0.036	0.032	-1.128	0.259	-0.036	-0.036
.I8	-0.005	0.032	-0.157	0.875	-0.005	-0.005
.I9	0.041	0.031	1.330	0.183	0.041	0.042
.I10	0.071	0.031	2.253	0.024	0.071	0.071
.I11	0.035	0.031	1.128	0.259	0.035	0.036
.I12	0.048	0.031	1.571	0.116	0.048	0.050
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Figure ESI4.7. R summary output for testing baseline model (Step 0) with unchanged non-STEM majors data highlighting standardized model parameters and intercepts.

It is important to note that this difference in intercept for I3 between the groups (Figures ESI4.6 & ESI4.7) did not affect the overall fit of each group (Figures ESI4.4 & ESI4.5) because the parameters in each group were allowed to vary as needed to best fit the model. The purpose of testing these baseline models is to ensure that each group has a reasonable fit to the model before constraining any parameters to be equal across groups.

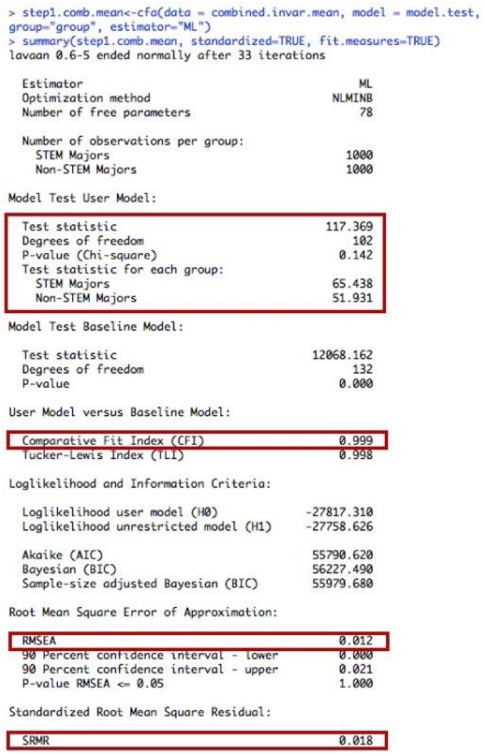
Step 1: Configural Invariance

The next step of invariance testing fits the model to both groups of data simultaneously. Within the `cfa()` function this is easily accomplished by specifying that groups are present and providing the name of the grouping variable (`group="group"`).

```
step1.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML")
```

```
summary(step1.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

Output from testing this model provides both an overall model chi square and the individual group chi square values obtained from Step 0 (Figure ESI4.8). The rest of the fit indices (CFI, RMSEA, and SRMR) are provided for the overall model. As show in Table 1 of the manuscript the fit indices for the configural model are essentially perfect. Further exploration of the model parameters shows that parameters for both groups have been estimated separately and match those in Step 0.



```
> step1.comb.mean<-cfa(data = combined.invar.mean, model = model.test,
group="group", estimator="ML")
> summary(step1.comb.mean, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 33 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         78

  Number of observations per group:
    STEM Majors                                   1000
    Non-STEM Majors                               1000

Model Test User Model:

  Test statistic                               117.369
  Degrees of freedom                               102
  P-value (Chi-square)                           0.142
  Test statistic for each group:
    STEM Majors                                 65.438
    Non-STEM Majors                             51.931

Model Test Baseline Model:

  Test statistic                             12068.162
  Degrees of freedom                               132
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.999
  Tucker-Lewis Index (TLI)                       0.998

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -27817.310
  Loglikelihood unrestricted model (H1)     -27758.626

  Akaike (AIC)                               55790.620
  Bayesian (BIC)                             56227.490
  Sample-size adjusted Bayesian (BIC)        55979.680

Root Mean Square Error of Approximation:

  RMSEA                                          0.012
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.021
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.018
```

Figure ESI4.8. R summary output for configural invariance model (Step 1) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

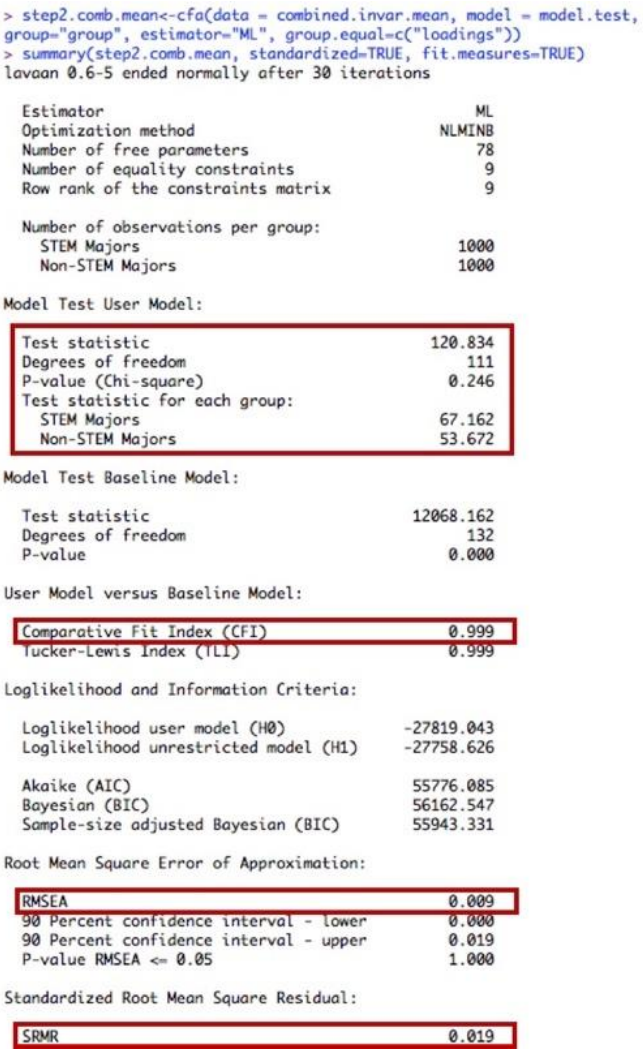
Step 2: Metric Invariance (Weak)

To test for metric invariance (weak) the `group.equal` argument is used to specify that the loadings must be held constant across the two groups.

```
step2.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML",
                     group.equal=c("loadings"))

summary(step2.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

The fit indices for the metric invariance model (Figure ESI4.9) again match Table 4.1 in the manuscript and show essentially perfect fit. As described in the manuscript the change in fit index values can be calculated by hand but the *p*-value for the Δchi square must be computed.



```
> step2.comb.mean<-cfa(data = combined.invar.mean, model = model.test,
group="group", estimator="ML", group.equal=c("loadings"))
> summary(step2.comb.mean, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 30 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         78
  Number of equality constraints                     9
  Row rank of the constraints matrix                 9

  Number of observations per group:
    STEM Majors                                   1000
    Non-STEM Majors                               1000

Model Test User Model:

  Test statistic                               120.834
  Degrees of freedom                               111
  P-value (Chi-square)                           0.246
  Test statistic for each group:
    STEM Majors                                 67.162
    Non-STEM Majors                             53.672

Model Test Baseline Model:

  Test statistic                             12068.162
  Degrees of freedom                               132
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.999
  Tucker-Lewis Index (TLI)                       0.999

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -27819.043
  Loglikelihood unrestricted model (H1)     -27758.626

  Akaike (AIC)                               55776.085
  Bayesian (BIC)                             56162.547
  Sample-size adjusted Bayesian (BIC)        55943.331

Root Mean Square Error of Approximation:

  RMSEA                                          0.009
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.019
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.019
```

Figure ESI4.9. R summary output for metric invariance model (Step 2) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

Examination of the model parameters is again done by groups (Figure ESI4.10) but shows that certain parameters have been constrained equal across the groups by assigning them a parameter name given in parenthesis (e.g., .p2.). Here the unstandardized loading values in the `Estimate` column are equal in both groups but the `Std.all` column values vary slightly. This is because the factors parameters (i.e., factor covariances) have not been constrained equal across groups and therefore affect the standardized loading values. Note that only the loadings have been assigned parameter names since these are the only parameters constrained equal across groups.

Group 1 [STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.770	0.791
I2	(.p2.)	1.034	0.028	36.926	0.000	0.796	0.799
I3	(.p3.)	1.025	0.028	36.999	0.000	0.789	0.796
I4	(.p4.)	1.053	0.028	37.170	0.000	0.811	0.812
CC =~							
I5		1.000				0.763	0.788
I6	(.p6.)	1.038	0.027	37.797	0.000	0.793	0.807
I7	(.p7.)	1.015	0.028	36.792	0.000	0.775	0.796
I8	(.p8.)	1.032	0.028	37.269	0.000	0.788	0.803
AC =~							
I9		1.000				0.759	0.784
I10	(.10.)	1.067	0.030	35.415	0.000	0.810	0.809
I11	(.11.)	1.016	0.030	34.431	0.000	0.771	0.787
I12	(.12.)	1.031	0.030	34.755	0.000	0.783	0.790

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.174	0.022	7.785	0.000	0.295	0.295
AC	0.119	0.022	5.501	0.000	0.204	0.204
CC ~~						
AC	0.084	0.021	3.976	0.000	0.146	0.146

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.004	0.031	-0.115	0.908	-0.004	-0.004
.I2	0.012	0.031	0.388	0.698	0.012	0.012
.I3	1.974	0.031	62.973	0.000	1.974	1.991
.I4	-0.003	0.032	-0.082	0.934	-0.003	-0.003
.I5	-0.009	0.031	-0.306	0.760	-0.009	-0.010
.I6	0.004	0.031	0.141	0.887	0.004	0.004
.I7	0.033	0.031	1.077	0.281	0.033	0.034
.I8	0.013	0.031	0.430	0.667	0.013	0.014
.I9	-0.018	0.031	-0.574	0.566	-0.018	-0.018
.I10	-0.006	0.032	-0.191	0.849	-0.006	-0.006
.I11	-0.001	0.031	-0.026	0.979	-0.001	-0.001
.I12	-0.021	0.031	-0.681	0.496	-0.021	-0.022
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Group 2 [Non-STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.781	0.784
I2	(.p2.)	1.034	0.028	36.926	0.000	0.807	0.806
I3	(.p3.)	1.025	0.028	36.999	0.000	0.800	0.812
I4	(.p4.)	1.053	0.028	37.170	0.000	0.822	0.804
CC =~							
I5		1.000				0.790	0.801
I6	(.p6.)	1.038	0.027	37.797	0.000	0.820	0.819
I7	(.p7.)	1.015	0.028	36.792	0.000	0.801	0.791
I8	(.p8.)	1.032	0.028	37.269	0.000	0.815	0.802
AC =~							
I9		1.000				0.742	0.762
I10	(.10.)	1.067	0.030	35.415	0.000	0.791	0.797
I11	(.11.)	1.016	0.030	34.431	0.000	0.754	0.773
I12	(.12.)	1.031	0.030	34.755	0.000	0.765	0.785

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.207	0.024	8.738	0.000	0.335	0.335
AC	0.145	0.022	6.625	0.000	0.250	0.250
CC ~~						
AC	0.104	0.022	4.804	0.000	0.178	0.178

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.008	0.031	-0.239	0.811	-0.008	-0.008
.I2	-0.019	0.032	-0.606	0.544	-0.019	-0.019
.I3	-0.028	0.031	-0.891	0.373	-0.028	-0.028
.I4	-0.050	0.032	-1.535	0.125	-0.050	-0.049
.I5	-0.023	0.031	-0.743	0.458	-0.023	-0.023
.I6	-0.053	0.032	-1.674	0.094	-0.053	-0.053
.I7	-0.036	0.032	-1.128	0.259	-0.036	-0.036
.I8	-0.005	0.032	-0.158	0.875	-0.005	-0.005
.I9	0.041	0.031	1.324	0.185	0.041	0.042
.I10	0.071	0.031	2.250	0.024	0.071	0.071
.I11	0.035	0.031	1.137	0.256	0.035	0.036
.I12	0.048	0.031	1.568	0.117	0.048	0.050
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Figure ESI4.10. R summary output for metric invariance model (Step 2) with STEM majors data having modified I3 intercept highlighting constraints on loading terms.

Step 3: Scalar Invariance (Strong)

Testing for scalar invariance only requires the addition of constraining the intercept terms to be equal, in addition to the loadings that were already constrained in Step 2.

```
step3.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML",
                     group.equal=c("loadings", "intercepts"))

summary(step3.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

Again, matching the values found in Table 4.1 of the manuscript, the fit indices for the strict invariance model (Figure ESI4.11) indicate poor data-model fit, which is to be expected since the intercept terms were not simulated to be equal across groups. Notice that the chi square values for the individual groups give some indication that the problem is in the STEM Majors group, as it has a much larger (worse) chi square value. Figure ESI4.12 shows that now the intercept terms are constrained to be equal across groups.

```
> step3.comb.mean<-cfa(data = combined.invar.mean, model = model.test,
group="group", estimator="ML", group.equal=c("loadings", "intercepts"))
> summary(step3.comb.mean, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 49 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         81
  Number of equality constraints                    21
  Row rank of the constraints matrix                21

  Number of observations per group:
    STEM Majors                                   1000
    Non-STEM Majors                               1000

Model Test User Model:

  Test statistic                              2267.834
  Degrees of freedom                           120
  P-value (Chi-square)                           0.000
  Test statistic for each group:
    STEM Majors                               2067.996
    Non-STEM Majors                            199.838

Model Test Baseline Model:

  Test statistic                             12068.162
  Degrees of freedom                           132
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.820
  Tucker-Lewis Index (TLI)                       0.802

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -28892.543
  Loglikelihood unrestricted model (H1)     -27758.626

  Akaike (AIC)                               57905.085
  Bayesian (BIC)                             58241.139
  Sample-size adjusted Bayesian (BIC)        58050.516

Root Mean Square Error of Approximation:

  RMSEA                                          0.134
  90 Percent confidence interval - lower         0.129
  90 Percent confidence interval - upper         0.139
  P-value RMSEA <= 0.05                          0.000

Standardized Root Mean Square Residual:

  SRMR                                           0.191
```

Figure ESI4.11. R summary output for metric invariance model (Step 3) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

Group 1 [STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.759	0.781
I2	(.p2.)	1.043	0.030	35.235	0.000	0.791	0.798
I3	(.p3.)	1.044	0.036	29.207	0.000	0.793	0.402
I4	(.p4.)	1.068	0.030	35.439	0.000	0.810	0.815
CC =~							
I5		1.000				0.763	0.788
I6	(.p6.)	1.039	0.027	37.804	0.000	0.793	0.807
I7	(.p7.)	1.015	0.028	36.786	0.000	0.775	0.796
I8	(.p8.)	1.032	0.028	37.252	0.000	0.787	0.803
AC =~							
I9		1.000				0.759	0.784
I10	(.10.)	1.067	0.030	35.442	0.000	0.810	0.809
I11	(.11.)	1.015	0.029	34.442	0.000	0.771	0.787
I12	(.12.)	1.032	0.030	34.788	0.000	0.783	0.790

Covariances:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~							
CC		0.170	0.022	7.612	0.000	0.294	0.294
AC		0.108	0.022	4.973	0.000	0.187	0.187
CC ~~							
AC		0.084	0.021	3.976	0.000	0.146	0.146

Intercepts:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	(.31.)	0.066	0.029	2.296	0.022	0.066	0.067
.I2	(.32.)	0.073	0.029	2.480	0.013	0.073	0.073
.I3	(.33.)	0.324	0.034	9.449	0.000	0.324	0.164
.I4	(.34.)	0.049	0.030	1.641	0.101	0.049	0.049
.I5	(.35.)	0.003	0.028	0.119	0.905	0.003	0.003
.I6	(.36.)	-0.004	0.029	-0.141	0.888	-0.004	-0.004
.I7	(.37.)	0.019	0.029	0.665	0.506	0.019	0.020
.I8	(.38.)	0.024	0.029	0.823	0.410	0.024	0.024
.I9	(.39.)	-0.018	0.028	-0.629	0.529	-0.018	-0.018
.I10	(.40.)	0.001	0.030	0.024	0.981	0.001	0.001
.I11	(.41.)	-0.013	0.029	-0.436	0.663	-0.013	-0.013
.I12	(.42.)	-0.017	0.029	-0.581	0.561	-0.017	-0.017
IC		0.000				0.000	0.000
CC		0.000				0.000	0.000
AC		0.000				0.000	0.000

Group 2 [Non-STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.770	0.777
I2	(.p2.)	1.043	0.030	35.235	0.000	0.803	0.804
I3	(.p3.)	1.044	0.036	29.207	0.000	0.805	0.789
I4	(.p4.)	1.068	0.030	35.439	0.000	0.822	0.804
CC =~							
I5		1.000				0.789	0.800
I6	(.p6.)	1.039	0.027	37.804	0.000	0.820	0.819
I7	(.p7.)	1.015	0.028	36.786	0.000	0.802	0.791
I8	(.p8.)	1.032	0.028	37.252	0.000	0.814	0.802
AC =~							
I9		1.000				0.742	0.762
I10	(.10.)	1.067	0.030	35.442	0.000	0.792	0.797
I11	(.11.)	1.015	0.029	34.442	0.000	0.753	0.773
I12	(.12.)	1.032	0.030	34.788	0.000	0.765	0.785

Covariances:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~							
CC		0.205	0.023	8.726	0.000	0.337	0.337
AC		0.144	0.022	6.636	0.000	0.252	0.252
CC ~~							
AC		0.104	0.022	4.805	0.000	0.178	0.178

Intercepts:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	(.31.)	0.066	0.029	2.296	0.022	0.066	0.066
.I2	(.32.)	0.073	0.029	2.480	0.013	0.073	0.073
.I3	(.33.)	0.324	0.034	9.449	0.000	0.324	0.318
.I4	(.34.)	0.049	0.030	1.641	0.101	0.049	0.048
.I5	(.35.)	0.003	0.028	0.119	0.905	0.003	0.003
.I6	(.36.)	-0.004	0.029	-0.141	0.888	-0.004	-0.004
.I7	(.37.)	0.019	0.029	0.665	0.506	0.019	0.019
.I8	(.38.)	0.024	0.029	0.823	0.410	0.024	0.023
.I9	(.39.)	-0.018	0.028	-0.629	0.529	-0.018	-0.018
.I10	(.40.)	0.001	0.030	0.024	0.981	0.001	0.001
.I11	(.41.)	-0.013	0.029	-0.436	0.663	-0.013	-0.013
.I12	(.42.)	-0.017	0.029	-0.581	0.561	-0.017	-0.017
IC		-0.146	0.037	-3.922	0.000	-0.189	-0.189
CC		-0.039	0.037	-1.053	0.292	-0.049	-0.049
AC		0.059	0.036	1.634	0.102	0.079	0.079

Figure ESI4.12. R summary output for scalar invariance model (Step 3) with STEM majors data having modified I3 intercept highlighting constraints on loading and intercept terms.

Step 4: Conservative Invariance (Strict)

Given the poor fit of the scalar invariance model, and out of range delta fit index values, it is not appropriate to go on to consider the strict invariance model. However, interested readers can test this model by adding "`residuals`" to the `group.equal` argument (residuals is another name for the error variance terms).

```
Step4.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML",
                     group.equal=c("loadings", "intercepts", "residuals"))

summary(step4.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

Exporting Data from R to Mplus

Data within R can be exported in a variety of familiar formats including txt, csv, and xlsx. Most conveniently for those working in Mplus there is also a package, `MplusAutomation`

(Hallquist and Wiley, 2018), that allows for direct export of data in the correct Mplus format, dat. The correct format for Mplus requires data to not have any header information, such as column names. The `MplusAutomation` package also generates appropriate code to communicate the structure of the file to Mplus. The R code below shows how to export the simulated PRCQ data to Mplus and request the input file, which provides the code to use within Mplus to import the dat file in the correct format to be read by Mplus. Note that the group variable had been stored as a categorical factor within R and must be changed to a numeric variable for export. In this case the first group (STEM majors) will become 1 and the second group will become 2. This can be confirmed with the `describeBy()` function.

```
library(MplusAutomation)

combined.invar.mean$group<-combined.invar.mean$group %>%
                           as.numeric()
describeBy(combined.invar.mean, group="group")

prepareMplusData(combined.invar.mean,
                 filename="InvarianceMean.dat", inpfile = TRUE,
                 keepCols=c("I1", "I2", "I3", "I4","I5", "I6",
                 "I7", "I8", "I9", "I10", "I11", "I12", "group"))
```

As a result of these commands R will create two new files, `InvarianceMean.dat` and `InvarianceMean.inp` in the working directory of your R session. If you are unsure of where your working directory resides, use the command `getwd()`.

Invariance Testing with Mplus – Continuous Data

Invariance testing in Mplus begins by opening the `inp` file generated previously or creating a new `inp` file for your own data. At the top of the `inp` file will be a title for the model being tested, the name of the data file, and the names of the variables in the data file. As before, the first step should be to test the model for each group individual. This is accomplished with the

command USEOBSERVATIONS. Then the model to be tested is specified, this step is similar to lavaan but uses the term BY instead of =~ to denote relations between items and factors.

```
TITLE: STEM Majors Group Step 0
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==1;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```

The output for this model provides the same fit indices and standardized model parameters (Figure ESI4.13) as produced in R (Figures ESI4.4 & ESI4.6) and shown in Table 4.1 of the manuscript.

MODEL FIT INFORMATION

Number of Free Parameters	39
Loglikelihood	
H0 Value	-13835.349
H1 Value	-13802.630
Information Criteria	
Akaike (AIC)	27748.698
Bayesian (BIC)	27940.100
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	27816.234
Chi-Square Test of Model Fit	
Value	65.438
Degrees of Freedom	51
P-Value	0.0841
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.017
90 Percent C.I.	0.000 0.028
Probability RMSEA <= .05	1.000
CFI/TLI	
CFI	0.998
TLI	0.997
Chi-Square Test of Model Fit for the Baseline Model	
Value	6052.309
Degrees of Freedom	66
P-Value	0.0000
SRMR (Standardized Root Mean Square Residual)	
Value	0.021

STDYX Standardization

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
IC	BY				
I1		0.787	0.015	52.142	0.000
I2		0.793	0.015	53.520	0.000
I3		0.802	0.014	55.343	0.000
I4		0.815	0.014	58.397	0.000
CC	BY				
I5		0.793	0.015	53.401	0.000
I6		0.805	0.014	55.833	0.000
I7		0.796	0.015	53.890	0.000
I8		0.801	0.015	54.917	0.000
AC	BY				
I9		0.787	0.015	51.270	0.000
I10		0.810	0.014	55.917	0.000
I11		0.781	0.016	50.057	0.000
I12		0.792	0.015	52.257	0.000
CC	WITH				
IC		0.295	0.033	8.851	0.000
AC	WITH				
IC		0.205	0.035	5.866	0.000
CC		0.146	0.036	4.107	0.000
Intercepts					
I1		-0.004	0.032	-0.116	0.908
I2		0.012	0.032	0.391	0.696
I3		1.976	0.054	36.366	0.000
I4		-0.003	0.032	-0.082	0.935
I5		-0.010	0.032	-0.304	0.761
I6		0.004	0.032	0.142	0.887
I7		0.034	0.032	1.077	0.282
I8		0.014	0.032	0.431	0.666
I9		-0.018	0.032	-0.571	0.568
I10		-0.006	0.032	-0.191	0.849
I11		-0.001	0.032	-0.026	0.979
I12		-0.021	0.032	-0.679	0.497

Figure ESI4.13. Mplus summary output baseline model (Step 0) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA, SRMR, and standardized model parameters.

Similar code can be used for the non-STEM majors group and again the results (Figure ESI4.14) will agree with the R output (Figures ESI4.15 & ESI4.17 as well as Table 4.1 of the manuscript.

```
TITLE: Non-STEM Majors Group Step 0
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==2;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```

```
MODEL FIT INFORMATION

Number of Free Parameters                       39

Loglikelihood

          H0 Value                       -13981.961
          H1 Value                       -13955.996

Information Criteria

          Akaike (AIC)                    28041.922
          Bayesian (BIC)                  28233.325
          Sample-Size Adjusted BIC        28109.459
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                              51.931
          Degrees of Freedom                     51
          P-Value                            0.4374

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.004
          90 Percent C.I.                    0.000  0.021
          Probability RMSEA <= .05           1.000

CFI/TLI

          CFI                                1.000
          TLI                                1.000

Chi-Square Test of Model Fit for the Baseline Model

          Value                           6015.854
          Degrees of Freedom                   66
          P-Value                          0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.016
```

```
STDYX Standardization

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

IC       BY
   I1                  0.789      0.015     52.836      0.000
   I2                  0.812      0.014     58.082      0.000
   I3                  0.806      0.014     56.771      0.000
   I4                  0.800      0.015     55.114      0.000

CC       BY
   I5                  0.796      0.015     54.388      0.000
   I6                  0.821      0.014     60.409      0.000
   I7                  0.791      0.015     53.320      0.000
   I8                  0.804      0.014     56.398      0.000

AC       BY
   I9                  0.759      0.017     44.993      0.000
   I10                 0.796      0.015     51.583      0.000
   I11                 0.780      0.016     48.554      0.000
   I12                 0.783      0.016     49.300      0.000

CC       WITH
   IC                  0.335      0.032     10.308      0.000

AC       WITH
   IC                  0.250      0.034      7.268      0.000
   CC                  0.178      0.035      5.041      0.000

Intercepts
   I1                 -0.008      0.032     -0.237      0.812
   I2                 -0.019      0.032     -0.602      0.547
   I3                 -0.028      0.032     -0.897      0.370
   I4                 -0.049      0.032     -1.542      0.123
   I5                 -0.024      0.032     -0.748      0.455
   I6                 -0.053      0.032     -1.668      0.095
   I7                 -0.036      0.032     -1.128      0.259
   I8                 -0.005      0.032     -0.157      0.875
   I9                  0.042      0.032      1.330      0.184
   I10                 0.071      0.032      2.250      0.024
   I11                 0.036      0.032      1.128      0.259
   I12                 0.050      0.032      1.570      0.116
```

Figure ESI4.14. Mplus summary output baseline model (Step 0) with Non-STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA, SRMR, and standardized model parameters.

Step 1: Configural Invariance

To test configural invariance within Mplus, the model is specified separately for each group. The `!` notation is used to insert comments within the Mplus model code. To provide results aligned with the R output the `@1` notation is used to identify the model by standardizing the loading for the first item on each factor. This is the default setting for the R `cfa()` function, but models in both programs can also be run by standardizing the factors instead of the loadings as a method of identifying the model.

Next the factor intercept is set to zero using brackets and `@0` notation. By default, Mplus assumes that item intercepts should be equal across groups, these can be freely estimated using the bracket notation. Item error variances are coded without the use of brackets. Specifying the same model for the second group will tell Mplus to estimate parameters for both models separately.

```
TITLE: Combined Dataset with Mean Differences Step 1 (Configural)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allowing item intercepts to be freely estimated
  [I1-I12];

! Allowing item error variances to be freely estimated
  I1-I12;

! Specifying the same model for the second group will cause
! all parameters to be freely estimated for the second group
```

```
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

  [IC@0];
  [CC@0];
  [AC@0];

  [I1-I12];

  I1-I12;
OUTPUT:
STANDARDIZED;
```

The output from this model (Figure ESI4.15) matches the fit indices in Table 4.1 of the manuscript for the configural model and both the unstandardized and standardized model parameters for the STEM majors group (Figure ESI4.16) and non-STEM majors group match those found using R (Figures ESI4.6 & ESI4.7).

```
MODEL FIT INFORMATION

Number of Free Parameters                       78

Loglikelihood

          H0 Value                       -27817.310
          H1 Value                       -27758.626

Information Criteria

          Akaike (AIC)                    55790.620
          Bayesian (BIC)                  56227.490
          Sample-Size Adjusted BIC        55979.680
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                             117.369
          Degrees of Freedom                    102
          P-Value                            0.1418

Chi-Square Contribution From Each Group

          STEM                               65.438
          NONSTEM                            51.931

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.012
          90 Percent C.I.                    0.000  0.021
          Probability RMSEA <= .05           1.000

CFI/TLI

          CFI                                0.999
          TLI                                0.998

Chi-Square Test of Model Fit for the Baseline Model

          Value                          12068.162
          Degrees of Freedom                   132
          P-Value                           0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.019
```

Figure ESI4.15. Mplus summary output for configural invariance (Step 1) with STEM majors data having modified I3 intercept highlighting fit information.

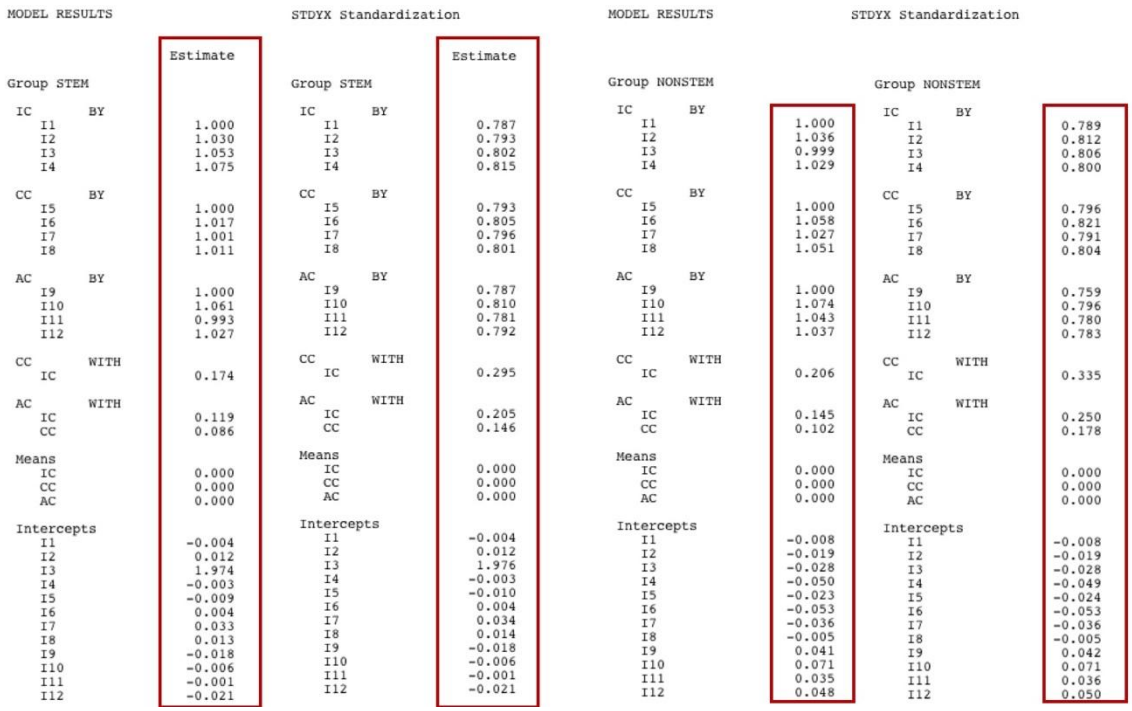


Figure ESI4.16. Mplus output for configural invariance (Step 1) with STEM majors data having modified I3 intercept highlighting unstandardized and standardized model parameters for both groups.

Step 2: Metric Invariance (Weak)

Metric invariance is tested by assigning the same parameter names to the loading terms in each group. In this example the names `L1-L12` are assigned to each of the loading parameters. Repeating this assignment in the second group will cause Mplus to set the unstandardized value of the parameters equal.

```
TITLE: Combined Dataset with Mean Differences Step 2 (Weak)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-L12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
```

```
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allowing item intercepts to be freely estimated
  [I1-I12];

! Allowing item error variances to be freely estimated
  I1-I12;

! Specifying the same model for the second group will force
! loadings to be equivalent across groups while other
! parameters are freely estimated
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

  [IC@0];
  [CC@0];
  [AC@0];

  [I1-I12];

  I1-I12;

OUTPUT:
STANDARDIZED;
```

The output from this model (Figure ESI4.17) matches the fit indices in Table 4.1 of the manuscript for the weak invariance model and now the unstandardized parameters are equal across groups (Figure ESI4.18) while the intercepts are allowed to differ. As before, the standardized parameters differ slightly, but are aligned with the R output (Figure ESI4.10).

MODEL FIT INFORMATION

Number of Free Parameters 69

Loglikelihood

H0 Value	-27819.043
H1 Value	-27758.626

Information Criteria

Akaike (AIC)	55776.085
Bayesian (BIC)	56162.547
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	55943.331

Chi-Square Test of Model Fit

Value	120.834
Degrees of Freedom	111
P-Value	0.2464

Chi-Square Contribution From Each Group

STEM	67.159
NONSTEM	53.674

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.009	
90 Percent C.I.	0.000	0.019
Probability RMSEA <= .05	1.000	

CFI/TLI

CFI	0.999
TLI	0.999

Chi-Square Test of Model Fit for the Baseline Model

Value	12068.162
Degrees of Freedom	132
P-Value	0.0000

SRMR (Standardized Root Mean Square Residual)

Value	0.019

Figure ESI4.17. Mplus summary output for metric invariance (Step 2) with STEM majors data having modified I3 intercept highlighting fit information.

MODEL RESULTS

		Estimate
Group STEM		
IC	BY	
I1		1.000
I2		1.034
I3		1.025
I4		1.053
CC	BY	
I5		1.000
I6		1.038
I7		1.015
I8		1.032
AC	BY	
I9		1.000
I10		1.067
I11		1.016
I12		1.031
CC	WITH	
IC		0.174
AC	WITH	
IC		0.119
CC		0.085
Means		
IC		0.000
CC		0.000
AC		0.000
Intercepts		
I1		-0.004
I2		0.012
I3		1.974
I4		-0.003
I5		-0.009
I6		0.004
I7		0.033
I8		0.013
I9		-0.018
I10		-0.006
I11		-0.001
I12		-0.021

STDYX Standardization

		Estimate
Group STEM		
IC	BY	
I1		0.791
I2		0.799
I3		0.796
I4		0.812
CC	BY	
I5		0.788
I6		0.807
I7		0.796
I8		0.803
AC	BY	
I9		0.784
I10		0.809
I11		0.787
I12		0.790
CC	WITH	
IC		0.295
AC	WITH	
IC		0.204
CC		0.146
Means		
IC		0.000
CC		0.000
AC		0.000
Intercepts		
I1		-0.004
I2		0.012
I3		1.991
I4		-0.003
I5		-0.010
I6		0.004
I7		0.034
I8		0.014
I9		-0.018
I10		-0.006
I11		-0.001
I12		-0.022

MODEL RESULTS

Group NONSTEM		
IC	BY	
I1		1.000
I2		1.034
I3		1.025
I4		1.053
CC	BY	
I5		1.000
I6		1.038
I7		1.015
I8		1.032
AC	BY	
I9		1.000
I10		1.067
I11		1.016
I12		1.031
CC	WITH	
IC		0.207
AC	WITH	
IC		0.145
CC		0.104
Means		
IC		0.000
CC		0.000
AC		0.000
Intercepts		
I1		-0.008
I2		-0.019
I3		-0.028
I4		-0.050
I5		-0.023
I6		-0.053
I7		-0.036
I8		-0.005
I9		0.041
I10		0.071
I11		0.035
I12		0.048

STDYX Standardization

Group NONSTEM		
IC	BY	
I1		0.784
I2		0.806
I3		0.812
I4		0.804
CC	BY	
I5		0.801
I6		0.819
I7		0.791
I8		0.802
AC	BY	
I9		0.762
I10		0.797
I11		0.773
I12		0.785
CC	WITH	
IC		0.335
AC	WITH	
IC		0.250
CC		0.178
Means		
IC		0.000
CC		0.000
AC		0.000
Intercepts		
I1		-0.008
I2		-0.019
I3		-0.028
I4		-0.049
I5		-0.023
I6		-0.053
I7		-0.036
I8		-0.005
I9		0.042
I10		0.071
I11		0.036
I12		0.050

Figure ESI4.18. Mplus output for metric invariance (Step 2) with STEM majors data having modified I3 intercept highlighting unstandardized and standardized model parameters for both groups.

Step 3: Scalar Invariance (Strong)

Scalar invariance is tested by assigning the same parameter names to the intercept terms in both groups while also removing the restrictions on the mean of the factor terms for the second group using the * notation. As seen in Table 4.1 of the manuscript and in the R output, this significantly worsens the value of all fit indices (Figure ESI4.19) indicating that scalar invariance has not been achieved due to differences in loadings across groups. As before, the Mplus model parameters (Figure ESI4.20) are similar to those produced by R (Figure ESI4.12).

```
TITLE: Combined Dataset with Mean Differences Step 3 (Strong)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allowing item intercepts to be freely estimated in one group
! assigning a parameter name so they will be equal across groups
  [I1-I12] (M1-M12);

! Allowing item error variances to be freely estimated
  I1-I12;

! Specifying the same model parameter names for the second group
! will cause loadings and item intercepts to be equivalent across
! groups while other parameters are freely estimated
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Allowing factor intercepts vary
  [IC*];
  [CC*];
```

```
  [AC*];

  [I1-I12] (M1-M12);

  I1-I12;

OUTPUT:
STANDARDIZED;
```

```
MODEL FIT INFORMATION

Number of Free Parameters                       60

Loglikelihood

          H0 Value                       -28912.498
          H1 Value                       -27758.626

Information Criteria

          Akaike (AIC)                    57944.997
          Bayesian (BIC)                  58281.051
          Sample-Size Adjusted BIC        58090.428
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                            2307.745
          Degrees of Freedom                   120
          P-Value                           0.0000

Chi-Square Contribution From Each Group

          STEM                              229.366
          NONSTEM                          2078.380

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.135
          90 Percent C.I.                    0.130  0.140
          Probability RMSEA <= .05           0.000

CFI/TLI

          CFI                                0.817
          TLI                                0.798

Chi-Square Test of Model Fit for the Baseline Model

          Value                            12068.162
          Degrees of Freedom                   132
          P-Value                           0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.238
```

Figure ESI4.19. Mplus summary output for scalar invariance (Step 3) with STEM majors data having modified I3 intercept highlighting fit information.

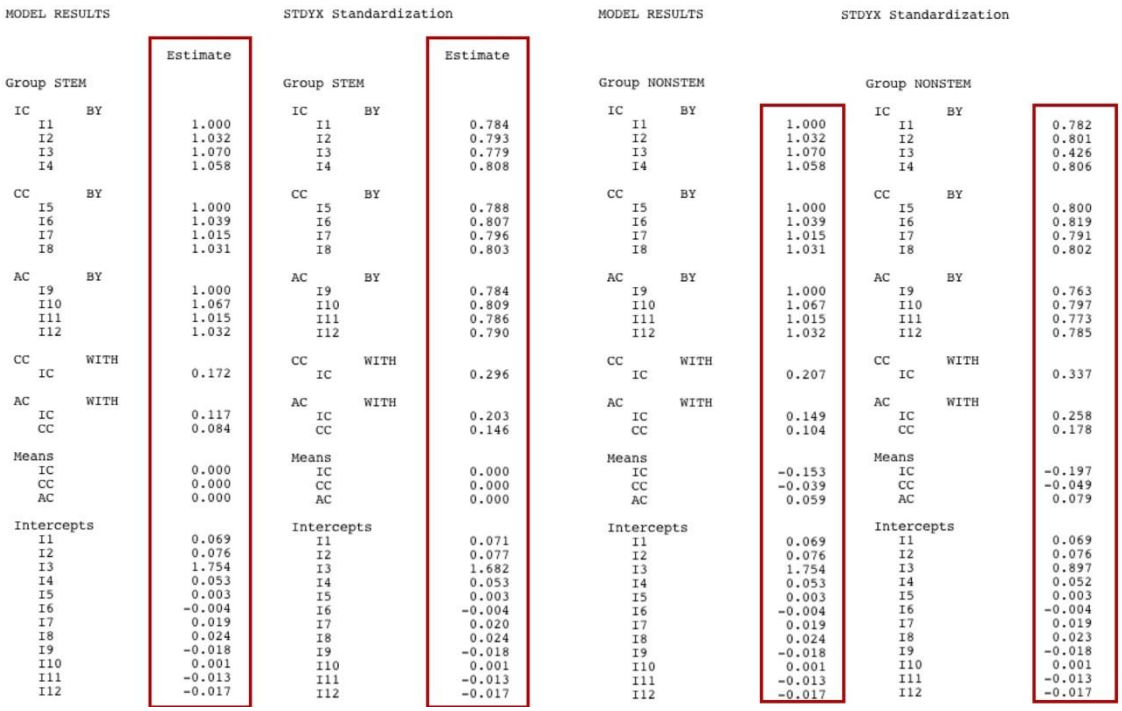


Figure ESI4.20. Mplus output for scalar invariance (Step 3) with STEM majors data having modified I3 intercept highlighting unstandardized and standardized model parameters for both groups.

Step 4: Conservative Invariance (Strict)

As noted previously, due to the poor fit of the scalar invariance model, you would stop at Step 3 and not go on to test Step 4 (conservative invariance with equal error variance terms). However, interested readers can test Step 4 in Mplus by providing the same name to the error variance parameters in both groups.

```
TITLE: Combined Dataset with Mean Differences Step 4 (Strict)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);
```

```
! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allow item intercepts to be freely estimated in one group but
! assigning a parameter name so they will be equal across groups
  [I1-I12] (M1-M12);

! Allow item error variances to be freely estimated but
! assigning a parameter name so they will be equal across groups
  I1-I12 (E1-E12);

! Specifying the same model parameter names for the second group
! will cause loadings and item intercepts to be equivalent across
! groups while other parameters are freely estimated
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Allowing factor intercepts vary
  [IC*];
  [CC*];
  [AC*];

  [I1-I12] (M1-M12);

  I1-I12(E1-E12);

OUTPUT:
STANDARDIZED;
```

Fit Indices for other Continuous Datasets

Tables ESI4.1 & ESI4.2 show the data-model fit output from R produced from following the previous steps with the two other continuous datasets: `combined` and `combined.invar.load`.

Table ESI4.1. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined` Simulated Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	SRMR	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔSRMR	ΔRMSEA
0	STEM majors Baseline	65	51	0.084	0.998	0.021	0.017	-	-	-	-	-	-
0	Non-STEM majors Baseline	52	51	0.437	1.000	0.016	0.004	-	-	-	-	-	-
1	Configural	117	102	0.142	0.999	0.018	0.012	-	-	-	-	-	-
2	Metric	120	111	0.245	0.999	0.019	0.009	3	9	0.964	0.000	0.001	0.003
3	Scalar	127	120	0.311	0.999	0.020	0.008	7	9	0.637	0.000	0.001	0.001
4	Conservative	135	132	0.417	1.000	0.020	0.005	8	12	0.786	0.001	0.000	0.003

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Table ESI4.2. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.invar.load` Simulated Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	SRMR	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔSRMR	ΔRMSEA
0	STEM majors Baseline	65	51	0.084	0.998	0.021	0.017	-	-	-	-	-	-
0	Non-STEM majors Baseline	66	51	0.081	0.997	0.017	0.017	-	-	-	-	-	-
1	Configural	131	102	0.028	0.997	0.019	0.017	-	-	-	-	-	-
2	Metric	305	111	< 0.001	0.983	0.051	0.042	101	9	< 0.001	0.014	0.032	0.025
3	Scalar	310	120	< 0.001	0.984	0.051	0.040	5	9	0.834	0.001	0.000	0.002
4	Conservative	433	132	< 0.001	0.974	0.043	0.048	123	12	< 0.001	0.010	0.008	0.008

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Creating Ordered Categorical Data in R

As seen in the previous examples, the data simulation function in R creates continuous data which may not be representative of data collected from instruments used in chemistry education

research, which often have five-point Likert-type scales. The code below is used to take the original simulated datasets and turn them into Likert-type data by collapsing the full ranges of data for each item into five bins using the `cut()` function. Note that this process of creating categorical data from continuous data ensures that each bin will be populated, but issues with testing models can arise if authentic categorical data are collected with empty bins (e.g., no responses in the 1 category).

```
STEM.ord<-STEM
for(i in 1:12){
  var[i]<-paste0("I", i)
  STEM.ord[[var[i]]]<-as.numeric(cut(STEM[[var[i]]], breaks=5))
                }

nonSTEM.ord<-nonSTEM
for(i in 1:12){
  var[i]<-paste0("I", i)
  nonSTEM.ord[[var[i]]]<-as.numeric(cut(nonSTEM[[var[i]]],
breaks=5))
                }
combined.ord<-rbind(STEM.ord, nonSTEM.ord)


nonSTEM.invar.load.ord<-nonSTEM.invar.load
for(i in 1:12){
  var[i]<-paste0("I", i)
  nonSTEM.invar.load.ord[[var[i]]]<-
as.numeric(cut(nonSTEM.invar.load[[var[i]]], breaks=5))
}
combined.invar.load.ord<-rbind(STEM.ord, nonSTEM.invar.load.ord)

STEM.invar.mean.ord<-STEM.invar.mean
for(i in 1:12){
  var[i]<-paste0("I", i)
  STEM.invar.mean.ord[[var[i]]]<-
as.numeric(cut(STEM.invar.mean[[var[i]]], breaks=5))
}
combined.invar.mean.ord<-rbind(STEM.invar.mean.ord, nonSTEM.ord)
```

When data collected on Likert-type scales have fewer than seven categories or the full range of the response scale is not used by most respondents (i.e. a ceiling or floor effect) it is often recommended to treat the data as ordinal categorical data rather than continuous. In a factor

analysis framework, this type of data is best modeled using a robust diagonally weighted least squares estimator, such as WLSMV (Finney and DiStefano, 2013). A noticeable difference in working with ordinal data the software will compute thresholds which are used to map the categorical variables onto an assumed underlying normal distribution of latent item responses and therefore create a set of latent correlations. This process is can be conceptualized as the reverse of the process used to create ordered categorical data from the original continuous data show in prior steps.

The concept of thresholds can be visualized by plotting the distribution of values for an item both in its continuous and categorical form. For this example, responses to I1 in the continuous data are visualized with a density plot (Figure ESI4.21a) and I1 responses in the categorical data are visualized with a bar plot (Figure ESI4.21b) using the code below.

```
plot(density(combined$I1),
     main="Density Plot for Combined Data Item I1 - Continuous",
     ylab="Frequency", xlab="Response")

barplot(prop.table(table(combined.ord$I1)),
        main="Frequency Plot for Combined Data Item I1 - Ordinal",
        ylab="Frequency", xlab="Response")
```

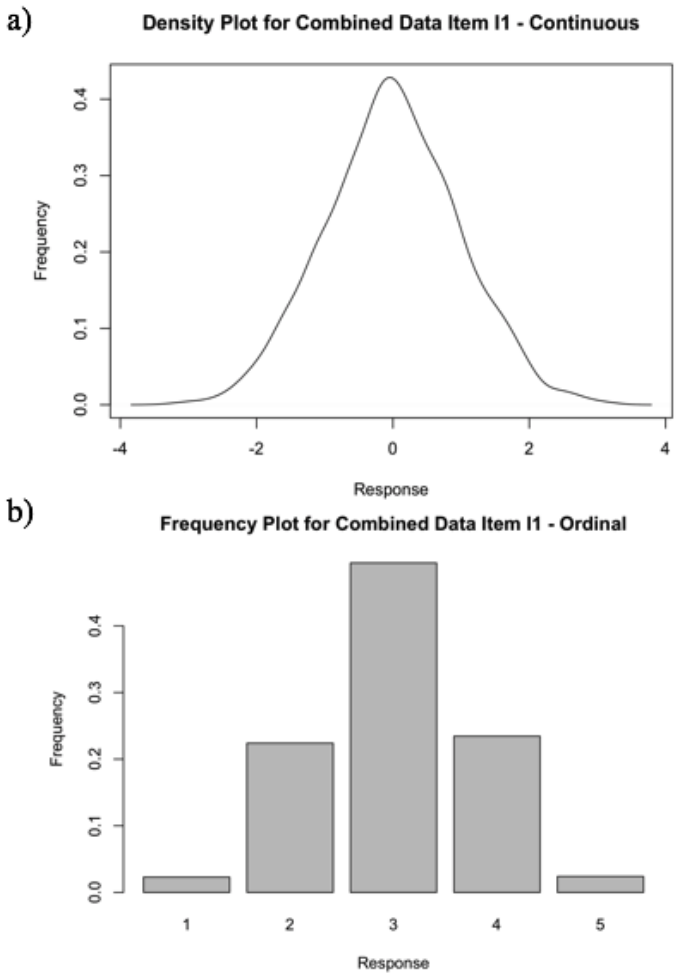


Figure ESI4.21. Density plot of continuous I1 responses (a) and frequency plot of categorical I1 responses (b)

Visual inspection of the two plots shows how the original continuous distribution aligns with the categorical data in that the middle responses have higher response frequencies and the extreme responses have lower response frequencies. When the ordinal data in Figure ESI21b are used to estimate a factor model, the software will assume the categorical data are representative of an underlying continuous variable (DiStefano and Morgan, 2014) and determine cut points, called thresholds, where the unobserved continuous distribution would have been divided to create the observed categorical distribution.

Since the categorical data used in this example were created from continuous data, we are able find the true cut points using the same code as before.

```
summary(cut(combined$I1, breaks=5))
```

Plotting these cut points (–1.97, –0.672, 0.624, and 1.92) on the continuous distribution (Figure ESI4.22) shows how the categorical data were simulated, and also provides insight into how the factor analysis itself will identify thresholds in the categorical data.

```
plot(density(combined$I1), main="Density Plot for Combined Data
     Item I1 - Continuous",ylab="Frequency", xlab="Response")
abline(v=c(-1.97, -0.672, 0.624, 1.92), col="grey")
```

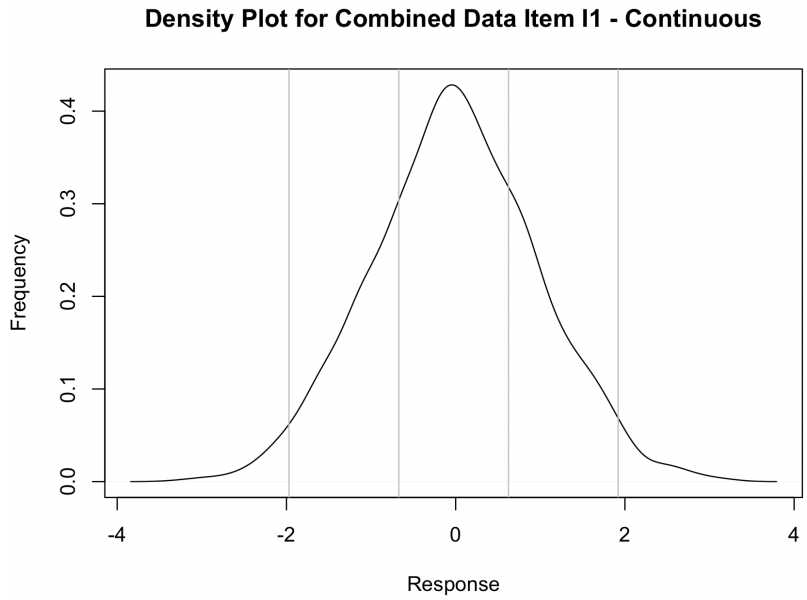


Figure ESI4.22. Density plot of continuous I1 responses showing cut points used to create categorical data.

Estimating Models with Ordered Categorical Data in R and Mplus

Running the factor models in R and also exporting the data for running in Mplus will provide an opportunity to see the threshold values established by the software. Full measurement

invariance testing steps will be described in later sections. Both programs will automatically switch to the correct estimator (WLSMV) when informed that the data are not continuous. In lavaan syntax the argument `ordered` is used.

```
combined.ord.cfa<-cfa(data = combined.ord, model = model.test,
                      ordered=c("I1", "I2", "I3", "I4", "I5",
                      "I6", "I7", "I8", "I9", "I10", "I11",
                      "I12"))
summary(combined.ord.cfa, standardized=TRUE, fit.measures=TRUE)

combined.ord$group<-combined.ord$group %>% as.factor() %>%
                    as.numeric()
prepareMplusData(combined.ord, filename="CombinedOrdinal.dat",
                 inpfile = T, keepCols=c("I1", "I2", "I3",
                 "I4","I5", "I6", "I7", "I8", "I9", "I10", "I11",
                 "I12", "group"))
```

In Mplus the variables are specified as `categorical`.

```
TITLE: Combined Ordinal Data - CFA Model
DATA: FILE = "CombinedOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
MISSING=.;

USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```

The full output of both programs can be examined to confirm similarities in how the data are treated as well as the matched fit indices and model parameters. Figure ESI4.23 shows the threshold values calculated by each program, indicated with the `t` notation in R and the `$` notation in Mplus. As expected, the thresholds for I1 are similar to those used to create the categorical data

from the continuous, even though neither R or Mplus had access to the continuous data when generating the threshold values.

a)

Thresholds:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
I1|t1	-1.919	0.058	-33.205	0.000	-1.919	-1.919
I1|t2	-0.645	0.030	-21.314	0.000	-0.645	-0.645
I1|t3	0.674	0.030	22.131	0.000	0.674	0.674
I1|t4	1.927	0.058	33.130	0.000	1.927	1.927

b)

Thresholds

I1$1	-1.919	0.058	-33.213	0.000
I1$2	-0.645	0.030	-21.319	0.000
I1$3	0.674	0.030	22.137	0.000
I1$4	1.927	0.058	33.138	0.000

Figure ESI4.23. Threshold values from R (a) and Mplus (b)

Data Model Fit for Ordered Categorical Data with WLSMV Estimator

The fit index cut off values recommended by Hu and Bentler (1999) were based on work using the maximum likelihood (ML) estimator which is appropriate for continuous data. Since a different estimator is used with categorical data, it is not appropriate to use the same Hu and Bentler recommendations for fit index cut off values. Simulation studies with the WLSMV estimator have indicated that more rigorous cut off values are best, particularly when the data contain a small number of categories or are severely nonnormal (Yu, 2002; Beauducel and Herzberg, 2006; DiStefano and Morgan, 2014). Recommendations for fit index values with the WLSMV estimator are CFI $\geq$ 0.95 and RMSEA $\leq$ 0.05. The SRMR is not recommended with the WLSMV estimator. In the context of invariance testing, less work has been done to determine recommended values for change in CFI and RMSEA values between models compared to the ML estimator. As with the fit indices themselves, simulation studies suggest either using more rigorous ΔCFI and ΔRMSEA values than those used with ML estimation or providing multiple sources of justification

for acceptable data-model fit potentially using different estimators to see if similar conclusions about invariance would be drawn (Sass *et al.*, 2014).

Invariance Testing with R – Ordered Categorical Data

Measurement invariance testing in R with categorical data can be conducted following similar steps as those used for continuous data. However, it should be noted that other researchers have advocated for a different order of steps or different sets of constraints when working with categorical data (Millsap and Yun-Tein, 2004; Wu and Estabrook, 2016; Svetina *et al.*, 2019). The primary differences when working with categorical data compared to continuous are that the ordinal nature of the data must be specified in order for the correct estimator to be used, and thresholds must be constrained along with other model parameters during invariance testing steps.

Also, unique to working with categorical data, a decision must be made about scaling of the underlying latent normal distribution for each set of item responses using either delta or theta scaling. In delta scaling the total variance of the latent response is set to 1 and in theta scaling the variance of the residual term is set to 1. These decisions primarily influence how the model parameters are identified. Theta scaling is appropriate for invariance research (Millsap and Yun-Tein, 2004) and was chosen for the analysis here, but it is possible to convert parameters between delta and theta scaling (Finney and DiStefano, 2013). Since theta scaling affects the residual terms, Step 4 of invariance testing (strict) is not necessary with categorical data when following this method.

The steps taken in this ESI will parallel those used previously for continuous data. The data used in this section are the categorical version of the continuous data used in previous examples where the mean for I3 was changed in the STEM majors group. The code for all steps of invariance testing in R with categorical data are specified below and the fit statistics are summarized in Table ESI4.3 using the WLSMV output from lavaan as given in the `Robust` column. Fit statistics for models using the other categorical datasets are provided in Tables ESI4.4 & ESI4.5.

Step 0: Establishing Baseline Model

The baseline model for each group is specified in the same way as the continuous data but now using the ordinal data set and specifying which variables are ordered categorical as well as the use of the theta parameterization. The same three factor model used for the continuous data is used for the categorical data.

```
STEM.step0.ord<-cfa(data = combined.invar.mean.ord %>%
                    filter(group==STEM), model=model.test,
                    ordered=c("I1", "I2", "I3", "I4", "I5", "I6",
                    "I7", "I8", "I9", "I10", "I11", "I12"),
                    parameterization="theta")

summary(STEM.step0.ord, standardized=TRUE, fit.measures=TRUE)

nonSTEM.step0.ord<-cfa(data=combined.invar.mean.ord %>%
                       filter(group=="nonSTEM"),
                       model=model.test, ordered=c("I1", "I2",
                       "I3", "I4", "I5", "I6", "I7", "I8", "I9",
                       "I10", "I11", "I12"),
                       parameterization="theta")

summary(nonSTEM.step0.ord, standardized=TRUE, fit.measures=TRUE)
```

Step 1: Configural Invariance

Configural invariance uses data from both groups while specifying the grouping variable.

```
step1.comb.mean.ord<-cfa(data=combined.invar.mean.ord,
                         group="group", model=model.test,
                         ordered=c("I1", "I2", "I3", "I4", "I5",
                         "I6", "I7", "I8", "I9", "I10", "I11",
                         "I12"), parameterization="theta")

 summary(step1.comb.mean.ord, standardized=TRUE,
         fit.measures=TRUE)
```

Step 2: Metric Invariance (Weak)

Metric invariance is tested by holding the loadings equal across groups.

```
step2.comb.mean.ord<-cfa(data=combined.invar.mean.ord,
                         group="group", model=model.test,
                         ordered=c("I1", "I2", "I3", "I4", "I5",
                         "I6", "I7", "I8", "I9", "I10", "I11",
                         "I12"), group.equal=c("loadings"),
                         parameterization="theta")

summary(step2.comb.mean.ord, standardized=TRUE,
fit.measures=TRUE)
```

Step 3: Scalar Invariance (Strong)

Adding the constraint of equal thresholds across groups is similar to holding intercepts equal to test for scalar invariance in continuous data.

```
step3.comb.mean.ord<-cfa(data=combined.invar.mean.ord,
                         group="group", model=model.test,
                         ordered=c("I1", "I2", "I3", "I4", "I5",
                         "I6", "I7", "I8", "I9", "I10", "I11",
                         "I12"), group.equal=c("loadings",
                         "thresholds"), parameterization="theta")

summary(step3.comb.mean.ord, standardized=TRUE,
fit.measures=TRUE)
```

Table ESI4.3. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.invar.mean` Simulated Categorical Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔRMSEA
0	STEM majors Baseline	81	51	0.005	0.996	0.024	-	-	-	-	-
0	Non-STEM majors Baseline	61	51	0.162	0.999	0.014	-	-	-	-	-
1	Configural	142	102	0.006	0.997	0.020	-	-	-	-	-
2	Metric	145	111	0.017	0.998	0.018	3	9	0.231	0.001	0.002
3	Scalar	869	144	< 0.001	0.953	0.071	724	9	< 0.001	0.045	0.053

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Invariance Testing with Mplus – Ordered Categorical Data

Following the previously shown steps, the categorical data in R are exported to Mplus by first converting the group variable from a text format into a numeric format.

```
combined.invar.mean.ord$group<-combined.ord$group %>% as.factor()
                                %>% as.numeric()

prepareMplusData(combined.invar.mean.ord,
                 filename="CombinedInvarMeanOrdinal.dat",
                 inpfile = T, keepCols=c("I1", "I2", "I3",
                 "I4","I5", "I6", "I7", "I8", "I9", "I10",
                 "I11", "I12", "group"))
```

As with lavaan, the default estimator in Mplus is ML but the software will adjust to an appropriate estimator for ordinal data (WLSMV) by specifying the item variables as categorical. The call for theta parameterization is also added and the models are specified separately for each group. Following these steps for R and Mplus should provide similar fit indices and model parameters.

Step 0: Establishing Baseline Model

```
TITLE: Categorical STEM Majors Group Step 0
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
MISSING=.;

USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==2;

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;


TITLE: Categorical Non-STEM Majors Group Step 0
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
MISSING=.;

USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==1;

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```

Step 1: Configural Invariance

By default, Mplus will constrain thresholds equal across groups so this must be released by freeing all thresholds for all variables. The notation to free the thresholds uses the $ character.

Four thresholds must be freed since four thresholds would be required to divide the underlying continuous distribution into five categories. As was done with the continuous data, the factor means are set to zero. The error variances are set to one for categorical data, in line with theta parameterization.

```
TITLE: Categorical Combined Dataset with Mean Differences Step 1
(Configural)
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = NonSTEM 2 = STEM);

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
! Model with standardized loading of first item on each factor
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

! Set error variances to 1
  I1-I12@1

MODEL STEM:
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
```

```
  [I1$4-I12$4*];


! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

! Set error variances to 1
  I1-I12@1

OUTPUT:
STANDARDIZED;
```

Step 2: Metric Invariance (Weak)

Loadings are constrained equal across groups by assigning the same name to the parameters in both groups. This is the same method used for invariance testing with the continuous data.

```
TITLE: Categorical Combined Dataset with Mean Differences Step 2
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = NonSTEM 2 = STEM);

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-L12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

! Set error variances to 1
  I1-I12@1

MODEL STEM:
```

```
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);


  ! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

  ! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

 ! Set error variances to 1
  I1-I12@1

OUTPUT:
STANDARDIZED;
```

Step 3: Scalar Invariance (Strong)

Mplus and lavaan differ in their default settings when thresholds are constrained equal across groups. To mimic the lavaan output the factor means and error variance terms for the second group are freed in the Mplus code. Freeing these parameters also aligns scalar invariance testing in the categorical data with the same step for the continuous data. Recall that the goal of Step 3 is to determine if the factors are being measured on the same scale in each group so that factor means can be compared across groups. Therefore, one group should have a mean of zero in order to function as a reference while the mean of the other group is freely estimated.

```
TITLE: Categorical Combined Dataset with Mean Differences Step 3
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = NonSTEM 2 = STEM);

ANALYSIS: PARAMETERIZATION=THETA;
```

```
MODEL:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

  [IC@0];
  [CC@0];
  [AC@0];

  I1-I12@1


MODEL STEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Fix thresholds equal by not specifying for this group

! Set factor means free
[IC*];
[CC*];
[AC*];

! Set error variances free
I1-I12*

OUTPUT:
STANDARDIZED;
```

Fit Indices for Invariance Testing Steps with other Simulated Categorical Data

Tables ESI4.4 & 4.5 show the data-model fit output from R produced from following the previous steps with the two other categorical datasets: `combined.ord` and `combined.invar.load.ord`.

Table ESI4.4. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.ord` Simulated Categorical Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔRMSEA
0	STEM majors Baseline	81	51	0.005	0.996	0.024	-	-	-	-	-
0	Non-STEM majors Baseline	61	51	0.162	0.999	0.014	-	-	-	-	-
1	Configural	142	102	0.006	0.997	0.020	-	-	-	-	-
2	Metric	145	111	0.017	0.998	0.018	3	9	0.964	0.001	0.002
3	Scalar	869	144	<0.001	0.953	0.071	724	33	< 0.001	0.045	0.053

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Table ESI4.5. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.invar.load.ord` Simulated Categorical Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔRMSEA
0	STEM majors Baseline	81	51	0.005	0.996	0.024	-	-	-	-	-
0	Non-STEM majors Baseline	40	51	0.869	1.000	0.000	-	-	-	-	-
1	Configural	119	102	0.120	0.999	0.013	-	-	-	-	-
2	Metric	383	111	< 0.001	0.982	0.050	264	9	< 0.001	0.017	0.037
3	Scalar	1305	144	< 0.001	0.925	0.090	922	33	< 0.001	0.057	0.040

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

ESI References

Beauducel A. and Herzberg P. Y., (2006), On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Struct. Equ. Model. A Multidiscip. J.*, **13**(2), 186–203.

Bontempo D. E. and Hofer S. M., (2007), Assessing Factorial Invariance in Cross-Sectional and Longitudinal Studies., in Ong A. D. and van Dulmen M. H. M. (eds.), *Series in positive psychology. oxford handbook of methods in positive psychology*. Oxford University Press, pp. 153–175.

Byrne B. M., (2004), Testing for multigroup invariance using AMOS Graphics: A road less traveled. *Struct. Equ. Model. A Multidiscip. J.*, **11**(2), 272–300.
DiStefano C. and Morgan G. B., (2014), A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Struct. Equ. Model.*, **21**, 425–438.
Finney S. J. and DiStefano C., (2013), Non-normal and categorical data in structural equation modeling., in Hancock G. R. and Mueller R. O. (eds.), *Structural equation modeling: a second course*. Charlotte, NC: Information Age Publishing, pp. 439–492.
Hallquist M. N. and Wiley J. F., (2018), MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Struct. Equ. Model.*, **25**(4), 621–638.
Hancock G. R., Stapleton L. M., and Arnold-Berkovits I., (2009), The tenuousness of invariance tests within multisample covariance and mean structure models., in *Structural equation modeling in educational research: concepts and applications*., pp. 137–174.
Hirschfeld G. and Von Brachel R., (2014), Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal. *Pract. Assessment, Res. Eval.*, **19**(7), 1–11.
Hu L. and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. A Multidiscip. J.*, **6**(1), 1–55.
Komperda R., (2017), Likert-Type Survey Data Analysis with R and RStudio., in Gupta T. (ed.), *Computer-aided data analysis in chemical education research (cadacer): advances and avenues*. Washington, DC: ACS Symposium Series; American Chemical Society, pp. 91–116.
Millsap R. E. and Yun-Tein J., (2004), Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behav. Res.*, **39**(3), 479–515.
Muthén L. K. and Muthén B. O., (2017), Mplus User's Guide, Eighth. Los Angeles, CA: Muthén & Muthén.
Narayanan A., (2012), A review of eight software packages for structural equation modeling. *Am. Stat.*, **66**(2), 129–138.
R Core Team, (2019), *R: A language and environment for statistical computing*, [Computer software].
Revelle W., (2018), *psych: procedures for psychological, psychometric, and personality research*, [Computer software].
Rosseel Y., (2020), The lavaan Project. http://lavaan.ugent.be/
Rosseel Y., (2012), lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.*, **48**(2), 1–36.
Sass D. A., Schmitt T. A., and Marsh H. W., (2014), Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Struct. Equ. Model.*, **21**(2), 167–180.
Schneider W. J., (2019), *simstandard: generate standardized data*, [Computer software].
Svetina D., Rutkowski L., and Rutkowski D., (2019), Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. *Struct. Equ. Model.*, **0**(0), 1–20.
Wei T. and Simko V., (2017), *R package "corrplot": visualization of a correlation matrix,* [Computer software].
Wickham H., (2007), Reshaping Data with the reshape Package. *J. Stat. Softw.*, **21**(12), 1–20.
Wickham H., (2016), ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag.

Wickham H., François R., Henry L., and Müller K., (2019), *dplyr: a grammar of data manipulation,* [Computer software].

Wu H. and Estabrook R., (2016), Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika*, **81**(4), 1014–1045.

Yu C.-Y., (2002), Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.

C.3. Chapter 5: Electronic Supplementary Information

Attitude toward the Subject of Chemistry Inventory version 2 (ASCIv2) presented in figure S5.1. This instrument is the original adaptation by Xu and Lewis in 2011.

ASCIv2

A list of opposing words appears below. Rate how well these words describe your feelings about chemistry. Think carefully and try not to include your feelings toward the chemistry teachers or chemistry courses. For each line, choose a position between the two words that describes exactly how you feel. The middle position is if you are undecided or have no feelings related to the terms on that line.

		Easy			Middle			**Hard**
1.	Chemistry is…	1	2	3	4	5	6	7
		Complicated						**Simple**
2.	Chemistry is…	1	2	3	4	5	6	7
		Confusing						**Clear**
3.	Chemistry is…	1	2	3	4	5	6	7
		Comfortable						**Uncomfortable**
4.	Chemistry is…	1	2	3	4	5	6	7
		Satisfying						**Frustrating**
5.	Chemistry is…	1	2	3	4	5	6	7
		Challenging						**Not challenging**
6.	Chemistry is…	1	2	3	4	5	6	7
		Pleasant						**Unpleasant**
7.	Chemistry is…	1	2	3	4	5	6	7
		Chaotic						**Organized**
8.	Chemistry is…	1	2	3	4	5	6	7

Figure S5.1: Attitude toward the Subject of Chemistry Inventory version 2 (ASCIv2). This is the original instrument developed by Xu & Lewis in 2011 as an adaptation of the original created by Bauer in 2008.

Descriptive Statistics

Table S5.1. Pre-exam 1 ASCIv2 Mean Scores in Organic Chemistry I Fall 2018

	White Female				Hispanic Female			
	Mean	S.D.	Skew.	Kurt.	Mean	S.D.	Skew.	Kurt.
1. Hard-Easy[a]	2.83	1.30	0.36	-0.50	2.85	1.26	0.68	0.82
2. Complicated - Simple	3.05	1.50	0.58	-0.49	2.96	1.44	0.81	0.08
3. Confusing-Clear	3.48	1.45	0.12	-0.77	3.18	1.39	0.32	-0.16
6. Challenging-Not Challenging	2.34	1.17	0.90	0.43	2.11	1.13	1.69[†]	4.56[†]
Intellectual Accessibility	**2.93**	**1.36**	**0.49**	**-0.33**	**2.78**	**1.31**	**0.60**	**0.25**
4. Uncomfortable-Comfortable[a]	3.47	1.46	0.02	-0.61	3.30	1.49	0.15	-0.31
5. Frustrating-Satisfying[a]	3.83	1.78	0.06	-1.02[†]	3.80	1.82	0.07	-0.89
7. Unpleasant-Pleasant[a]	3.84	1.51	-0.07	-0.36	3.63	1.52	-0.24	-0.39
8. Chaotic - Organized	4.23	1.74	-0.33	-0.78	3.94	1.79	-0.07	-0.94
Emotional Satisfaction	**3.84**	**1.62**	**-0.08**	**-0.58**	**3.67**	**1.66**	**-0.02**	**-0.63**

S. D. = Standard deviation. These scores are only for Pre-exam 1 in Organic Chemistry I for White female students (n = 170) and Hispanic female students (n = 84). Each score ranges from 1 to 7, with 4 being the midpoint. High scores mean students feel that chemistry is intellectually accessible or emotionally satisfying. [a]Items 1, 4, 5 and 7 were reverse coded for ease of interpretation. These items appear in reverse on the instrument. [†]Value outside of acceptable range.

Table S5.2. Pre-exam 2 ASCIv2 Mean Scores in Organic Chemistry I Fall 2018

	White Female				Hispanic Female			
	Mean	S.D.	Skew.	Kurt.	Mean	S.D.	Skew.	Kurt.
1. Hard-Easy[a]	2.73	1.25	0.55	-0.11	2.41	1.28	0.74	0.03
2. Complicated - Simple	3.04	1.45	0.67	-0.19	2.90	1.59	0.56	-0.61
3. Confusing-Clear	3.38	1.58	0.37	-0.31	3.14	1.65	0.31	-0.99
6. Challenging-Not Challenging	2.33	1.24	1.31[†]	2.61[†]	1.87	1.06	1.58[†]	3.01[†]
Intellectual Accessibility	**2.87**	**1.38**	**0.53**	**-0.20**	**2.58**	**1.40**	**0.54**	**-0.52**
4. Uncomfortable-Comfortable[a]	3.39	1.48	0.28	-0.73	3.18	1.64	0.10	-1.00
5. Frustrating-Satisfying[a]	3.72	1.77	0.10	-0.90	3.25	1.90	0.49	-0.92
7. Unpleasant-Pleasant[a]	3.56	1.48	-0.09	-0.58	3.26	1.52	-0.23	-1.10[†]
8. Chaotic - Organized	4.08	1.62	-0.18	-0.72	3.91	1.89	-0.13	-1.14[†]
Emotional Satisfaction	**3.69**	**1.59**	**0.03**	**-0.73**	**3.40**	**1.74**	**0.06**	**-0.96**

S. D. = Standard deviation. These scores are only for Pre-exam 2 in Organic Chemistry I for White female students (n = 170) and Hispanic female students (n = 84). Each score ranges from 1 to 7, with 4 being the midpoint. High scores mean students feel that chemistry is intellectually accessible or emotionally satisfying. [a]Items 1, 4, 5 and 7 were reverse coded for ease of interpretation. These items appear in reverse on the instrument. [†]Value outside of acceptable range.

Table S5.3. Pre-exam 3 ASCIv2 Mean Scores in Organic Chemistry I Fall 2018

	White Female				Hispanic Female			
	Mean	S.D.	Skew.	Kurt.	Mean	S.D.	Skew.	Kurt.
1. Hard-Easy[a]	2.42	1.08	0.37	-0.39	2.09	1.11	0.96	0.41
2. Complicated - Simple	3.02	1.57	0.60	-0.66	2.67	1.55	0.50	-0.98
3. Confusing-Clear	3.07	1.55	0.27	-0.88	2.82	1.52	0.43	-1.15[†]
6. Challenging-Not Challenging	2.22	1.28	1.33[†]	1.77[†]	1.92	1.19	1.28[†]	0.83
Intellectual Accessibility	**2.68**	**1.37**	**0.41**	**-0.64**	**2.38**	**1.34**	**0.63**	**0.09**
4. Uncomfortable-Comfortable[a]	3.08	1.55	0.32	-0.90	2.65	1.46	0.46	-0.99
5. Frustrating-Satisfying[a]	3.24	1.79	0.37	-0.85	2.87	1.72	0.50	-0.92
7. Unpleasant-Pleasant[a]	3.22	1.61	0.33	-0.60	2.85	1.62	0.46	-0.74
8. Chaotic - Organized	3.62	1.84	-0.09	-1.24[†]	3.22	1.81	0.24	-1.21[†]
Emotional Satisfaction	**3.29**	**1.70**	**0.23**	**-0.78**	**2.90**	**1.65**	**0.42**	**-0.88**

S. D. = Standard deviation. These scores are only for Pre-exam 3 in Organic Chemistry I for White female students (n = 170) and Hispanic female students (n = 84). Each score ranges from 1 to 7, with 4 being the midpoint. High scores mean students feel that chemistry is intellectually accessible or emotionally satisfying. [a]Items 1, 4, 5 and 7 were reverse coded for ease of interpretation. These items appear in reverse on the instrument. [†]Value outside of acceptable range.

Confirmatory Factor Analysis

The initial CFA without modification did not yield acceptable fit (*i.e.*, Hispanic female Pre-exam 2: χ^2 (n = 84, df = 19, $p < 0.001$) = 51.039; CFI = 0.848; SRMR = 0.077; RMSEA= 0.154 White female Pre-exam 1: χ^2 (n = 170, df = 19, $p < 0.001$) = 53.448; CFI = 0.909; SRMR = 0.059; RMSEA= 0.109. Therefore, as suggested by Wang and Wang (2012) we examined the modification index suggestions provided in the output in both the statistical and theoretical sense. The modification that we chose to examine further is a correlation between error variances of Item 2 (Complicated-Simple) and Item 3 (Confusing-Clear). These items are next to each other chronologically and are part of a cluster of three items that belong to the same factor (Items 1, 2, and 3) and thus perhaps produce a priming effect in this short instrument (Xu, 2010), indicating that the unique variance of these items may be somewhat linked because of the item order (Xu, 2010). In the past, this peculiarity was tested by swapping Item 2 with Item 8 (Chaotic-Organized),

which belongs to a different factor, testing for a priming effect (Xu 2010; Rocabado et al., 2019).

The findings of these investigations yielded characteristic results that revealed a possibility of the priming effect in this instrument, and in particular for these items (Xu 2010; Rocabado et al., 2019). The addition of the correlated error variance between Items 2 and 3 was tested each instance for each group at each time point. All of the results for CFA that we present in this study contain this modification for each group at each time point. Given the consistency of this particular modification to the model, we proceeded to evaluate the model fit for each group at each time point.

Table S5.4. ASCIv2 CFA for Hispanic female students in Organic Chemistry I in Fall 2018

	N	χ^2	*df*	*p*	CFI	SRMR	RMSEA	Omega IA	Omega ES
Pre-Exam 1	84	47.485	18	<0.001	0.903	0.073	0.148	0.807	0.871
Pre-Exam 2	84	34.642	18	0.011	0.921	0.056	0.114	0.796	0.869
Pre-Exam 3	84	23.078	18	0.188	0.980	0.036	0.065	0.866	0.911

IA = Intellectual Accessibility. ES = Emotional Satisfaction.

Table S5.4 shows data-model fit for Hispanic female students only, resulting in acceptable model fit at each of the three time points for this study (Hu and Bentler, 1999). As mentioned previously, the RMSEA displays irregular behavior with short instruments like the ASCIv2, often indicating poor fit (Kenny, Kaniskan, and McCoach 2015). Although the fit for pre-exam 1 is not strong, it is acceptable, and the reliability of each factor is also strong. We see improved fit statistics in pre-exam 2 and 3 together with strong reliabilities for each factor as well.

Table S5.5. ASCIv2 CFA for White female students in Organic Chemistry I in Fall 2018

	N	χ^2	*df*	*p*	CFI	SRMR	RMSEA	Omega IA	Omega ES
Pre-Exam 1	170	44.208	18	<0.001	0.931	0.053	0.098	0.752	0.870
Pre-Exam 2	170	28.818	18	0.051	0.963	0.054	0.064	0.751	0.871
Pre-Exam 3	170	26.406	18	0.091	0.974	0.043	0.061	0.782	0.891

IA = Intellectual Accessibility. ES = Emotional Satisfaction.

In Table S5.5 we show the White female group data, which displays good model fit for each time point throughout the semester. Additionally, reliability values are strong for both factors.

Measurement Invariance Testing

The following tables (S5.6-S5.9) contain the results of measurement invariance testing between Hispanic female and White female students at pre-exam 1, 2 and 3, respectively. Following the steps suggested by Rocabado and colleagues (2020), we performed configural, metric, scalar, and strict invariance tests. Note that the fit indices suggest appropriate data-model fit (Hu and Bentler 1999), as well as the change in fit indices from one model to the next (Chen 2007).

Table S5.6. Measurement Invariance Testing to Support Comparisons Between Hispanic and White Female Students at Time of Pre-Exam 1

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	89.403	36	<0.001	0.922	0.061	-	-	-	-	-
Metric	95.378	42	<0.001	0.922	0.067	5.975	6	0.426	0.000	0.006
Scalar	102.898	48	<0.001	0.920	0.068	7.520	6	0.275	0.002	0.001
Strict	102.913	56	< 0.001	0.932	0.071	0.015	8	>0.999	0.012	0.003

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Hispanic female students (n = 84) and White female students (n = 170). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. The scalar model adds the constraint of equal intercepts for both groups. The strict models adds the constraint of equal error variances. Each constraint is added one at a time. *df* = degrees of freedom.

Table S5.7. Measurement Invariance Testing to Support Comparisons Between Hispanic and White Female Students at Time of Pre-Exam 2

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	63.021	36	0.004	0.946	0.055	-	-	-	-	-
Metric	68.622	42	0.006	0.947	0.066	5.601	6	0.469	0.001	0.011
Scalar	74.733	48	0.008	0.947	0.064	6.111	6	0.411	0.000	0.002
Strict	73.105	56	0.062	0.966	0.072	1.628	8	0.990	0.019	0.008

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Hispanic female students (n = 84) and White female students (n = 170). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. The scalar model adds the constraint of equal intercepts for both groups. The strict model adds the constraint of equal error variances. Each constraint is added one at a time. *df* = degrees of freedom.

Table S5.8. Measurement Invariance Testing to Support Comparisons Between Hispanic and White Female Students at Time of Pre-Exam 3

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	46.985	36	0.104	0.981	0.039	-	-	-	-	-
Metric	50.711	42	0.168	0.985	0.044	3.726	6	0.714	0.004	0.005
Scalar	55.439	48	0.215	0.987	0.046	4.728	6	0.579	0.002	0.002
Strict	56.104	56	0.471	1.000	0.052	0.665	8	0.999	0.013	0.006

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are Hispanic female students (n = 84) and White female students (n = 170). The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. The scalar model adds the constraint of equal intercepts for both groups. The strict model add the constraint of equal error variances. Each constraint is added one at a time. *df* = degrees of freedom.

Additionally, we performed longitudinal measurement invariance testing within the combined groups of students (Hispanic and White female), to support comparisons throughout the semester. The process of conducting longitudinal measurement invariance follows the same steps as previously discussed with between group testing, thus the interpretation of the results follows the same patterns as well.

Table S5.9. Measurement Invariance Testing to Support Longitudinal Comparisons Between Pre exam 1 and Pre-exam 3 for White and Hispanic Female Students

	χ^2	*df*	*p*	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR
Configural	194.435	96	<0.001	0.937	0.052	-	-	-	-	-
Metric	204.844	102	<0.001	0.934	0.060	10.409	6	0.108	0.003	0.008
Scalar	217.655	108	<0.001	0.930	0.060	12.811	6	0.046	0.004	0.000
Strict	219.180	116	<0.001	0.934	0.066	1.525	8	0.992	0.004	0.006

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are White female and Hispanic female students combined (n = 254) for pre-exam 1 and pre-exam 3. The configural model is a comparison model for both groups without constraints. The metric model adds the constraint of equal factor loadings for both groups. The scalar model adds the constraint of equal intercepts for both groups. The strict model adds the constraint of equal error variances. Each constraint is added one at a time. *df* = degrees of freedom.

On Table S5.9, we can see that measurement invariance holds at each level. One note is that the delta Chi-square at the scalar level yields a significant (<0.05) value; however, looking at the other fit indices, we see that the change between metric and scalar models is within the cutoffs prescribed (Chen 2007). This result provides evidence that inferences made from comparing attitude scores across the semester for this group of students can be meaningful.

Meta-Analysis of ASCIv2 Studies With and Without Interventions

Through a systematic search of the literature we have found 6 articles that met our criteria that utilized the ASCIv2 across a semester with at least two observations. Within some of these articles several groups of students were represented and results were obtained separately for group comparisons. Many of these separate groups fell into the categories of intervention, meaning students who experience pedagogical interventions outside of traditional lecture classrooms, and no intervention, meaning traditional classrooms. We performed separate meta-analyses for the different groups experiencing different classrooms since our interest was to examine attitude change at baseline (no intervention). The Tables S5.10 and S5.11 contain the descriptive values used in the meta-analysis. The rows highlighted in gray are the values we used for the no intervention meta-analysis and the rows without background are the values we used in the intervention meta-analysis.

Table S5.10. Meta-Analysis of *Intellectual Accessibility*

#	Article	Group	Mean	S.D.	*n* (T1)	Mean	S.D.	*n* (T2)	Effect Size	Std. Error
1	Mooring et al., 2016	Organic I Flipped Classroom	11.60	4.10	134	13.90	4.60	134	0.53	0.03
		Organic I Traditional	11.40	4.20	160	11.40	4.70	160	0.00	0.01
		Organic II Active Learning	11.80	4.50	57	11.60	4.30	57	-0.05	0.02
		Organic II Traditional	11.50	4.10	81	11.30	3.80	81	-0.05	0.01
2	Brandriet, Ward, and Bretz, 2013	Gen. Chem. POGIL Recitation	2.78	1.25	123	2.95	1.49	89	0.13	0.02
3	Brandriet, Xu, Bretz, and Lewis, 2011	Lab (T1), Discussion (T2)	2.76	1.38	148	2.99	1.41	148	0.16	0.01
		At-Risk	2.77	1.28	87	2.94	1.46	87	0.12	0.02
4	Nenning et al., 2019	Online	13.44	3.31	16	13.19	2.83	16	-0.08	0.03
		Face-2-Face	14.27	3.40	37	14.49	3.55	37	0.06	0.02
5	Stanich et al., 2018	STEM-Dawgs	12.40	3.90	146	12.40	4.00	146	0.00	0.01
		Volunteers	11.90	3.20	55	12.10	4.00	55	0.06	0.02
		Gen. Chem.	13.10	3.90	1489	13.40	4.30	1489	0.07	0.00
6	Vishnumolakala et al., 2017	Semester 1	3.75	0.72	213	4.19	1.06	213	0.49	0.02
		Semester 2	3.45	1.18	67	3.72	0.99	67	0.25	0.03
7	Present Study	Organic Chemistry I								

Note. Shaded cells indicate values for groups considered no-intervention (control).

Table S5.11. Meta-Analysis of *Emotional Satisfaction*

#	Article	Group	Mean	S.D.	*n* (T1)	Mean	S.D.	*n* (T2)	Effect Size	Std. Error
1	Mooring et al., 2016	Organic I Flipped Classroom	15.90	4.50	134	17.40	4.90	134	0.32	0.02
		Organic I Traditional	15.50	4.20	160	15.30	5.30	160	-0.04	0.01
		Organic II Active Learning	16.10	5.60	57	16.40	5.70	57	0.05	0.01
		Organic II Traditional	15.80	5.00	81	15.10	5.10	81	-0.14	0.01
2	Brandriet, Ward, and Bretz, 2013	Gen. Chem. POGIL Recitation	3.91	1.39	123	3.77	1.56	89	-0.10	0.01
3	Brandriet, Xu, Bretz, and Lewis, 2011	Lab (T1), Discussion (T2)	3.82	1.50	148	3.76	1.47	148	-0.04	0.01
		At-Risk	3.94	1.41	87	3.79	1.55	87	-0.10	0.01
4	Nenning et al., 2019	Online	20.56	3.56	16	20.38	4.05	16	-0.05	0.02
		Face-2-Face	20.76	4.20	37	20.95	2.96	37	0.05	0.01
5	Stanich et al., 2018	STEM-Dawgs	17.80	3.60	146	17.80	4.10	146	0.00	0.01
		Volunteers	17.40	2.80	55	16.10	4.00	55	-0.38	0.04
		Gen. Chem.	17.80	4.00	1489	17.50	4.40	1489	-0.07	0.00
6	Vishnumolakala et al., 2017	Semester 1	4.10	0.88	213	4.41	0.98	213	0.33	0.02
		Semester 2	3.87	0.92	67	4.17	0.99	67	0.31	0.03
7	Present Study	Organic Chemistry I								

Note. Shaded cells indicate values for groups considered no-intervention (control).

References

Brandriet A. R., Ward R. M. and Bretz S. L., (2013), Modeling meaningful learning in chemistry using structural equation modeling, *Chem. Educ. Res. Pract.*, **14**, 421-430.

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

Chen F. F., (2007), Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance, *Struct. Equ. Modeling*, **14**(3), 464-504.

Hu L. T. and Bentler P. M., (1999), Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives, *Struct. Equ. Modeling,* **6**(1), 283-292.

Kenny D. A., Kaniskan B. and McCoach D. B., (2015), The Performance of RMSEA in Models with Small Degrees of Freedom, *Sociol. Method. Res*., **44**(3), 486-507.

Mooring S. R., Mitchell C. E. and Burrows, N. L., (2016), Evaluation of a Flipped, Large Enrollment Organic Chemistry Course on Student Attitude and Achievement, *J. Chem. Educ.*, **93**, 1972−1883.

Rocabado G. A., Komperda R., Lewis J. E. and Barbera J., (2020), Addressing diversity and social inclusion through groups comparisons: A primer on measurement invariance testing, *Chem. Educ. Res. Pract.,* **21**, 969-988.

Nenning H. T., Idarraga K. L., Salzer L. D., Blaske-Rechek A. and Theisen R. M., (2020), Comparison of student attitudes and performance in an online and face-to-face inorganic chemistry course, *Chem. Educ. Res. Pract*., **21**, 168-177.

Stanich C. A., Pelch M. A., Theobald E. J. and Freeman S., (2018), A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women, *Chem. Educ. Res. Pract*., **19**, 846-866.

Vishnumolakala V. R., Southam D. C., Treagust D. F., Mocerino M. and Qureshi S. (2017), Students' attitudes, self-efficacy and experiences in a modified process-oriented guided inquiry learning undergraduate chemistry classroom, *Chem. Educ. Res. Pract*., **18**, 340-352.

Wang J. and Wang X., (2012), *Structural Equation Modeling: Applications Using Mplus*. Wiley: Chichester, West Sussex, UK.

Xu X. (2010), Refinement of a Chemistry Attitude Measure for College Students. Dissertation, University of South Florida, Tampa, FL.

Xu X. and Lewis J., (2011), Refinement of a Chemistry Attitude Measure for College Students, *J. Chem. Educ.,* **88**, 561-568.

C.4. Chapter 6: Supplemental Information

Instruments

In this chapter I have presented data collected with two distinct instruments, the ASCI-UE and PC. The ASCI-UE was used in English and Spanish. Figures S6.1-S6.3 display the instruments used in this study.

Figure S6.1: Attitude toward the Subject of Chemistry Inventory – Utility and Emotional (ASCI-UE).

A list of opposing words appears below. Rate how well these words describe your feelings about chemistry. Think carefully and try not to include your feelings toward the chemistry teachers or chemistry courses. For each line, choose a position between the two words that describes exactly how you feel. The middle position is if you are undecided or have no feelings related to the terms on that line.

		Relevant		Middle			**Irrelevant**
9.	Chemistry is…	1	2	3	4	5	6	7
		Depressing					**Exciting**
10.	Chemistry is…	1	2	3	4	5	6	7
		Unnecessary					**Essential**
11.	Chemistry is…	1	2	3	4	5	6	7
		Pleasant					**Unpleasant**
12.	Chemistry is…	1	2	3	4	5	6	7
		Overwhelming					**Manageable**
13.	Chemistry is…	1	2	3	4	5	6	7
		Applicable					**Not Applicable**
14.	Chemistry is…	1	2	3	4	5	6	7
		Satisfying					**Frustrating**
15.	Chemistry is…	1	2	3	4	5	6	7
		Not Important					**Important**
16.	Chemistry is…	1	2	3	4	5	6	7
		Enjoyable					**Dull**
17.	Chemistry is…	1	2	3	4	5	6	7

Figure S6.2: Attitude toward the Subject of Chemistry Inventory – Utility and Emotional (ASCI-UE) Spanish version.

A continuación, se presenta una lista de palabras opestas. Califique qué tan bien estas palabras describen sus sentimientos hacia la química. Piensa cuidadosamente e intenta no incluir tus semtimientos hacia los profesores de química o los cursos de química. Para cada línea, elija una posición entre las dos palabras que describan exactamente cómo se siente. La posición media es si está indeciso o no tiene sentimientos relacionados con los términos de esa línea.

		Relevante			Medio			**Irrelevante**
1.	La química es…	1	2	3	4	5	6	7
		Deprimente						**Emocionante**
2.	La química es…	1	2	3	4	5	6	7
		Innecesaria						**Esencial**
3.	La química es…	1	2	3	4	5	6	7
		Agradable						**Desagradable**
4.	La química es…	1	2	3	4	5	6	7
		Abrumadora						**Manejable**
5.	La química es…	1	2	3	4	5	6	7
		Aplicable						**Inaplicable**
6.	La química es…	1	2	3	4	5	6	7
		Satisfactoria						**Frustrante**
7.	La química es…	1	2	3	4	5	6	7
		Insignificante						**Importante**
8.	La química es…	1	2	3	4	5	6	7
		Divertida						**Aburrida**
9.	La química es…	1	2	3	4	5	6	7

The process of translation of this instrument was carried out in parallel to item generation with the help of an expert linguist in Spanish and English and a native Chilean who was familiar with the language use among Chilean university students. This expert provided insight into the appropriate translations of the adjectives that would be well-understood by the Chilean university students. Their insights were invaluable as the instrument language was finalized in each version. For example we initially thought to use the adjective pair "Familiar-Foreign", but in Spanish the

word “foreign” is more often attached to people who come from a different geographical space, not to concepts or topics that are unfamiliar. Therefore, while this word pair was used in one of the English versions of the instrument, it was quickly removed and replaced with other adjective pairs that could be used in Spanish with Chilean university students.

Figure S6.3. Perceived Competence for Learning Scale

Please respond to each of the following items in terms of how true it is for you with respect to your learning in this course. Use the scale:

1	2	3	4	5	6	7
not at all true			somewhat true			very true

1. I feel confident in my ability to learn this material.
2. I am capable of learning the material in this course.
3. I am able to achieve my goals in this course.
4. I feel able to meet the challenge of performing well in this course.

Process of Instrument Development and Refinement of the ASCI-UE

The ASCI-UE was developed from data collected in cognitive interviews and expert panel review. Details of the interviews and expert panel suggestions will be given later in this document. Following are some important steps in the process of development of the instrument.

At the beginning stages of this project following the cognitive interviews with students in Chile and the U.S. my collaborators and I generated 20 items including the original eight items of the ASCIv2 in hopes to create a four-factor instrument, each factor with five items. The theorized -factor instrument contained two emotional factors (*comfortability* and *emotions*), and two factors associated with cognitive mental processes, namely *intellectual accessibility* and *utility*. The four-factor solution was not acceptable in CFA, and an EFA revealed a three-factor structure instead.

Based on the EFA this three-factor model should work statistically; however, some items were not in line with the theory. Therefore, shuffling them around to match a theoretical basis we kept 15 items and 3 factors and did one more round of expert panel review with the following items, factors. Some of the items switched which word appears first, and some changed one of the adjectives in the pair to a better word. We brought back some items that are more in line with theory and discarded other items that are not so in line with the operationalization of the factors.

Table S6.1. Three-Factor Solution from EFA

MR1	MR2	MR3	Communality	Items
0.897	-0.091	-0.099	0.631	Soothing-Irritating
0.864	-0.030	-0.065	0.651	Infuriating - Calming
0.759	0.214	-0.094	0.610	Exciting-Depressing
0.623	-0.149	0.193	0.617	Confusing-Clear
0.557	-0.168	0.290	0.673	Peaceful – Horrific*
0.521	0.184	0.144	0.585	Enjoyable – Dull*
0.507	0.039	0.231	0.597	Pleasant – Unpleasant*
-0.104	0.909	0.075	0.813	Important - Not Important*
0.248	0.881	-0.236	0.744	Essential - Unnecessary
-0.057	0.715	0.139	0.588	Relevant-Irrelevant*
0.001	0.695	0.172	0.637	Applicable - Not Applicable*
0.072	0.001	0.603	0.639	Manageable – Overwhelming*
-0.080	0.288	0.588	0.670	Understandable-Incomprehensible*
0.180	-0.228	0.527	0.539	Easy-Hard*
0.080	0.128	0.448	0.460	Familiar-Foreign*

*Indicates item was reverse coded for ease of interpretation.
Principal Axis – Oblimin rotation. Rotated solution – ordered. Analysis done in R.

Variances with rotation

	MR1	MR2	MR3	
Proper value	4.50	3.19	1.76	
Proper variance	0.30	0.21	0.12	
Accumulated variance	0.30	0.51	0.63	63% of variance explained

After refining the items, the 15-item instrument was administered in the same courses (OCII and GCII). Yet, based on CFA results the three-factor solution was not a good fit. The emotional and intellectual factors were too highly correlated to the point that the correlation was above 1, rendering a solution that was not possible. Good fit was impossible to attain with this solution; therefore, a two factor solution was explored as per the expert panel's suggestion.

In the summer of 2019 in GCII and OCII courses I piloted a 11-item, two-factor instrument (ES and U) with the following item pairs.

1. Boring -Interesting
2. Relevant-Irrelevant
3. Calming Infuriating
4. Depressing-Exciting
5. Unnecessary-Essential
6. Overwhelming-Manageable
7. Applicable-Not Applicable
8. Pleasant-Unpleasant
9. Satisfying-Frustrating
10. Not Important-Important
11. Enjoyable-Dull

Through iterative refinement based on expert panel suggestions and statistical analysis we refined some of these items, exchanged some of the adjectives, and removed two items permanently resulting in the ASCI-UE instrument presented in this work. The resulting two-factor, nine-item ASCI-UE was then administered in fall 2019 in two sections of OCII and is the data I have presented in this chapter.

Table S6.2. Descriptive Statistics for All Students in OCII - ASCI-UE and PC

		Utility		Emotional Satisfaction		Perceived Confidence	
	N	Mean	S.D	Mean	SD	Mean	SD
Pre Exam 1	291	5.81	1.19	4.00	1.28	5.00	1.29
Pre-Exam 2	285	5.71	1.24	3.88	1.24	4.72	1.40
Pre-Exam 3	262	5.68	1.22	3.79	1.37	4.48	1.49
Pre-Final Exam	249	5.83	1.26	3.92	1.35	4.54	1.54

Table S6.3. Descriptive Statistics for High- and Low-Achievers at Times 2 and 3 for ASCI-UE

	U_2		ES_2		U_3		ES_3	
Achievement	Mean	S.D	Mean	SD	Mean	S.D	Mean	S.D
Low[a]	5.58	1.23	3.69	1.20	5.50	1.32	3.41	1.29
High[b]	5.98	1.14	4.33	1.15	5.94	1.02	4.34	1.29

[a]Low-achievement group 2 (n = 157); 3 (n = 154).
[b]High-achievement group 2 (n =106); 3 (n = 107).

Table S6.4. CFA and Reliability at Times 2 and 3 for ASCI-UE

	N	**χ^2**	***df***	***p***	**CFI**	**SRMR**	**RMSEA**	**Omega U**	**Omega ES**
2	285	91.571	26	<0.001	0.914	0.055	0.094	0.839	0.859
3	262	72.317	26	<0.001	0.934	0.050	0.082	0.780	0.890

U = Utility. ES = Emotional Satisfaction.

Table S6.5. Longitudinal Measurement Invariance for PC

	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural	69.342	19	<0.001	0.955	0.031	0.094	-	-	-	-	-	-
Metric	75.696	22	<0.001	0.952	0.047	0.090	6.354	3	0.096	0.003	0.014	0.004
Scalar	92.947	25	<0.001	0.939	0.060	0.095	17.251	3	0.001	0.013	0.013	0.005

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are all students combined (n = 291) for pre-exam 1 and pre-exam 4. The configural model is a comparison model without constraints. The metric model adds the constraint of equal factor loadings. The scalar model adds the constraint of equal intercepts. The strict model adds the constraint of equal error variances. Each constraint was added one at a time. *df* = degrees of freedom.

Table S6.6. Paired Samples *t*-test for ASCI-UE.

	Mean	Std. Deviation	t	df	Sig. (2-tailed)
U_Pre - U_Post	-0.01255	1.10639	-0.175	238	0.861
ES_Pre - ES_Post	0.13808	1.02335	2.086	238	0.038

After Bonferroni adjustment, no significant difference observed.

Table S6.7. MANOVA of ASCI-UE Between High- and Low-Achievement Groups at the Beginning and End of Semester

Tests of Between-Subjects Effects

Source	Construct	Type III Sum of Squares	*df*	Mean Square	F	Sig.
Corrected Model	U_Pre	.098	1	0.098	0.068	0.794
	ES_Pre	19.968	1	19.968	14.342	0.000
	U_Post	17.632	1	17.632	11.276	0.001
	ES_Post	73.720	1	73.720	49.223	0.000
Intercept	U_Pre	7935.961	1	7935.961	5560.946	0.000
	ES_Pre	3938.872	1	3938.872	2829.000	0.000
	U_Post	8072.152	1	8072.152	5162.309	0.000
	ES_Post	3751.785	1	3751.785	2505.109	0.000
Achievement	U_Pre	0.098	1	0.098	0.068	0.794
	ES_Pre	19.968	1	19.968	14.342	0.000
	U_Post	17.632	1	17.632	11.276	0.001
	ES_Post	73.720	1	73.720	49.223	0.000
Error	U_Pre	338.220	237	1.427		
	ES_Pre	329.980	237	1.392		
	U_Post	370.590	237	1.564		
	ES_Post	354.944	237	1.498		
Total	U_Pre	8439.875	239			
	ES_Pre	4291.640	239			
	U_Post	8524.750	239			
	ES_Post	4106.880	239			
Corrected Total	U_Pre	338.317	238			
	ES_Pre	349.948	238			
	U_Post	388.222	238			
	ES_Post	428.663	238			

Measurement invariance testing between high- and low-achievement groups conducted for PC. Results indicate that comparisons are not supported between these groups for PC at any time point.

Table S6.8. Measurement Invariance Testing for PC for High- and Low-Achievers at the Beginning of the Semester

	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural	21.520	4	<0.001	0.967	0.027	0.183	-	-	-	-	-	-
Metric	30.053	7	<0.001	0.956	0.096	0.159	8.533	3	0.036	0.011	0.069	0.024

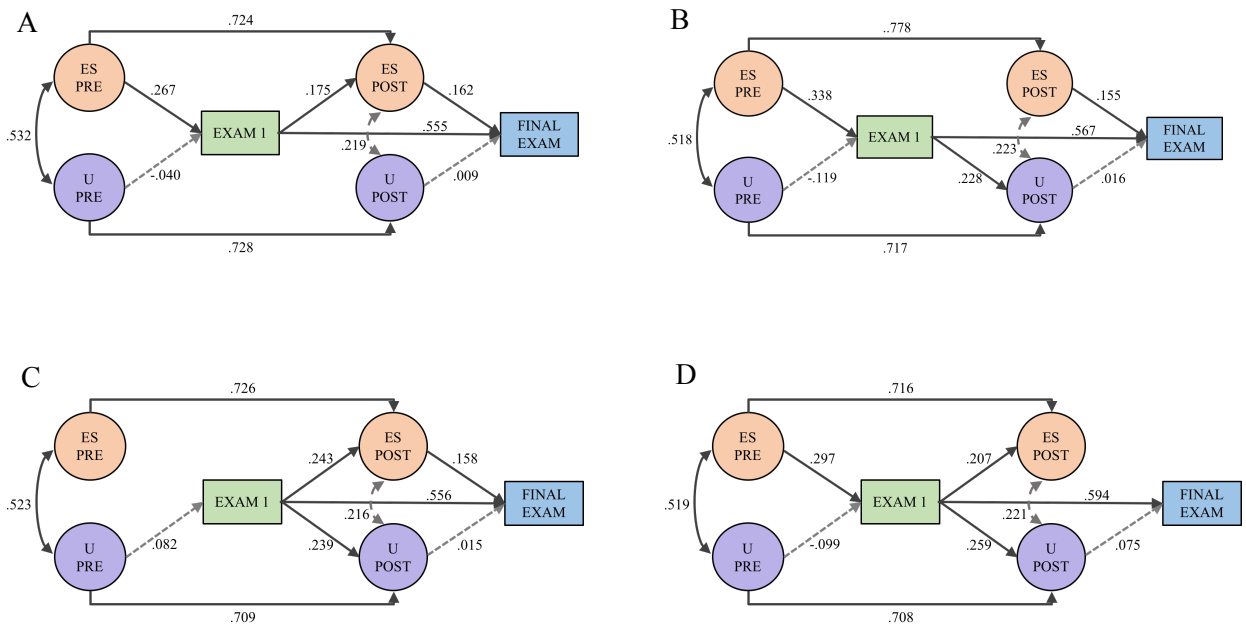
Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups high-achievers (n = 105) and low achievers (n = 157) for pre-exam 1. The configural model is a comparison model without constraints. The metric model adds the constraint of equal factor loadings. The scalar model adds the constraint of equal intercepts. The strict model adds the constraint of equal error variances. Each constraint was added one at a time. *df* = degrees of freedom.

Table S6.9. Measurement Invariance Testing for PC for High- and Low-Achievers at the End of the Semester

	χ^2	*df*	*p*	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔSRMR	ΔRMSEA
Configural	36.946	4	<0.001	0.895	0.038	0.257	-	-	-	-	-	-

Model fit statistics using maximum likelihood robust (MLR) estimator. Note that the comparison groups are high achievers (n = 106) and low achievers (n= 143) for pre-exam 4. The configural model is a comparison model without constraints. The metric model adds the constraint of equal factor loadings. The scalar model adds the constraint of equal intercepts. The strict model adds the constraint of equal error variances. Each constraint was added one at a time. *df* = degrees of freedom.

Figure S6.4. SEM Reciprocal Causation Nested Models



Cognitive Interview Data for ASCIv2 (*n* = 11)

Item 1: Easy-Hard

Students thought this set of items belonged in the *Intellectual Accessibility* (IA) factor because the adjective elicit a cognitive mental process about content understanding. Here are a few quotes:

"I think that there are, there are like patterns ..." – Student 1

"There is a lot of rules to remember and that can make it difficult and there's a lot of equations." – Student 2

"Like you don't really have to, like kind of think twice about it. So it just kind of like, it registers with you immediately rather than like, hard. It would be kind, kind of like you hear it once within, you're still kind of like, I still don't understand what that means." – Student 3

"Because from what I've gone through, it's a very fast-paced subject and there's a lot of memorization involved juggling that against other courses as well. It's quite difficult" – Student 4

"So understanding chemistry is not the simplest task. When I think of easy, I think have the least amount of work necessary to complete something and the ability to understand something immediately without complex thought or higher level like, um, reflection, chemistry in my aspect, I would not consider it easy because you have to be able to visualize and understand a lot more than you would normally if you look at something because some things take a lot more factors to be able to understand the base or something." – Student 5

"Like it's, it's explained to you, it's kind of given to you in like bite size pieces. Everything adds up to the next thing that you do and like things just like keep, like adding onto one another so you learn the rest of the material and like if you don't get it you're going to fail the rest. So it's like you better learn it now before it comes at you later or it goes the same way, but it's not given to you and like the small bite size pieces." – Student 6

"It's easy because I can understand it. Yeah. Like it's, it makes sense to me. Okay. So for the most part, yeah, it makes sense to me. I think that's logical. Like I can work out the problems. So that's why it's easy to me." – Student 7

Item 2: Complicated-Simple

Students thought this item also belonged to the IA factor. Here are a few quotes:

"So complicated because like you learn all these things and they're like here's this thing and it's always this way except for these five exceptions or like, and then they test you on the exceptions and you're like, what?" – Student 1

"How to figure out the polarity of stuff. Lewis structures. It's a little complicated." – Student 2

"So I would say like complicated would be like, there's just like a lot of information, there's a lot of steps and simple would just be like, it's just kind of, like straightforward" – Student 3

"I think that it's complicated, but if you break it apart, it's quite simple. Really. Like chemistry, if I'm working on a problem and like it's a paragraph long, it seems really complicated, but if I just single out like what I need to find in what I'm given, the problem becomes quite simple and then I can solve for what I need sometimes like problems have extra info that you don't need and just being able to see what you actually need to solve for is a good tactic" – Student 4

"So complexity to me just means a lot of things are going on and you have to be able to generally organize and separate what those things are as in if there's a multitude of steps happening at the same time, you have to understand how those relate to each other." – Student 5

"Do I find certain things simple or not? Yes. Like when we're talking about like bond types, like when molecules have like ionic bonds and covalent bonds, like okay, that's pretty simple, but like objectively, the theories behind it and why it works that way. Not Simple. It's a complicated thing because everything is just, it's lots of things happening at once and it's like, okay, well you're looking at the chart of like covalent bonds and like into ionic and you're looking at like the electronegativity differences. Like there was a lot that goes into determining what the bond is. Right? Some of you are like on the cusp, there's just a lot. It's a lot of mental work to figure some of those out so I wouldn't like objectively call chemistry simple" – Student 6

"I think the math aspect of chemistry to me that simple because I'm good at math, but like nomenclature and mechanisms, those are more complicated and complex because there's a lot of things to consider versus with math I just matched up units and then I figured it out." – Student 7

Item 3: Confusing-Clear

Students thought this item also belonged to the IA factor. They used many words that could potentially be interpreted as emotional or affective; however, their overall description more closely matched cognitive processes. Here are a few quotes.

"I think like there are certain rules like for everything, even for the exceptions, like when you add this to this, it adds in this, it has the stereochemistry, it has this regio-selectivity and that's it. Like always. And then if there's an exception, this is how the exception goes. Like I think it's very clear. I think that it can get confusing, like with like in the way that it's taught." – Student 1

"It's a little confusing…. It's just like all of the different rules and stuff. It's mostly just like all the different rules" – Student 2

"I would say like, it's just, you don't really have to think more about it, like kind of like, it's just there, like you don't have to decipher anymore and then confusing is, you might not get it the first time. It might not register with you right away." – Student 3

"I think it's pretty clear. The only time it's confusing as if like I'm not paying attention and I don't like I miss pieces of info. That's only when it's confusing, but it's all laid out. And all the laws and the rules are written and the formulas are like given you just need to know how and when to apply them and like look in the text to see like what requires what. I think it's pretty good" – Student 4

"So I would say chemistry itself isn't confusing, but I think the math necessary to complete it can be confusing because of the amount of things that do and don't follow rules. There are so many exceptions and so many different smaller steps you have to do just to be able to complete something so that you have the base of knowledge that you can get tripped up in the aspect of you can lose yourself in what you're doing" – Student 5

"That's more of like something that like I feel like I don't understand something, so I find it confusing, but I could talk to somebody else and be like, no, okay, that's not confusing. Like it's pretty easy. Here's the explanation and I'd be like, Oh, I'm less confused" – Student 6

"I think when you're overwhelmed or when there's a lot to think about, not necessarily like working out the problems. I feel like when you're overwhelmed everything is confusing and nothing makes sense." – Student 7

Item 4: Comfortable-Uncomfortable

Students thought for the most part that this item belonged to the *Emotional Satisfaction* (ES) factor; however there were a couple of students that could also link this item to cognitive processes (Students 1, 3, and 5). Yet, in these instances, the students described their feeling when something was not clear. Therefore, I concluded that this item belongs in the ES factor based on students' descriptions. Some students (5 and 6) thought that this item was not a good description or attitude toward chemistry. Here are a few quotes.

"It's uncomfortable because it challenges you to think in 3D especially in orgo." – Student 1

"I think of something that I am good at, something that I will be able to, like if you gave me a problem I'd be able to solve it without any help. And then uncomfortable would be like, you don't really know what you're doing. You might need someone to like help you, guide you to the answer" – Student 3

"Yeah, there's a lot of things like I don't know and sometimes I don't know how to do something. I'll just feel like very powerless" – Student 4

"Uncomfortable in this sense is kind of confusing in the fact that it, it would make you, I guess in an academic sense, it kind of degrades where you, what you feel about yourself moving forward

in terms of knowledge, but in terms of chemistry itself, as a discipline being uncomfortable? ... It creates insecurity versus security" – Student 5

Item 5: Satisfying-Frustrating

All the students thought this item belonged in the ES factor, although some had some reservation about the adjective "frustrating." Yet, they all described affective processes when talking about these two adjectives. Here are a few quotes.

"It's like a rollercoaster, like you could be like in the satisfied I'm satisfied phase for like one concept, like for NMR spectroscopy or whatever. And then you could be in the frustrated stage of like diels alder" – Student 1

"I would say it is kind of nice to see everything, like go through all of the equations and then get their right answer and being able to help other people. And that was pretty nice too" – Student 2

"So we have like that peer-leading thing in class and, so like for example, like some of the problems might be kind of difficult and you can't feel like you keep coming up with an answer and then the peer leader person would like check it and like it would be wrong and then they took it again. It's so wrong. So like that might be frustrating. But then satisfying would be like, you do the problem, you get an answer and then they say that's right. So then you're satisfied. Yeah. You're like, yes, I did something great." – Student 3

"I think it's pretty satisfying, like being able to solve problems, being able to know like how things work. It's pretty satisfying. Just like being able to think in terms of chemistry is pretty fun." – Student 4

"So satisfaction, a lot of times comes just from the work that you put into something. So if you put a lot of work into something and you get the desired result, then you're usually satisfied with what

you do. Also just satisfaction comes from the importance you put on something. So if something goes wrong but you didn't care so much about where you wanted it to be, it's more like goal setting and the aspect that if you set a goal and you can reach that goal or go above and beyond it, you're happy with what you go on if otherwise you're not." – Student 5

"I feel like when you figure out orgo it is like really satisfying and like when you figure out like all those stoichiometry problems and like the gas law problems in Gen chem like that's satisfying because you know what you're doing. Yeah. And then you get the next one wrong. You just want to cry." – Student 6

"Some people could understand chemistry and still not find it satisfying, like get chemistry. It's just annoying." – Student 8

Item 6: Challenging-Not Challenging

This item was described in terms of cognitive processes for most students. However, there were some students (4 and 6) that also described an affective aspect for this item. It was also very clear that students viewed "challenge" as both a positive and negative characteristic of chemistry in the sense that chemistry encourages effort and growth. These two reasons are evidence for the idiosyncratic behavior of this item in some studies including Chapter 3 in this work. Here are some illustrative quotes.

"I think chemistry is challenging for sure. Know what's it like in the fact that it causes you to think differently than you have had to before. Yeah and like you have to, I think you have to work at it more than you do for other subjects." – Student 1

"It's a little challenging. I do have to put in work because it doesn't come to me. It's not too bad because it really isn't. If I just study, which you're supposed to study anyway, if I just studied, it's not too bad." – Student 2

"So challenging I would say like makes you think like you have to work for it kind of and then not challenging would just be like, it just, you just immediately now how to do it. You don't have any problem with it" – Student 3

"Going back to like what I said before, there's a lot to learn and there's a lot to take in and applying all the concepts and using them for the exams. It's quite overwhelming. Also just learning new concepts in general. It's kind of difficult and being able to fully understand and comprehend. It's tough" – Student 4

"So challenging to me is when something can feel difficult and how much effort you put into it, but it gives you the ability to actually rise to that level and you have the actual ability to get what you want out of it. If you put in effort, if you put in focus, if you put it in time, you will be able to do whatever you want moving forward. The difference I guess, between challenging and hard is that you have that ability and challenging while I was in hard, it's not really about how much you put in. It's about the material itself and the ability to visualize or conceptualize what they're talking about." – Student 5

"This one is objective. To me, challenging is an objective word because it's like complex, challenging and complex kind of fall in the same. Like if were to have like a Venn diagram of these words, like they fall in the same category. I'm like, chemistry is objectively challenging, like it's not, it's not for the faint of heart let me tell you." – Student 6

"So I think chemistry is challenging because of how like how much work you have to put into it and how much you have to know. It's challenging because it requires a lot of concepts. Connection." – Student 7

"I feel like when I think challenging, I think that critical thinking and stuff like that and like easy or hard something can be hard but it's not really like critical thinking wise. Like if something is

hard, like you could be like trying to, I don't know, push a heavy object that's hard, but like it's not really challenged. I feel like it's not as challenging" – Student 8

Item 7: Pleasant-Unpleasant

All students described this item belonging to the ES factor. They described the feeling after succeeding in a task. Students 3 and 5 also connected this item to the applicability or usefulness of the discipline in one's life. Here are some quotes.

"It's not unpleasant. It's just kind of boring to me. ... It's a little bit better than just total boring but not my thing" – Student 2

"I don't hate it." – Student 4

"Something unpleasant, at least in student aspect, is something that you are actively dedicating time and energy into that you don't see in your life in any way or see yourself using because you feel more so is if you're kind of wasting your time in a way. And the aspect of if you're learning something it shouldn't be of use, but if you have no use for it, then you're just kind of taking up space as well as taking up time." – Student 5

"Is it pleasant to get something right? In chemistry ... absolutely. I'm very pleased when that happens." – Student 6

"It's how I would feel after like understanding something or after working out something." – Student 7

"You can understand it and still think chemistry is unpleasant or you could, you know, hate chemistry and be like, it's still cool though, like pleasing when you get it when you get it right. Yeah. It's so pleasing when you get it right." – Student 8

Item 8: Chaotic-Organized

Without hesitation all students described cognitive processes for this item, although the factor structure indicates that this item belong to the affective ES factor. Additionally, students linked this item to how chemistry is presented in the class or the complicated processes to solve problems. Some students also described "chaotic" as somewhat of an affective process when they are confused by the steps, yet, these descriptions were also accompanied by descriptions that matched cognitive processes more closely. This conflated result provides evidence of the idiosyncratic behavior observed with this item including in Chapter 3 of this work. Here are some illustrative quotes.

"I think it's very organized. Like in the, like these are the steps for this. I guess like if you're confused it can be chaotic, it can be like I don't get it, what's going on? Like, but I don't really think it's chaotic. Okay. Yeah, I think that there's like structure to chemistry." – Student 1

"And it's like just having rules in general is very orderly and organized." – Student 2

"So organized would be like everything has like a certain, like step and then chaotic would be like, there's no instructions is kind of just like, you have to think of them on your own" – Student 3

"You have all the rules laid out and you know what applies to what and why things work like that because you have theories and laws backing those up and you have examples, too. I wouldn't say it's chaotic because just because I don't view it as all over the place." – Student 4

"Chemistry as a discipline in and of itself is like mostly organized, right? Because it gets very research-based. It's very fact-based like, Hey, this is what happens, here's why. So yeah, it's organized. It's never, if you find chemistry to be chaotic, you're not learning it, right? So you need to go back and you re-read the book or something or talk to your professor." – Student 6

"It deals with how it's like complex. And was that the one challenging? Well, not really challenging. It's more like how it's complex because there's a lot that you have to do and sometimes like it can feel like a whole bunch of chaos." – Student 7

"If you present it in a poor way then it could be chaotic. Like if you like didn't have sections, they're like you didn't like keep it together. Like if you threw things that weren't in the same section into each other, then it can be chaotic. So I guess it depends on like how it's presented." – Student 8

One student also made the explicit connection between items 2 and 3 that has been observed consistently throughout my work with this instrument. "Because if you're confused, you're feeling confused. So and usually kinda ties hand in hand with if it's complicated or as simple as well, like if it's complicated or as simple as well, like if it's complicated it's probably gonna be more confused. So those, these two pretty tied together." – Student 8

Other Potential Adjectives to Consider

Fun

Some students described chemistry as "fun" and posed this adjective as an item that could belong in an affective factor. However, the students also linked this adjective to other adjectives such as "interesting." It seemed as though chemistry can be fun only when one finds it interesting, or only when one understands it, or only when one likes the subject. Therefore, this adjective in and of itself might not be an evaluative judgement on its own. One student (5) equated this item to "engaging." Here are few quotes.

"It's like fun going through the learning process of like where you go from not understanding to understanding" – Student 1

"I like to learn like new things, so like new things are fun to me, but it might not be fun, like everyone might not think that chemistry is fun, but because I like science because it's something that I like to learn about, then I would say that it's fun." – Student 3

"It's just fun to learn chemistry and like, I like my classes, they're very interesting and just like I'm always like wide awake and so I like learning about it and I think it's really fun to understand more about the world." – Student 4

"If it's engaging then it can be fun. The engagement and the ability to do things outside of strictly mental perspective in my opinion is what makes things fun." – Student 5

Enjoyable

This adjective was linked to an affective mental process as well. Although this adjective is similar to "fun" some student's made an explicit distinction between these adjectives and concluded that this adjective can stand on its own or be linked to other processes like "understanding." This item became a new item in the ES scale of the ASCI-UE. Here are a few quotes.

"I remember I had chem lab today and we put two solutions together and they are both clear I think I'm pretty sure that they mixed them together. There was an obvious reaction that was pretty cool. So like that kind of stuff, like seeing, like not just measuring out stuff and trying to make a or seeing how much calcium is in the thing, but like seeing the different reactions is really cool." – Student 2

"The more you learn and if you understand it then it's more enjoyable." – Student 4

"I feel like it's more of like a 'do you find this type of thing enjoyable?' Because I could also do stoichiometry all day, but if you asked me to do like, oh, it was ozonolosis mechanisms all day, I'd probably just walk out of the room. I'd be like, no, screw that. I don't even, I don't even want the compensation. Just leave me alone. Yeah. That's how I'd feel about that." – Student 6

Relevant

Many students talked about this adjective as an important way to view or judge chemistry. All students discussed this item in terms of usefulness in their lives in a broad/global or individual way. This is an item in the *Utility* (U) scale of the ASCI-UE described in this chapter. Here are a few quotes.

"I guess I'm like, that is chemistry's important because if I didn't have chemistry I wouldn't really understand that stuff or if I didn't have like a basic knowledge of chemistry, I wouldn't get that stuff." – Student 1

"Relevant, that kind of ties in with the usefulness because I'm like, it really just depends on what your career, what your interests are. Like it's not really relevant or useful to like a writer or an artist or maybe an engineer, but that's still kind of science. Like it's not really useful for anyone out of STEM. It's not really relevant, like they're not going to use it" – Student 2

"I think [chemistry] is definitely relevant to almost everything because chemistry is just, that is the world , like chemistry is based off of the natural world." – Student 4

"So relevance is just kind of how you apply to your life, but also what's occurring are like not just in your life but in the world." – Student 5

"If you're taking the course, you need to find it relevant. So it's kind of like the other thing where it's like the organized chaotic. If you are presented with something that feels irrelevant to you, then

obviously you're not going to be interested in it. You're going to want to learn it. I'm like I said before, like chemistry is one of those things like you need to learn, like you need to want to learn it. You need to learn how to want to learn it." – Student 6

"Definitely useful. Definitely relevant with everything really. I'm literally quite literally and more than like the common definition of everything, you know. So definitely useful for me. Especially going into the medical field. Definitely relevant. There's like we were saying about the antibiotic thing that's relevant. We need to figure that [explicit word] out because we're gonna die soon. It's going to be at like a mega bacteria. We're all gonna, there's going to be like a bacteria that nothing will ever kill it ever. And then we're going to be extinct. And you ignored chemistry." – Student 10

Interesting

Many students chose this adjective to describe chemistry. Most students explained this adjective in terms of the utility of chemistry; however, some (2, 3, and 6) also described this item in terms of affective or cognitive processes. Because of the conflated descriptions, this item did not make the cut for the new instrument, although most students gave this description. Here are a few quotes.

"I think it's really interesting. Yeah, I love it. I really, really enjoyed connecting the dots from one course to another. And then just like learning, like how things work, like in the very beginning when you learn that like water is polar and oil is not polar, like you see that in your everyday life. Like why does this part of my salad dressing always sink to the bottom, like I always have to shake it, you know what I mean? So I think it's cool learning the science behind those things even if you don't. And I guess that goes along with relevant, like even if you don't think like, oh I'll never need times in my life but I want to know why these two don't mix." – Student 1

"Ties into enjoyable. So feeling like if I'll enjoy things that are enjoyable sometimes, like, like learning about how much food dyes and Gatorade is kind of interesting, but it's not really fun." – Student 2

"I mean, I guess it depends, like it depends on like what you find interesting. So like one person might not find that to be interesting. That might be like, oh, that's dumb, but if they liked chemistry maybe. And if they're good at chemistry, chemistry comes easy to them, then they'll think that it's interesting because it's something that they're good at, but like if it's hard for you then you're like, you're automatically, I'm just going to hate it and you won't try to like it or find it interesting." – Student 3

"Just learning about it how things work and why certain molecules do things and applying them to larger scales, like with fire, or if you light something on fire it burns and why does it? Like, what's the reaction?" – Student 4

"So I find it interesting. It doesn't make it easier. So I think objectively chemistry is interesting whether or not you understand it, but it's interesting regardless. It's like art history is also interesting whether or not you like we'll pursue that if there's another story. But it's also a feeling like, am I interested in finishing this mechanism? Am I interested in pursuing that as a group, are interested in learning this reaction? So that is a feeling that's more of like a, no, I don't really want to do it, but like I know I have to so I shouldn't find some interest in it." – Student 6

Overwhelming

Many students talked about how they feel when they think of chemistry, and "overwhelming" was a common way to describe their feelings toward the discipline. This adjective

was described as an affective process and I added it to the ES scale in the ASCI-UE because of its prevalence in students' evaluative judgements. Here are a few quotes.

"I think [chemistry can be overwhelming]. Yeah, but I think I might be saying that like in the context of my own life, like where I want to go to office hours, I want to do all the extra credit, I want to do every single homework problem like correctly and like go through the answers and like walked through everything but like in context with like my other like seven classes that I have and um, like it's just like, uh, where can I find the time for that even though it's something that I want to do." – Student 1

"[Overwhelming], I would just say like the all like trying to learn all the information, make sure you know how to do all of it, be able to apply it. All of the assignments. Tests. It's just like I'm packing all the information into your head and knowing it, so not just like listening, but also like being able to understand and apply it." – Student 3

Applicable

Many students spoke of the utility of the discipline and gave this adjective to describe what they meant. It was apparent that students found significance in learning a subject that they could apply in their lives, future careers, and could see its utility in a global perspective. This item was added to the U factor of the new ASCI-UE. Here are a few quotes.

"Applicable, that really does tie into useful and relevant. Okay. Because if something's applicable, it's useful for me." – Student 2

"I guess would that like kind of coincide with the relevant you'd be able to like apply that in your life." – Student 3

"Kind of goes with relevant in the sense that like I was talking about with the fire, it can apply what you learned in chemistry, too. Like with Global warming and stuff that you can understand like how the gases affect the earth in the atmosphere." – Student 4

"Applicability, I guess it's genuinely understanding something and then its use so utility once more, but like it has to do with the person and the situation itself because if the person has that in their life or all those other factors line up for them, it could be ethical. That's why things are more applicable to some groups than others and that's why I guess for marketing or any other form of advertisement, they do like those focus groups in those things because they had to figure out what applies to everybody. It's based off the person and the factors that affect them." – Student 5

"It's the application part happens when you actually see it happening in labs. Okay. Um, okay, that's cool." – Student 9

Boring

Some students talked about chemistry being "boring." This item was explored in the pilot studies, however, after review of the pilot data and with the expert panel it was decided that the word "dull" better described the opposing adjective to "enjoyable." Here are a few quotes.

"It's not my thing, it really isn't, but I can see where it be other people's things, like I've said, yeah, it's, it's, it's kind of tedious at times it feels like because like, especially just to do practices, you have to do the same thing over and over and over again. That's kind of boring." – Student 2

"Like I do sometimes find myself just like sitting there and like spacing out and then like whenever like pay attention, like I just missed a huge portion so then they ended up having to go back and reteach myself, you know." – Student 3

Encouraging – Discouraging

This item came from a student who described the ambivalence of certain tasks in chemistry showing that students can feel both of these opposite feeling within the discipline. This item did not make the final cut of the new instrument because based on the students' descriptions and the expert panel, the value of this item is tied explicitly to task in the course rather than the discipline. Here quotes from one of the students.

"Reiterating the fact that like if you don't find it interesting then or maybe it's, it can be a little bit like discouraging, like on my part. Like I would say like, like how I, for example, like I did bad on the test and then this weekend during class, like I have just not been paying attention because I was just discouraged" – Student 3

"Again with like the peer leading thing, I think that can be a very encouraging thing because I'm like, my peer leader will be like, oh, like if you understand this, then explain it to your partner. I think it's encouraging, like in that sense" - Student 3

One student also made the connection between some of the items that eventually made the cut to be included in the U factor of the new ASCI-UE. *"I think it's fun and interesting. Yeah, I think chemistry is interesting and fun and useful. Like the real world application of like chemistry, like in the chemistry labs for example, like when you do like your, like we did water hardness testing and our last lab and like roommate, we made kidney stones in the lab if we just did that. So we're about to try and figure out how to dissolve them, like they will help you. It's useful and like other than just like a classroom, this is what happens. ...chemistry's really important in a lot of aspects. Like if you're a doctor, it's important in pharmaceuticals and stuff and like all the medicines that you're going to be prescribing, you're going to have to understand what you're*

giving to your patients. So I was, I think you should understand it. ...I think chemistry is fun just because it's interesting to learn about how things were kind of molecular level and like I guess that ties in with how it's interesting just because like you find out how things connect. Like I said, I like things to connect, like that's like most satisfying thing is when things connect and so I think it's fun and interesting when you learn new things that can help you like connect things together and it's useful because it can help you, like in your field of study. Like for me, like it'll help with pharmaceuticals and if I wanted to do research it could also help me if I like, wanted to try and create my own drug or something like that. So I'm into use. Yeah, it's useful." – Student 8

APPENDIX D

INSTITUTIONAL REVIEW BOARD APPROVALS

D.1. Pro00020840

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

2/4/2015

Li Ye, M.S.
USF CITRUS - Center for the Improvement of Teaching and Research in Undergraduate STEM Education
4202 East Fowler Ave., CHE205
Tampa, FL 33620

RE: **Expedited Approval for Initial Review**
IRB#: Pro00020840
Title: Investigating Evidence for the Validity of Chemistry Assessments Methods

Study Approval Period: 2/4/2015 to 2/4/2016

Dear Ms. Ye:

On 2/4/2015, the Institutional Review Board (IRB) reviewed and **APPROVED** the above application and all documents outlined below.

Approved Item(s):
Protocol Document(s):
IRB Protocol.pdf

Consent/Assent Document(s)*:

Informed Consent B.docx.pdf

*Please use only the official IRB stamped informed consent/assent document(s) found under the "Attachments" tab. Please note, these consent/assent document(s) are only valid during the approval period indicated at the top of the form(s).

It was the determination of the IRB that your study qualified for expedited review which includes activities that (1) present no more than minimal risk to human subjects, and (2) involve only procedures listed in one or more of the categories outlined below. The IRB may review research through the expedited review procedure authorized by 45CFR46.110 and 21 CFR 56.110. The research proposed in this study is categorized under the following expedited review category:

(6) Collection of data from voice, video, digital, or image recordings made for research purposes.

(7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

As the principal investigator of this study, it is your responsibility to conduct this study in accordance with IRB policies and procedures and as approved by the IRB. Any changes to the approved research must be submitted to the IRB for review and approval by an amendment.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

Kristen Salomon, Ph.D., Vice Chairperson
USF Institutional Review Board