Chemistry Education Research and Practice

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/cerp

Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry

Brent Ferrell and Jack Barbera

Department of Chemistry and Biochemistry, University of Northern Colorado, Greeley, CO 80639
Email: jack.barbera@unco.edu
Key Words: motivation, self-efficacy, interest, effort beliefs, confirmatory factor analysis

Abstract

Research in academic motivation has highlighted a number of salient constructs that are predictive of positive learning strategies and academic success. Most of this research has centered on college-level social sciences or secondary school student populations. The main purpose of this study was to adapt existing measures of personal interest and effort beliefs to a college chemistry context. In addition, a chemistry-specific measure of self-efficacy was evaluated in a modified form. This set of scales was initially administered at two time points in a first-semester general chemistry course to a sample of undergraduates (n_1 = 373, n_2 = 294). Confirmatory factor analyses (CFA) were conducted to determine whether the scales were functional in a chemistry context. Following revision of the scales, all CFA models demonstrated acceptable fit to the data. Cross-validation of the revised scales was performed using two different populations (n = 432, n = 728), with both studies producing similar model fits. Furthermore, our data shows that chemistry majors reported higher self-efficacy and interest than non-science majors. Cronbach's alpha estimates ranged from 0.75 to 0.92 for the revised scales across all studies. This set of scales could provide useful tools for assessing general chemistry students' motivation and the motivational impacts of various teaching practices.

Introduction

Introductory chemistry is a course required by several disciplines. At most institutions, enrollment overwhelmingly consists of students outside of the discipline of chemistry. Often, many students struggle through chemistry and are unsuccessful due to the complexity and abstract nature of the content (Nakhleh, 1992). The combination of content difficulty and the fact that most students are fulfilling a credit requirement for their non-chemistry majors generates an interesting classroom environment for the introductory-level chemistry course. Many factors can influence whether a student is successful in chemistry. There are some obvious characteristics of students including: inherent aptitude and prior experience in chemistry, which can be predictive of success in chemistry (Tai *et al.*, 2005). However, research has shown that cognitive factors such as these are not sufficient to predict achievement, but must be augmented by adaptive motivational processes (Dweck, 1986; McCoach and Siegle, 2003; Zusho *et al.*, 2003).

The importance of motivation for learning and achievement in any classroom setting is indisputable (Dweck, 1986; Schunk, 1991; Ames, 1992; Hidi and Harackiewicz, 2000; Singh *et al.*, 2002; Zusho *et al.*, 2003). Student motivation has been described as a theoretical construct that can explain "the degree to which students invest attention and effort in various pursuits" (Brophy, 2010). Motivation and ability are the two major components for academic success among students (Hidi and Harackiewicz, 2000). Despite this, students' motivation can easily be ignored due to its complexity, or oversimplified as an unchanging facet of one's character. Motivation in the classroom is very complex, and can fluctuate in different situations (Nicholls *et al.*, 1989; Pintrich, 2003) and among different subjects (Guay *et al.*, 2008). Students' motivation toward school tends to become highly differentiated throughout grade school, as individuals encounter various situations and experiences that shape their interests and conceptions of ability. As such, it is important for researchers to study motivation in specific contexts (Pintrich, 2003).

Although much research has been conducted to understand the motivational and affective factors that influence performance and student engagement in the college

classroom, only a small fraction of this has been directed towards science classrooms. In 2012, the National Research Council called for a collective effort on the part of science education experts across many disciplines to put together a report that would highlight current research areas in education, as well as areas that are lacking across these disciplines. As pointed out in this Discipline-Based Education Research (DBER) report, students' dispositions and motivations to learn science and engineering are largely understudied and are of "central importance" (National Research Council, 2012). Nevertheless, many researchers have explored specific motivational constructs and processes in the context of college chemistry both prior to and since this report (Dalgety *et al.*, 2003; Zusho *et al.*, 2003; Bauer, 2005; Dalgety and Coll, 2006; Taasoobshirazi and Glynn, 2009; Uzuntiryaki and Aydin, 2009; Xu *et al.*, 2013; Villafane *et al.*, 2014).

Our aim in this study is to expand upon the current body of research directed at understanding motivational processes among students in college chemistry. In particular, we were interested in modifying existing measures of motivational processes for use in college chemistry. We chose to examine three distinct constructs that have been linked to motivation in students: self-efficacy, interest, and effort beliefs (Zimmerman, 2000; Hidi and Renninger, 2006; Jones *et al.*, 2012). We chose these three variables both for their salience within motivation literature and their influence on student performance and retention. Self-efficacy has been found to be positively correlated to achievement outcomes in many studies, as well as the adaptive motivational processes, effort and persistence (Pintrich and De Groot, 1990; Multon *et al.*, 1991; Pajares and Miller, 1994; Zusho *et al.*, 2003). Individual interest is less consistently correlated with performance, but is more strongly linked to positive learning strategies, such as mastery goals and attention (Hidi and Renninger, 2006). Furthermore, discipline-specific individual interest has been positively correlated with choice of major and number of classes taken within that discipline (Harackiewicz *et al.*, 2008). Positive effort beliefs are very highly correlated with an incremental theory of intelligence, which states that competence is not fixed, but malleable (Blackwell *et al.*, 2007; Dweck, 2012; Jones *et al.*, 2012). Effort is absolutely vital for success in any college classroom. However, preceding the action of effort itself must be a positive, adaptive belief about the potential impacts of the action. Hence, effort beliefs are indicators about how a student perceives the impact of their effort on learning and performance, and students who believe their competence can be changed through effort tend to be more motivated and perform at higher levels than those who do not (Dweck, 2000).

The emphasis in academic motivational research has shifted from behavioral aspects such as drive and reinforcement to beliefs, goals, and expectations over the last 40 years (Wigfield and Eccles, 2002). These modern theories of motivation point to the critical role that expectancies and beliefs play in adaptive learning patterns among students (Eccles and Wigfield, 2002; Zeldin *et al.*, 2008). By researching these beliefs among students in college chemistry, we can have a more clear understanding of the basis for motivational processes that exist in our classrooms. However, to do this we first need measures of the various motivational aspects that have been tested within the target population.

Self-Efficacy

Self-efficacy is rooted in social cognitive theory and is defined as the self-appraisal of one's capacity to execute a specific task (Bandura, 1977). Efficacy beliefs or expectations are self-referent and guide an individual toward certain actions and away from others (Pajares, 1996). Self-efficacy must be distinguished from two theoretically similar constructs, outcome expectations and self-concept. Outcome expectations are beliefs that certain behaviors will lead to certain outcomes. Both self-efficacy and outcome expectations influence motivation, but self-efficacy is thought to play a larger role in predicting

achievement (Zimmerman, 2000). While both self-efficacy and outcome expectations are task-specific, self-concept is much more broadly defined, and has more to do with one's beliefs about their self-worth and competence within a domain (Pajares and Miller, 1994). Moreover, self-concept has been defined in many different ways, is tied to affective as well as cognitive judgments, and is inherently norm-referenced (Hansford and Hattie, 1982; Bong and Clark, 1999). In contrast, the prevailing definition of self-efficacy has remained virtually unchanged since its inception (Bandura, 1977). Also, self-efficacy judgments are less likely to be influenced by social comparisons and affective swings due to the task-specific, objective nature of the construct (Bong and Clark, 1999). For these reasons, and the notable presence of self-efficacy in the literature, we chose to measure self-efficacy instead of other related constructs.

Bandura (1986) argued that efficacy expectations are drawn from four sources of information: performance accomplishments, vicarious experiences, verbal persuasion, and physiological states. The most dominant of these sources are performance accomplishments, because they are founded upon personal mastery experiences (Bandura, 1977). However, some suggest the relative salience of sources for self-efficacy beliefs may be different for males and females (Zeldin *et al.*, 2008). Nevertheless, a student in chemistry is more likely to have high efficacy expectations for a particular task if he or she has already successfully completed that task. When asked to explain their self-efficacy in college chemistry, students noted their prior success in chemistry as a common theme (Dalgety and Coll, 2006). From a quantitative approach, Lopez and Lent (1992) found that prior experience in math explained the most variance in self-efficacy scores when considering math self-concept, interest in math, and perceived value of mathematics. Equally important to considering the sources of self-efficacy, is the influence that self-efficacy has on the student.

Self-efficacy is hypothesized to have far-reaching implications in academics by influencing students' effort, perseverance, and emotional reactions to specific tasks in school (Lent *et al.*, 1984; Lopez and Lent, 1992; Pajares and Kranzler, 1995; Pajares, 1996). Lent et al. (1984) investigated the relationship of students' self-efficacy beliefs to persistence in technical and science majors. They found that students who reported higher self-efficacy scores for completing their educational requirements were more likely to persist in their major. This supports Hackett and Betz's (1981) hypothesis that self-efficacy is linked to persistence in career goals.

Other studies across several disciplines suggest that self-efficacy is related to academic performance, problem solving, college entrance, and college major choice (Betz and Hackett, 1983; Lent and Hackett, 1987; Lopez and Lent, 1992; Pajares and Kranzler, 1995; Andrew, 1998; Britner and Pajares, 2001; Schunk and Pajares, 2002; Zusho *et al.*, 2003; Parker *et al.*, 2014). Zimmerman, Bandura, and Martinez-Pons (1992) found that students' academic self-efficacy scores significantly predicted ($\beta = .39$) their final grade in a high school social studies class. In a college setting, it was found that students' self-efficacy beliefs were significantly correlated to their grade-point average (GPA) and accounted for more variation in GPA than ACT scores (Gore, 2006).

Generalization can also occur with efficacy expectations, such that a student who has experienced mastery with one task could report high efficacy expectations for a similar task (Bandura *et al.*, 1969). However, the degree of generalization is limited to the domain of functioning (Zimmerman, 2000). For example, one cannot assume that just because a student has high self-efficacy in biology, he or she will also have high self-efficacy in chemistry. Thus, it is important, when measuring and describing students' self-efficacy, to ensure that it is domain-specific.

Chemistry has been referred to as the "central science" and it is believed that mastery of chemistry concepts is influential for success in later science courses (Tai *et al.*, 2005).

Although chemistry course requirements are not as widespread as those of math, there are many professions unrelated to chemistry, particularly in health care, that require a background in chemistry (Brown *et al.*, 2012). In fact, students pursuing degrees other than chemistry make up the bulk of the enrollment in introductory chemistry at larger universities, and it is these students who are least likely to exhibit high self-efficacy in chemistry (Uzuntiryaki and Aydin, 2009). Therefore, since self-efficacy could have a substantial influence on students' achievement and retention in college chemistry, it is an important factor to consider in chemical education research (Zusho *et al.*, 2003).

Several researchers have investigated chemistry-specific self-efficacy in university-level classes (Dalgety *et al.*, 2003; Zusho *et al.*, 2003; Taasoobshirazi and Glynn, 2009; Uzuntiryaki and Aydin, 2009; Villafane *et al.*, 2014). Taasoobshirazi and Glynn (2009) conducted a study where undergraduate introductory chemistry students were asked to solve a series of problems along with judging their self-efficacy in chemistry. It was found that students who reported high self-efficacy were more likely to use forward-working strategies and obtain the problem solution than students who reported low self-efficacy. In a related study, Zusho et al. (2003) investigated the role of several motivational processes on achievement in chemistry, as well as the correlations between them. They found that self-efficacy was the best predictor of final course grade, when measuring SAT-math, task value, and rehearsal strategies. Both of these studies support Bandura's (1986) notion that self-efficacy is positively related to achievement through persistence and effort. Villafañe et al. (2014) explored student self-efficacy trajectories across a semester of preparatory chemistry. Their results suggest that individual characteristics (race/ethnicity and gender) could influence the degree to which students show an increase in self-efficacy across the semester. Although these studies have addressed important gaps in the literature, more research is needed that investigates the interplay between different motivational variables.

Interest

Interest, in academic settings, can be described as a psychological state where the student is engaging or has a predisposition to engage with content over time (Hidi and Renninger, 2006). Hidi and Baird (1986) argued that interest is more than "arousal," but must be considered as a process. As a process, interest is said to endure and persist through time. There are two types of interest outlined in the literature: situational interest and individual interest (Hidi, 1990). Situational interest refers to an interest that is triggered spontaneously through interaction with the environment (Harackiewicz *et al.*, 2008). This type of interest can be considered to deliver a sense of enjoyment and curiosity, but may or may not persist within the person (Renninger, 2000). Individual interest refers to a relatively stable interest that has developed over time and is associated with an enduring predisposition for the person to reengage with specific topics, subject areas, or activities (Hidi, 1990; Schiefele, 1991; Hidi and Renninger, 2006). Individual interest can be further broken down into two components: feeling-related and value-related (Schiefele, 1999). Feeling-related interest (e.g., Chemistry is fascinating to me) is tied to stimulation and enjoyment. Value-related interest (e.g.. The material we are learning in chemistry is important for me to know) is associated with importance and personal significance and has been positively correlated with performance in academic contexts (Hulleman *et al.*, 2008).

Hidi and Renninger (2006) postulated that interest is a developmental process, which occurs in four phases: triggered situational interest, maintained situational interest, emerging individual interest, and well-developed individual interest. Triggered situational interest results from temporary cognitive or affective changes in the individual (Hidi and Baird, 1986). Maintained situational interest occurs after a triggered event, where the individual's interest is "held" and involves focused attention that endures for a period of time. If

maintained situational interest persists long enough, it becomes emerging situational interest, characterized by stored value, positive feelings, and consistent reengagement with the material or activity. At some point, emerging interest can become well-developed interest following substantial reengagement with the material. Individuals with well-developed interest seek out answers to questions, and are likely to be resourceful when answers are not easily found. Also, it is possible for these individuals to expend effort, but feel as though the task is "effortless" (Renninger and Hidi, 2002). Although the development of interest is not the same for all individuals, Hidi and Renninger (2006) argued that there is no evidence suggesting well-developed interest can spawn without the individuals first being exposed to the area and experiencing triggered interest.

In the same vein as efficacy beliefs, interest is content-specific and represents a personal significance between the individual and the object of his or her interest (Schiefele, 1991; Renninger and Hidi, 2002). Science, as a subject matter in schools, is somewhat broad and diffuse in elementary and middle school years, but becomes differentiated and more focused during high school and college years. In light of this, several studies on interest in science and/or students' perceptions of the value of science at lower grade levels have focused on "science" as a whole (Anderman and Young, 1994; Singh *et al.*, 2002; Tuan *et al.*, 2005). However, studies on interest in science dealing with high school and college student populations tend to center on specific disciplines, such as chemistry, physics, and biology (Dalgety *et al.*, 2003; Uitto *et al.*, 2006; Gungor *et al.*, 2007; Barbera *et al.*, 2008). These studies underscore the importance of developing measures to target individual interest as a domain-specific construct. Hence, to understand more about interest as a component of motivation in the college chemistry classroom, the focus of the interest measure must be specifically directed towards the discipline of chemistry.

Effort Beliefs

Effort can be considered as part of the attributional theory of motivation, and from this theoretical standpoint, is intimately tied to conceptions of ability (Weiner, 1985). Weiner (1985) pointed out that ability and effort are the most salient causal ascriptions to achievement. In short, students believe that those who have high ability and display high effort will be more successful than students who have low ability and display low effort. The notion that effort and persistence has a positive effect on the academic outcome of a student has been supported by empirical studies (Stipek and Gralinski, 1996; Elliot, 1999). For example, Elliot (1999) found that self-reported persistence and effort were positive predictors of academic performance. Effort was found to be a mediator between adaptive mastery goals and academic performance. Another study revealed that students who expressed positive beliefs toward the value of effort do not necessarily show increased performance, but do tend to focus more on mastery and the development of their abilities (Stipek and Gralinski, 1996). Effort is certainly a key component in the academic success of students. It must be considered when making judgments about academic performance, due to how it mediates the link between motivational constructs and academic outcomes (Elliot, 1999; Goodman *et al.*, 2011)

Students' beliefs about effort are a precedent to effortful actions and are highly correlated with their conceptions about intelligence and ability (Blackwell *et al.*, 2007; Jones *et al.*, 2012). These conceptions are referred to as implicit theories of intelligence. Implicit theories are "beliefs about the nature of human attributes," two implicit theories of intelligence have been described in the literature: incremental and entity (Dweck, 2012). An incremental theory of intelligence is characterized by the view that intelligence is malleable and can change over time with the expenditure of effort. Conversely, individuals who hold an entity view of intelligence see it as fixed and unchanging, independent of effort (Dweck

and Leggett, 1988). Students holding an incremental view of intelligence are more likely to see effort as enhancing ability and apply more effort to overcome obstacles. On the contrary, those endorsing an entity view of intelligence are less likely to put forth effort in the face of failure, less likely to be interested in a subject, and exhibit achievement gaps when compared to incremental theorists (Dweck and Sorich, 1999; Hong *et al.*, 2004; Dweck, 2012). Effort beliefs and implicit theories of intelligence are fundamentally different constructs, but deeply related (Blackwell et al., 2007; Jones, Wilkins, Long, & Wang, 2012). Between these two, the vast majority of research has centered on implicit theories, leaving a gap with respect to effort beliefs. No studies were found on effort beliefs in a college setting, but a few studies have investigated effort beliefs in secondary school (Blackwell et al., 2007; Jones et al., 2012).

It is important that students feel they can improve upon their abilities with persistence and effort. This could be a challenge to get across to college students, as it is likely that their beliefs about intelligence and effort have already developed by the time they reach the college classroom. However, if we understand more about effort beliefs and how they can change among college-age students, new instructional strategies can be implemented so that more students will endorse positive effort beliefs in their classes. This is particularly important in math and science, because most students have experienced both prior to coming to college, and have likely developed beliefs about their abilities in those domains.

Purpose and Rationale for Current Study

There is a need in the chemistry education community to understand some of the motivational and affective components of students enrolled in chemistry courses. What is true of one college subject is not necessarily true of another, and given the lack of motivational research in college-level sciences, it is important that this need be addressed.

In order to effectively assess a large classroom of students on their motivational characteristics and dispositions, instructors and researchers must rely upon easy to administer self-report scales or instruments, consisting of items targeted at measuring a specific latent trait or group of traits. As with any scale or instrument, either in the physical or social sciences, the quality of data that can be produced from it depends largely on the quality of the data generated with the target population (Barbera and VandenPlas, 2011; Heredia and Lewis, 2012). Thus, steps must be taken to ensure that a scale or instrument will produce valid and reliable results when used with the target population.

While general and science-specific motivation instruments exist, such as the Motivated Strategies and Learning Questionnaire (MSLQ), and the Science Motivation Questionnaire (SMQ), the current availability of individual scales that measure motivational variables in college chemistry is limited (Pintrich *et al.*, 1993; Glynn *et al.*, 2009). In 2003, Bauer and colleagues published the Chemistry Self-Concept Inventory (CSCI), a 40-item survey that was adapted to measure students' self-concept in five domains: math, chemistry, academic, academic enjoyment, and creativity. There are two instruments (Colorado Learning Attitudes about Science Survey (CLASS) and Chemistry Attitude and Experiences Questionnaire (CAEQ)), which measure interest in chemistry (Dalgety *et al.*, 2003; Barbera *et al.*, 2008). The MSLQ and SMQ also have general academic interest scales, which could be adapted for a chemistry context. However, none of these instruments were designed based on the prevailing theoretical underpinnings of interest theory from educational psychology (Schiefele, 1991; Renninger, 2000). The CAEQ, being the largest instrument for motivation and affect in college chemistry, also has a self-efficacy scale. The CAEQ together with the College Chemistry Self-Efficacy Scale (CCSS) represent the only two instruments designed specifically to measure self-efficacy in a college chemistry setting. Both of these were designed using Bandura's (1986) theory of self-efficacy, which is widely accepted among

researchers across many disciplines. The MSLQ and SMQ also have self-efficacy scales. However, the items in both scales are very general, relating more to the course as a whole than to tasks within the course. As mentioned above, self-efficacy is conceptualized as confidence at the task-level. Thus, scales purported to measure self-efficacy in a particular academic domain should be written with items targeting specific tasks encountered within that academic domain. We carefully examined the self-efficacy items from those that had been used in chemistry and related disciplines for wording that was most appropriate for an introductory chemistry class in a college setting. The CCSS contained items that were relevant, task-specific, and readable for college students in introductory chemistry. No chemistry-specific measures for effort beliefs, or implicit theories of intelligence were found.

The goal of the present study was to identify, utilize, and evaluate existing measures of motivational constructs (initial interest, maintained interest, self-efficacy, and effort beliefs) in a college chemistry setting. As doing this required modification of items and scales, this manuscript represents evidence of validity and reliability for these measures prior to their use in future studies of instructional styles. Items from existing scales must be modified to be discipline-specific, as the constructs under investigation are operationalized around a meaningful connection between the subject area and the individual (Bandura, 1986; Schiefele, 1991). Modifying items, scales, and instruments for subject-specific language and meaning is common practice in science and math education (Barbera *et al.*, 2008; Taasoobshirazi and Glynn, 2009; Jones *et al.*, 2012). However, any modifications to an item, scale, or instrument must be followed up by an investigation for validity and reliability evidence (Barbera and VandenPlas, 2011).

Items were taken from three independently published scales, each measuring a separate motivational construct (self-efficacy, interest, or effort beliefs) in academic settings. The items from the effort beliefs and interest scales were adapted to fit a college chemistry context. The self-efficacy items were originally developed for college chemistry and did not need modification. All scales were subjected to confirmatory factor analysis to determine if the structure of each scale matched that proposed by the individual scale developers. In addition, indicators of global and component model fit were used to assess whether any of the scales should be modified. Student interviews were also conducted and used as validity evidence for potential modifications of items and scales. Interviews with the target population are a vital part of evaluating an item, scale, or instrument for response process validity (Arjoon *et al.*, 2013; Wren and Barbera, 2013). The data collection and subsequent analyses were guided by the following research questions:

(1) What modifications are needed to produce brief, chemistry-specific scales of self-efficacy, interest, and effort beliefs?

(2) What evidence supports the functioning of each of the modified scales?

(3) To what extent do students' interest, effort beliefs, and self-efficacy change across a semester of college chemistry?

The current study will contribute to the growing body of literature centered on motivational and affective processes among students in college chemistry courses. In addition, brief chemistry-specific scales for measuring three salient motivational beliefs will be made available to educators and researchers interested in gauging the motivational climate of their chemistry classrooms. These scales will be used in a follow-up study as variables in a path analysis to investigate a set of *a priori* motivation models that will include measures of academic performance in general chemistry. These scales provide important tools for educators who plan to implement new teaching strategies, and are interested in more than just the performance outcomes that result from those strategies. The follow-up study will provide detailed connections among the scales and student performance.

Methods

Participants

Quantitative Studies. Participants for the initial study were recruited from all first-semester general chemistry laboratory sections at a mid-sized Rocky Mountain region university during the fall of 2013. This course is required by several science and health majors, and represents the first of two courses in the general chemistry sequence. Additional participants for the cross-validation studies were recruited during the fall semester of 2014 at the same institution as well as at a second institution in the same US region.

Qualitative Studies. Interviews were conducted with students from the target population (first-semester general chemistry students) to gather evidence for the response process validity of the modified items and scales (AERA, APA, NCME, 1999). In the fall of 2013, when the scales were given to lab sections, the last item asked students to indicate whether or not they would participate in a short interview regarding the survey. Students who indicated interest were contacted through the school e-mail address they provided on the survey. In the spring of 2014, additional participants were recruited via an announcement during lecture. Interested students volunteered by adding their name to a sign-up sheet passed out and collected by one of the authors.

In accordance with Institutional Review Board (IRB) policy, students in both the quantitative and qualitative studies were informed that their participation had no impact on their course grade and that they would be volunteering for a research study regarding their academic motivation. Standard university policies for confidentiality and data handling were utilized throughout the study. Participation in the studies was voluntary and no incentives were given to students for participating.

Data Collection

Quantitative. Data for the initial study were collected in all laboratory sections of the course at two time points. At the pre-semester data collection (time 1), only students who were enrolled in a lab section and were present in lab during the 1st week of the semester could be included in the study. At the post-semester data collection (time 2), only students who were present in lab during the 13th week of the semester could be included in the study. At each time point, the teaching assistants for the labs gave a prescribed announcement regarding the purpose of the study and instructions for completing the survey. Following that, students were administered a packet containing the survey with all scales and a set of demographic items (see Appendix for survey and demographic items). Each student was asked to provide an identifier, so that time 2 responses could be matched to that specific student. Students were given approximately 15 minutes to complete the items on the survey and six demographic items. All students were required to take the survey, but were informed that their data would only be used if they signed the consent form. The consent form covered both the time 1 and time 2 data collections and was only offered during time 1; thus, scores from the time 2 administration were not retained unless they could be matched to consent forms from the time 1 administration. In addition to individual time 1 and time 2 analyses, a matched-pair sample was also used for analyses involving changes in students' scores across the semester. Hence, two different sample sizes were used in the study. The total number of complete data sets from the time 1 administration was 373. The total number of complete data sets from time 2 (and hence the matched-data set pairs) was 294. All available and complete sets of scores from the start of the semester were retained, as we wanted to test the functionality of the scales with the entire incoming population, regardless of their future trajectory in the course. Data for the cross-validation studies were collected following the

same protocols as noted above, with two exceptions: demographic data was not collected for the entire sample, and time 2 data was only collected at the main institution.

Scales

Preliminary wording changes. In order to appropriately assess chemistry students on the three latent traits (self-efficacy, interest, and effort beliefs) being measured, we made minor wording changes and adjustments to the measurement scales where needed. These preliminary modifications are described below for the original scales. All quantitative and qualitative study participants were given these modified scales (hereafter referred to as "scales"). Changes made to the scales after administration to the students and interviews were data-driven. The resulting scales following these data-driven revisions hereafter are referred to as "revised" scales.

Chemistry self-efficacy. The self-efficacy scale was taken from the College Chemistry Self-Efficacy Scale (CCSS; (Uzuntiryaki and Aydin, 2009). These items are designed to measure a student's perception of his or her ability to complete a given task in a chemistry course. The original instrument (21 items) has three subscales of chemistry self-efficacy: self-efficacy for cognitive skills (12 items), self-efficacy for psychomotor skills (5 items), and self-efficacy for everyday application (4 items). The original items are on a 9-point Likert-type scale ranging from "very poorly" to "very well." Select items from the self-efficacy for cognitive skills subscale were evaluated for this study. Uzuntiryaki & Aydin (2009) reported a Cronbach's alpha estimate of 0.92 for this subscale. These items are intended to measure students' belief in their ability to work through intellectual operations in chemistry (Uzuntiryaki and Aydin, 2009). Our main interest in students' self-efficacy is related to chemistry problems encountered during the lecture portion of class; thus, items related to the laboratory or the nature of science were excluded (see Appendix for items). An example item that was excluded is: "How well can you write a lab report summarizing main findings?" Example items that were retained include: "To what extent can you explain chemical laws and theories" and "How well can you read the formulas of elements and compounds?" In the original instrument developed by Uzuntiryaki & Aydin (2009), nine numerical choices were given, but only five delineated categorical choices, ranging from "very poorly" to "very well", were placed above the numbers. The student then, is left with multiple numerical choices per category. The meaning of the difference between the two numerical choices is therefore lost. Hence, we argue that for clarity, if only five categories are given, then only five numerical choices are necessary. As there was no compelling reason to retain the nine options, and to allow for electronic scoring, we changed the nine-point Likert-type scale used in the original instrument to a five-point Likert-type scale. Since we are testing the internal structure of the scale, a CFA result consistent with the CFA from the original authors (Uzuntiryaki and Aydin, 2009) will provide favorable support for the condensed number of response options. In addition, there is evidence to support that changing scale length does little to affect the distribution about the mean, skewness, or kurtosis (Dawes, 2008).

Initial and Maintained Interest. The original initial interest and maintained interest scales were adapted from a survey developed by Harackiewicz et al. (2008). The initial interest items were designed to measure a student's interest in psychology at the beginning of an introductory undergraduate psychology course. The maintained interest items were given to students at week 13 of the semester and were designed to measure the "hold" component of situational interest. Exploratory and confirmatory factor analysis performed with these items confirmed the distinction between the "catch" and "hold" components of situational interest

(Linnenbrink et al., 2007 reported in Harackiewicz et al. 2008, see also Linnenbrink et al. 2010). The original scales for initial interest and maintained interest have seven and nine items, respectively. Both original scales are measured on a seven-point Likert-type scale ranging from "not at all true of me" to "very true of me." The Cronbach's alpha estimates ($\alpha = 0.90$ for the initial interest scale and $\alpha = 0.95$ for the maintained interest scale) for scores based on a sample of 1,265 college students in an introductory psychology class were deemed acceptable (Harackiewicz *et al.*, 2008). We modified the wording of these items slightly to fit the context of a chemistry course, mostly by just replacing the word "psychology" with the word "chemistry." For example, the item, "I am really looking forward to learning more about psychology" was changed to "I am really looking forward to learning more about chemistry." As the authors provided no rationale for retaining the 7-point scale, and to keep the number of response options consistent across all measures, we adjusted the scale to 5-points. Lastly, the responses on the scale were changed to from "true of me" statements to "agree" statements (see Appendix for items) as agree-type response options better fit the wording of the items.

Effort beliefs. The original items for the effort beliefs scale were developed by Sorich and Dweck (1997) and first used in Blackwell's (2002) unpublished doctoral dissertation study, which involved seventh grade students. The nine-item effort beliefs scale was designed to measure the degree to which students believe their effort will lead to positive outcomes. These items were then adapted for use in a motivational study involving ninth grade math students (Jones *et al.*, 2012). The effort beliefs scale used by Jones et al. (2012) consisted of nine items measured by a six-point Likert-type scale ranging from "strongly disagree" to "strongly agree." Jones et al. (2012) found the Cronbach's alpha ($\alpha = .77$) estimate for a sample of 163 ninth-grade math students acceptable. We used the exact wording of each item (based on Jones et al. (2012) version), except for substituting the word "chemistry" for the word "math." The scale consists of four positive items ("If a chemistry assignment is hard, I'll probably learn a lot doing it"), and five negative items ("To tell the truth, when I work hard at chemistry, it makes me feel like I'm not very smart"). In addition, we adjusted the scale range from a 6-point to a 5-point Likert scale.

Qualitative

Interview Protocol. All interviews took place in a private interview room to ensure both participant confidentiality and audio quality. Prior to starting the interview, each participant was informed about the purpose of the study, the interview procedure, and the protocols for confidentiality. Following that, the participants signed a consent form approved by the IRB. Since students interviewed in the fall 2013 had already completed the scales at the start of the semester, they were provided a copy of their original answer choices for each scale. Students who were interviewed in the spring of 2014 were asked to complete the scales prior to the start of the interview. All interviews were audio recorded. A verbal probing interview approach was used, whereby students were asked to read each item out-loud, explain the reasoning behind the answer choice they made, and comment on the readability of the items (Knafl *et al.*, 2007). If a student's reasoning did not match their answer choice or was unclear to the researcher, additional probing questions were asked to clarify their interpretation of the item. This methodology is important in establishing the response process validity (Arjoon *et al.*, 2013) for the modified items and response scales, ensuring proper readability and consistency between students' answer choices and reasoning among the target population (Barbera and VandenPlas, 2011).

Quantitative Data Analysis

Descriptives. Descriptive statistics were analyzed on all data to check for skew and kurtosis as well as to determine means and standard deviations. In line with what is commonly accepted, we considered acceptable skew and kurtosis values as falling in the range of -1 to +1 (Huck, 2012). Reliability estimates for internal consistency (Cronbach's α) were calculated for each scale as well. Cronbach's α is an estimate of the internal consistency in the responses and should be reported with respect to each scale (AERA, APA, NCME, 1999). A value of 0.70 is considered acceptable for classroom multiple-choice tests and rating scales (Murphy and Davidshofer, 2005). Statistical Package for the Social Sciences (SPSS) 20.0 software was used for these analyses.

Time 1 and Time 2 Measurements. Analysis of variance (ANOVA) tests were performed on each scale at both time points separately to determine if any significant differences existed among students' scores based on major choice. Paired samples *t*-tests were conducted to determine if any changes were significant in students' scores across the semester. All item scores from each scale were aggregated to produce a scale score. In congruence with what has been commonly reported in the field of chemical education and among authors of the scales used in this paper, a mean score for each scale, based on the raw item scores, was produced (Blackwell, 2002; Dalgety and Salter, 2002; Zusho *et al.*, 2003; Lewis *et al.*, 2009; Jones *et al.*, 2012). The mean score aggregation method was used as each scale contains a different number of items. Therefore, this method allows for consistent interpretation across scales and will lead to less variance when utilizing scale scores in future path analyses (Kline, 2011). To further assess if these changes were different by major, mixed-between-within ANOVA tests were employed for the interest and self-efficacy scales. As we do not have any theoretical underpinnings or prior studies to support the existence of differences in effort beliefs by academic major, we cannot use this type of comparison to provide supporting evidence of validity for the effort beliefs scale. Only students who took the survey at both time points in the initial study were included in this analysis (n = 294). All tests were evaluated at $p < 0.05$.

Confirmatory Factor Analysis. Confirmatory factor analysis (CFA) is a powerful tool for assessing how well a proposed model fits a set of measured variables. To date, a few studies in chemical education have used CFA during scale development and validation (Uzuntiryaki and Aydin, 2009; Xu and Lewis, 2011; Raker *et al.*, 2013). Each scale (i.e., self-efficacy, initial interest, and effort beliefs) was considered a latent variable (or factor) and each item, an indicator for its respective scale. To identify any problematic items that should be considered for deletion and ascertain the fit of each indicator to the appropriate latent variable, a one-factor CFA was conducted for each scale. One item per latent variable was set to unity. Only complete data sets were included in the analyses; thus, list-wise deletion was used for any missing data. All CFAs were performed using LISREL version 8.80 (Jöreskog and Sörbom, 2006). Analyses were based on the robust maximum likelihood (RML) estimator, as the data were treated as ordinal and were non-normal. The commonly used maximum likelihood (ML) estimator is not appropriate for these analyses because the data must be continuous and normal for the ML estimator to be unbiased with respect to fit indices, parameter estimates, and standard errors (Finney & DiStefano, 2006).

Global fit of each one-factor model was analyzed based on several indices including: Satorra-Bentler (SB) scaled chi-square (Satorra and Bentler, 1994), root mean squared error of approximation (RMSEA; (Steiger, 1990), non-normed fit index (TLI; (Tucker and Lewis, 1973), standardized root mean squared residual (SRMR) , and comparative fit index (CFI; (Bentler, 1990). The SB scaled chi-square is a test for exact model fit, where the population

covariances are fully reproduced by the hypothesized model. A non-significant result is desired and indicates that there are not significant discrepancies between the population covariances and those predicted by the model (Kline, 2011). However, the chi-square test is sensitive to sample size and will often produce a significant result for very small deviations in model fit. Thus, other descriptive test statistics are used to assess the fit of the model (Schermelleh-Engel, Moosbrugger, & Müller, 2003).

There are two types of fit indices commonly used in the literature to assess model fit: absolute and incremental. Absolute fit indices (i.e., chi-square, RMSEA, and SRMR) are estimates of how well an *a priori* model fits the data. Incremental, or comparative fit indices (i.e., TLI and CFI) reflect improvement of model fit compared to a baseline model (Kline, 2011). The RMSEA can range from 0 to infinity and is a measure of the approximate model fit in the population (Steiger, 1990). Because exact fit of the model in the population is impractical, the RMSEA is a measure of "close fit," and in general, values < 0.05 are considered good and those < 0.08 are considered reasonable (Browne and Cudeck, 1992). The SRMR value ranges from 0 to 1 and is a "badness of fit" measure based on the standardized fitted residuals. By standardizing the residuals, the scale of the variables is taken into account (Schermelleh-Engel *et al.*, 2003). Hu and Bentler (1995) suggested that an SRMR value of < 0.05 is indicative of good fit and < 0.10 is acceptable. The TLI and CFI both take into account the chi-square values of the proposed model and the null baseline model (Brown, 2006). The TLI and CFI values are normed and range from 0 to 1, with values ≥ 0.95 indicating good fit (Hu and Bentler, 1999). Only when several fit indices (both incremental and absolute) are considered together can the quality model fit be assessed with reasonable propriety (Brown, 2006). Based on what is commonly accepted in the literature, we used the following cut-off values as an evaluation of acceptable model fit beyond the chi-square test statistic: RMSEA ≤ 0.05, SRMR ≤ 0.10, TLI and CFI ≥ 0.95 (Hu and Bentler, 1999; Hooper *et al.*, 2008).

Component model fit was evaluated based on statistical significance ($p < 0.05$) and reasonable parameter estimates. In addition, modification indices were considered when significant. The modification index (MI) is represented as a one degree of freedom chi-square statistic that estimates the difference between two nested models. Modification indices are parameter-specific and reflect the approximate decrease in the model chi-square statistic when the fixed parameter is allowed to be freely estimated (Brown, 2006). With regard to CFAs, MI values are most commonly associated with correlated residuals between two indicator variables or between an indicator and two factors in the model. Thus, the higher the MI, the more likely it is that a particular indicator either is redundant or belongs with another factor. Modification indices are part of the evidence used to assess whether an indicator (item) should be dropped, aggregated with another indicator, or relocated to a different factor (latent trait). When the MI value exceeds the critical χ^2 value ($\alpha = 0.05$) using the degrees of freedom (*df)* from the model, then the appropriate modification should be considered (Hancock, 1999).

The one-factor CFAs were conducted on data collected at each time point. At time 1, all complete data sets were included for those students who consented to the study ($n = 373$). At time 2, only matched data sets were used in the analysis ($n = 294$). Although more than 294 students participated at the end of the semester, consent forms were only issued at the beginning of the semester, precluding the use of data from students who might have been absent at the first data collection or added the class late.

Qualitative Data Analysis

All interviews were transcribed and coded for significant statements and emergent themes, based on each item and its corresponding scale (Creswell, 2013). The strategy for

coding centered on readability and the degree of consistency among participants' interpretations of the items. If students repeatedly report dissimilar interpretations of an item, then there can be no consensus on what the score of that item really means. This is problematic, as it negatively affects the validity of the scale and the inferences that can be drawn from the scores.

Results

Interview results

A total of nine interviews (2 males and 7 females) were conducted in the fall of 2013 and five additional interviews (4 males and 1 female) in the spring of 2014. We felt that this number of interviews was sufficient to reach consensus on the items, especially given that we were not designing them from scratch. The students who participated in the fall 2013 interviews were asked to comment on the initial interest, self-efficacy, and effort beliefs scales. The interviews were conducted during the middle of the semester; therefore, students had not yet been given the maintained interest scale. For this reason, and to solidify the results from the self-efficacy scale, the interviews in the spring of 2014 covered the maintained interest and self-efficacy items. The effort beliefs scale and initial interest scales were not included in the spring interviews because we observed consistent responses from participants regarding the meaning and interpretability of items during the fall interviews.

In-depth student interviews were conducted using the full versions of the scales. The results from the interviews were used in conjunction with quantitative results from the CFAs as support to flag any items that should be considered for modification or removal.

Readability and Interpretation. Overall, the items in all scales showed good readability during the interviews. Participants read most of the items without stumbling and seemed to have a good grasp on the flow of each statement. However, there was one item in the effort beliefs scale that failed to show adequate readability and was confusing for many of the participants. Item 7 (see Appendix for items) reads, *If you don't do well in chemistry and put in a lot of effort, you won't do well*. During the interview, many participants had to read the item at least twice before explaining their reasoning for the answer they chose. Several participants regarded the item as "confusing". Below is an example of how one student had to double back on his answer choice.

Oh, it was a little confusing, I guess, yea. I would have said – I kind of connected it with the other ones and just said strongly agree. But I guess – but now I would say I disagree because…Oh, wait, let me re-phrase that. Yeah, it's a little confusing, the wording. I strongly agree with that too.

This student started off with an affirmative response, then he switched his answer to a negative response, before returning to his original choice. The confusion over the wording of this item was consistent throughout the interviews.

In addition to readability, we were interested in how students attributed meaning to the items. Toward this end, participants were asked to explain their reasoning behind the answer choice they made. Most participants gave plausible reasons for the answers they chose, and provided rational explanations. However, there were two items from the self-efficacy scale (Items 3 and 8) for which the agreement of the meaning differed among several students. Item 3 (*How well can you describe the structure of an atom?*) was problematic because there were differing opinions about what it meant to "describe the structure of an atom." Some students reported describing the structure of an atom simply meant knowing "the positions of things and charges". Others reported that "interactions" and valence shell theory were part of the description. We observed, in several cases, that this item could be

interpreted to varying degrees of depth and understanding. Item 8 (*How well can you solve chemistry problems?)* was also problematic for a similar reason. Participants regarded this item as "broad", "vague", and "depending on the problem". Clearly, there are many types of problems students encounter in first-semester general chemistry. Diffuse tasks such as these can lead to problems when a student tries to self-appraise their ability to complete the task (Bandura, 1986). Additionally, item 8 is somewhat redundant in that every item that precedes it represents some type of chemistry problem.

Feeling-related interest versus Value-related interest. Individual interest is conceptualized as having both feeling-related (emotional arousal) and value-related (importance/utility) components (Schiefele, 1999). The scale used in this study was designed to measure initial interest, thus items to measure both components were incorporated into the scale. However, in the original study from which the items were adapted, initial interest was presented as a single factor (Harackiewicz *et al.*, 2008). In a later study on situational interest, a similar set of items were grouped into two factors of interest: feeling-related and value-related (Linnenbrink-Garcia *et al.*, 2010). As the factor structure is an important part of a scale and the validity of the gathered data, we were concerned with how students responded to feeling-related versus value-related items. If similar reasoning were given for all interest items, then the qualitative evidence to split the scale into two factors would be missing. If, however, there was a clear demarcation between reasons used for answers to feeling-related versus value-related items, then two factors might be a more valid interpretation of the scale.

We found that participants during the fall and spring interviews used dissimilar language when describing their reasons for answers to feeling-related items versus value-related items. Examples of feeling-related items from the initial interest scale are: *I am fascinated by chemistry* and *I chose to take general chemistry because I'm really interested in the topic*. Participants cited reasons for choosing their answers by using words and phrases such as: "I'm naturally gifted", "I connect with the material", "It excites me", "Interested", "Fascinated". These words are evoked from "feelings of involvement, stimulation, and enjoyment" toward the topic of chemistry, and is exactly the type of interest that is characterized by feeling-related items (Schiefele, 1999). On the other hand, participants explaining their answer choices for value-related items used entirely different language. Examples of value-related items from the initial interest scale are: *I think what we will study in general chemistry will be important for me to know* and *I think the field of chemistry is an important discipline*. Participants commented on their answer choices by using phrases such as: "I'm going to be building off this", "it will obviously be important in my field", "this class will…help with future chemistry classes", "chemistry is…everything around us". Schiefele (1999) describes value-related interest as being directed toward something that is personally significant and important to the individual. The statements made by the participants during the interviews regarding value-related items were indicative of a personal significance and importance, as opposed to feelings of excitement or enjoyment. Based on the overwhelming difference we found in how participants described their interest using feeling-related items versus value-related items, the scale was tested as both a 1 and 2-factor model.

Data Screening and Descriptive statistics

Prior to analysis, all data sets were screened for careless responses from students (i.e., students selecting all of one response option). Only one case was found that exhibited this pattern. Missing item-level data were also screened for patterns. The only consistently missed item was the last item on the list. Of 37 total cases that had missing data, 17 of them

failed to respond to the last item. As list-wise deletion was implemented, all cases with missing data were removed from each data set prior to analysis.

Mean, standard deviation, skew and kurtosis were evaluated for each item on each scale (see Appendix). Most items had skew and kurtosis values that were within acceptable ranges to be considered normal (-1 to +1) (Huck, 2012). However, some items were outside of this range with negative skew values down to ~ -1.5 and kurtosis values up to 3.5. Due to these deviations from normality, the robust maximum likelihood (RML) estimator was used in all CFA runs. The RML estimator utilizes the Satorra-Bentler scaled chi-square statistic, and is robust with respect to non-normal data (Chou *et al.*, 1991).

Demographics. Demographic data was collected from all participants in the initial study during the fall of 2013. In the fall of 2014, a cross-validation study was conducted (n = 1160), but demographic data was only collected from a sub-sample of these participants (n = 175), those for whom both pre- and post-semester data was gathered. Of the participants in the initial study (n = 373, pre-semester), most were female (67%). Nearly half (46%) were non-science majors (nursing, sports and exercise science, statistics, earth science), 32% were other science majors (biology, physics, or mathematics), 20% were chemistry majors, and 2% were undeclared. In this sample, the majority of students were first-year (60%) and second-year (13%) university students. Most took a chemistry course in secondary school (89%). Similar demographic breakdowns from the initial study were observed for the matched-pair sample (n = 294), with all categories within 1 percentage point of the reported statistics. Demographic data from the cross-validation study (n = 175) showed a similar breakdown. Most of the students were female (73%), and most reported taking chemistry in secondary school (93%). Non-science majors made up the bulk of the sample (62%), followed by other science (24%), and chemistry (13%). Nearly 80% of the sample was first and second-year university students.

Reliability analysis. The internal consistency estimates (Cronbach's α) for each scale were acceptable to high, ranging from 0.75 to 0.88 (see Tables 1 and 2). As revisions were made to the scales, based on qualitative data, fit indices, and factor loadings from the single factor CFAs, the alpha values dropped for two of the scales (self-efficacy and initial interest), and increased for the effort beliefs scale. Despite this, all alpha values remained acceptable to high (0.77 to 0.89).

Time 1 CFAs. Each scale was evaluated using a single-factor CFA using a sample of first-semester general chemistry students at the start of the semester (n = 373). The goals of the analyses were to substantiate each scales structure and to seek possibilities to shorten the scales. From the quantitative side, revisions to the model (e.g., dropping items) were guided by the fit indices as well as the modification indexes for each scale. We considered parallel qualitative evidence together with the CFA results before making decisions about model revisions. Table 1 shows the values of the fit indices for each time 1 CFA (χ^2, RMSEA, NFI, CFI, and SRMR) before and after revision of the model.

Table 1. *CFA fit indices and reliability estimates of preliminary and revised scales at time 1 for the initial sample (n = 373)*

Scale	# of items	χ^2 value	*df*[a]	*p*-value	RMSEA	TLI	CFI	SRMR	α[b]
Initial interest	7	158.81	14	< 0.001	0.17	0.93	0.95	0.09	0.88

Initial feeling	4	23.84	13	0.033	0.05	0.99	1.00	0.03	0.90
Initial value (revised)	3								0.79
Effort beliefs	9	148.25	27	< 0.001	0.11	0.92	0.93	0.10	0.75
Effort beliefs (revised)	6	16.27	9	0.26	0.05	0.99	0.99	0.05	0.77
Self-efficacy	8	96.49	20	< 0.001	0.10	0.97	0.98	0.06	0.89
Self-efficacy (revised)	6	43.32	9	< 0.001	0.10	0.97	0.98	0.05	0.85

[a]Degrees of freedom (*df)* are based on RML estimator [b]Cronbach's alpha
χ^2 – likelihood ratio test, RMSEA – root mean squared error of approximation, TLI – Tucker-Lewis index, CFI – comparative fit index, SRMR – standardized root mean squared residual

Initial interest. The model for initial interest was based on seven indicators for the 1-factor solution (see Appendix for all tested models). The standardized factor loadings for all seven indicators were significant ($p < 0.05$). However, after analyzing the fit and modification indices, it was clear that the model did not adequately fit the data. The global fit index, SB-scaled chi-square test, indicated inadequate fit of the model to the data χ^2 (14, n = 373) = 158.81, $p < 0.001$. However, a significant chi-square test is very common with large sample sizes. Component fit indices, RMSEA value (0.17) and SRMR (0.094), also suggested inadequate fit (see Table 1). In addition, items 5, 6, and 7 displayed markedly lower standardized factor loadings than items 1 – 4. Based on these results, and the qualitative evidence suggesting students use different language when describing feeling-related versus value-related interest, the scale was split into two factors. A second CFA was run with the scale split into feeling-related and value-related factors. All of the reported fit indices improved for the 2-factor model, meeting the acceptable cut-off values. Although the SB-scaled chi-square test remained significant, χ^2 (13, n = 373) = 23.84, $p < 0.033$, the improvement of other fit indices following revision, suggested reasonable fit.

Effort beliefs. The 1-factor model for effort beliefs was composed of nine indicators, five of which were negatively worded and were reverse-coded for analysis (see Appendix for all tested models). The RMSEA (0.11), TLI (0.92), and CFI (0.93), as well as low standardized factor loadings (< 0.40) for some items, suggested inadequate fit of the model. Three items (1, 7, and 9) were dropped based on low factor loadings (< 0.40). Moreover, items 7 and 9 had large modification indices with other items in the scale, suggesting correlated error among those items. Once these three items were removed, a second 1-factor CFA was run and all of the reported fit indices improved to be within the appropriate ranges considered acceptable. In addition, the SB-scaled chi-square statistic, χ^2 (9, n = 373) = 162.7, p = 0.061, was not statistically significant. Taken together, these results suggest the revised model fits the data well.

Self-efficacy. The 1-factor model for self-efficacy was composed of eight indicators (see Appendix for all tested models). The RMSEA (0.10) and SRMR (0.06) values, as well as large modification indices for several items, suggested poor fit of the model. Items 2, 3 and 8 all had high modification indices with at least two other items. Additionally, items 3 and 8 were found to be problematic in the student interviews due to ambiguity in the meaning that students attributed to them. Hence, both items were removed from the model. Item 2 was

left in the model as a high modification index should not be the only criteria for removing an indicator from a model and no other quantitative or qualitative results supported removal. A second 1-factor CFA was run with the revised scale, but the fit indices did not suggest improved model fit. The removal of these items was neither an improvement nor a detriment to the model fit. In spite of this, we chose to retain the revised scale. We feel that the qualitative data and high MI values are sufficient reasons to justify removing these items, and therefore shortening, the scale. Although the original authors of this scale had a larger CFA model with two additional subscales, our CFA results for selected items from the self-efficacy for cognitive skills (SCS) subscale are consistent with those from the authors (Uzuntiryaki and Aydin, 2009). The only exception to this is the RMSEA value from our model, which was slightly inflated (0.10) compared to the original authors' model (0.08).

Time 2 CFAs. The self-efficacy and effort beliefs scales consisted of the same items from time 1 to time 2. CFAs, using the revised models from time 1, were run on these two scales to confirm the revised scale structures and functionality of items. We found that the two revised models fit the time 2 data adequately (see Table 2). The items for maintained interest were not identical to those of initial interest, thus a 1-factor model was evaluated and subsequent revisions made.

Maintained interest. The 1-factor model for maintained interest was based on eight indicators (see Appendix for all tested models). As with the initial interest scale, the maintained interest items were composed of both feeling-related and value-related interest. One item on the maintained interest scale was negatively worded and was reverse coded prior to analysis. When the 1-factor model was run, the fit was very poor, χ^2 (20, n = 294) = 222.34, $p < 0.001$, RMSEA (0.19), TLI (0.92), CFI (0.94), and SRMR (0.12). All reported global and component fit indices were outside of the acceptable ranges. When the model was split into two factors (feeling-related and value-related), the fit improved dramatically. The SB-scaled chi-square statistic was not significant (p = 0.26), suggesting adequate global model fit. Additionally, all component fit indices improved, falling within good to acceptable ranges. Standardized factor loadings for the indicators were significant ($p < 0.05$) with both models.

Table 2. *CFA fit indices and reliability estimates of preliminary and revised scales at time 2 for the initial sample (n = 294)*

Scale	# of items	χ^2 value	*df*[a]	*p*-value	RMSEA	TLI	CFI	SRMR	α[b]
Maintained interest	8	222.34	20	< 0.001	0.19	0.92	0.94	0.12	0.91
Maintained feeling	4	22.49	19	0.26	0.03	1.00	1.00	0.04	0.92
Maintained value (revised)	4								0.87
Effort beliefs (revised)	6	15.49	9	0.08	0.05	0.99	0.99	0.03	0.82
Self-efficacy (revised)	6	36.03	9	< 0.001	0.10	0.97	0.98	0.04	0.87

[a]Degrees of freedom (*df*) are based on RML estimator [b]Cronbach's alpha
χ^2 – likelihood ratio test, RMSEA – root mean squared error of approximation, TLI – Tucker-Lewis index, CFI – comparative fit index, SRMR – standardized root mean squared residual

Cross-validation of revised scales. Supporting evidence for the structural validity of the revised scales is provided by evaluating the model fit with alternate samples from the same target population (Kline, 2011). It is advised that anytime a model is revised, the revised model be cross-validated with an independent sample. Samples for cross-validation studies could come from the original data set, if the sample size is large enough, or from a completely separate data collection (Brown, 2006, p. 124). As the initial data from fall 2013 was used to make revisions to each scale, cross-validation samples were collected in fall 2014 and used to further validate the revised scales. The first sample (cross-validation 1) was collected at the same institution as the initial data set, a second sample (cross-validation 2) was collected at a different institution. Due to administration constraints, post-semester data was not collected from the second sample; therefore, the maintained interest items were not cross-validated with this population. A comparison of fit indices and reliability estimates for all samples is presented in Table 3.

Table 3. *CFA fit indices and reliability estimates for initial and cross-validation samples*

Fit Index	Revised scales (n = 373)				Cross-validation 1 (n = 432)				Cross-validation 2 (n = 728)		
	II	**MI**	**EB**	**SE**	**II**	**MI[c]**	**EB**	**SE**	**II**	**EB**	**SE**
χ^2	23.84	22.49	16.27	31.31	49.02	18.31	60.01	62.74	80.44	48.61	66.78
***df*[a]**	13	19	9	9	13	19	9	9	13	9	9
***p*-value**	0.03	0.26	0.26	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
RMSEA	0.05	0.03	0.05	0.08	0.08	0.00	0.11	0.12	0.08	0.08	0.09
TLI	0.99	1.00	0.99	0.98	0.99	1.00	0.96	0.96	0.98	0.97	0.96
CFI	1.00	1.00	0.99	0.99	0.99	1.00	0.98	0.98	0.99	0.98	0.98
SRMR	0.03	0.04	0.05	0.04	0.05	0.04	0.08	0.04	0.04	0.06	0.05
α[b]	0.90 0.79	0.92 0.87	0.77	0.85	0.91 0.90	0.94 0.86	0.84	0.88	0.90 0.84	0.75	0.82

[a]Degrees of freedom (*df*) are based on RML estimator [b]Cronbach's alpha [c]Post-semester n=175
χ^2 – likelihood ratio test, RMSEA – root mean squared error of approximation, TLI – Tucker-Lewis index, CFI – comparative fit index, SRMR – standardized root mean squared residual

Fit indices and alpha values from the cross-validation samples confirm the validity and reliability of the revised scales. All scales had consistently high Cronbach's alpha values across all samples, indicating similar reliability for each administration. With only a few exceptions, the fit indices were within the range of acceptable values. In the cross-validation 1 sample, the effort beliefs scale produced SRMR (0.08) and RMSEA (0.11) values above the standard cut-offs (≤0.06 and ≤0.08). However, Hu and Bentler report that acceptable SRMR values can be as high as 0.09 with CFI and TLI values >0.95 (Hu and Bentler, 1999). Therefore, with only the RMSEA value being out of range, the revised scale model has acceptable fit to the data from this sample. The self-efficacy scale produced inflated RMSEA values (0.12 and 0.09) in both cross-validation samples; however, the other indices were well within the acceptable ranges. Therefore, the revised self-efficacy scale is deemed to have acceptable fit to both populations. As both additional data sets were larger than the original data set, it is expected that the SB-scaled chi-squared values would be significant as the chi-square test is highly dependent on sample size (Brown, 2006). The only exception to this was the maintained interest scale, as this post-semester data set only contained 175 students. Taken together, the consistent fit indices and alpha values across all samples provide supporting evidence for the validity and reliability of the revised scales across three different samples from two different institutions.

Pre- and post-semester comparisons. The comparison of pre-semester (time 1) with post-semester (time 2) scores was conducted using only the matched sample data sets from the initial sample (n = 294) and the cross-validation study sub-sample (n = 175). To check for data patterns from those students who did not have time 2 responses, demographic items and item-level means from time 1 were compared between those with and without time 2 data. The frequencies of responses to all demographic items and item-level means appeared very similar, indicating that the two groups of students were likely from the same population. The initial interest and maintained interest scales address related traits, however, due to wording differences each scale consisted of different items from time 1 to time 2. Hence, the mean scores from the scales (Table 4) cannot be directly compared with a t-test. However, the scales could be used to compare sub-groups of students (e.g., major choice, ethnicity, or gender) based on how their interest changed relative to one another. In a future study, we will be using these scales as part of a path analysis to investigate the connections between these motivational factors and student performance. There was a significant drop in effort beliefs (M_1 = 3.95, M_2 = 3.77, see Table 5) among all students, $t(293) = 4.35$, $p < 0.001$. However, the effect size (d = 0.29) was small (Cohen, 1992). This trend was observed for the cross-validation sub-sample as well (M_1 = 4.03, M_2 = 3.91) with a significant drop in effort beliefs scores, $t(174) = 2.30$, $p = 0.02$. The effect size (d = 0.18) of this difference was also small. Of the three scales, the most change across the semester and variation by group in students' scores was observed for the self-efficacy scale (Table 5).

Table 4. *Mean values of scores for initial and maintained interest at times 1 and 2 for all students from the initial sample (n = 294)*

Scale	Mean (SD) time 1	Mean (SD) time 2
Intitial interest (feeling)	3.46 (0.84)	-
Initial interest (value)	4.06 (0.67)	-
Maintained interest (feeling)	-	3.23 (0.90)
Maintained interest (value)	-	3.61 (0.83)

Note: Interest scales were different from time 1 to time 2 and cannot be directly compared.

Self-efficacy overall. Self-efficacy is one's self-appraisal of ability to complete a task. Our measure of chemistry self-efficacy included tasks that would be commonly encountered in a first-semester chemistry class, such as: explaining the structure of an atom, or choosing an appropriate formula to solve a problem. As instructors, we would expect our students to improve upon these tasks during the course of a semester and, we would expect their self-appraisals of ability to improve as well.

Our results suggest that students' chemistry self-efficacy increased across the semester for both the initial sample and cross-validation sub-sample (see Table 5). Self-efficacy scores were based on a mean composite score of the revised scale (6 items). The mean difference in scores for the initial study (M_1 = 3.29, M_2 = 3.60) for all students across the semester was significant, $t(293) = 8.23$, $p < 0.001$). The effect size (d = 0.50) for this comparison was medium (Cohen, 1992). Data from the cross-validation study sub-sample showed a similar trend (M_1 = 2.87, M_2 = 3.69), with the difference also being significant, $t(174) = 14.83$, $p < 0.001$). The effect size (d = 1.19) for this test was large (Cohen, 1992).

On average, students from both samples felt more confident in their abilities to solve chemistry problems at the end of the semester than at the beginning of the semester.

While our results suggest an increase in self-efficacy for our overall sample, this trend might not hold for all students in the sample. For example, Villafane *et al.* (2014) reported differing trajectories in chemistry self-efficacy based upon ethnic group. Similar trends were observed by Zusho *et al.* (2003) with regard to performance in chemistry. They reported that the self-efficacy of students who were "low achievers" in chemistry dropped sharply across a semester, and those who were "average achievers" dropped slightly. In contrast, students who were "high achievers" reported higher self-efficacy at the end of the semester than at the start. Collectively, these two studies demonstrate that students' self-efficacy trends across a semester depend on several factors, some of which may change during the semester. Hence, instructors should be aware and expectant of such differing trends in self-efficacy among their students, especially when evaluating the effectiveness of a novel approach to instruction.

Table 5. *Mean values of scores for self-efficacy and effort beliefs at times 1 and 2 for all students from both the initial sample and cross-validation sub-sample*

Study	Scale	Mean (SD) time 1	Mean (SD) time 2	Mean difference[a] (effect size[b])	*p*-value
Initial (*n* = 294)	Effort beliefs	3.95 (0.57)	3.77 (0.68)	-0.18 (0.29)	< 0.001
	Self-efficacy	3.29 (0.60)	3.60 (0.65)	0.31 (0.50)	< 0.001
Cross-validation (*n* = 175)	Effort beliefs	4.03 (0.57)	3.91 (0.74)	-0.13 (0.18)	0.022
	Self-efficacy	2.87 (0.73)	3.69 (0.64)	0.71 (1.19)	< 0.001

[a]Based on paired samples *t*-test [b]Effect size represented by *Cohen's d* – small (0.20), medium (0.50), large (0.80) (Cohen, 1992)

Self-efficacy and interest by major. We were interested in whether declared chemistry majors differed from those in other majors on self-efficacy and interest. We expected chemistry majors to score differently than other majors based on the nature of interest and self-efficacy in academic choice (Lent *et al.*, 1994). Barbera et al. (2008) found that chemistry majors were more interested in chemistry than non-majors. To confirm this and test the notion that chemistry majors would also be more likely to report higher self-efficacy than non-majors, we performed ANOVA tests on both time 1 and time 2 data (see Table 6). All majors (chemistry, other science, non-science, and undeclared) were compared in the ANOVA test; only the post-hoc results on chemistry versus non-science majors are reported in Table 6. The overall ANOVA model for the pre-semester self-efficacy data was significant, $F(3,293) = 4.20$, $p = 0.006$. The assumption of homogeneity of variances was not violated, as indicated by a non-significant result of Levene's test ($p > 0.05$). Post-hoc analysis with the Tukey test showed that chemistry majors' self-efficacy ($M = 3.49$, $SD = 0.62$) was higher than non-science majors' ($M = 3.17$, $SD = 0.59$). This difference was significant at $p < 0.01$. The overall ANOVA model for the post-semester self-efficacy data was not significant, indicating that chemistry majors did not differ from other majors at the end of the semester.

The overall models for the two components of initial interest (feeling- and value-related) were significant for the initial sample, $F(3,293) = 20.87$, $p < 0.001$, and $F(3, 293) = 7.12$, $p < 0.001$, respectively. The assumption of homogeneity of variances was not violated for any of the tests performed, as indicated by non-significant results using Levene's test ($p >$

0.05). Post-hoc analyses with Tukey tests revealed that chemistry majors reported higher feeling- (M_1 = 4.14, SD_1 = 0.84) and value-related interest (M_2 = 4.38, SD_2 = 0.77) than non-science majors (M_1 = 3.19, SD_1 = 0.74, M_2 = 3.91, SD_2 = 0.62) (see Table 6). The same trend was evident in the post-semester data with the two components of maintained interest being significant for the initial sample, $F(3,293) = 12.93$, $p < 0.001$, and $F(3,293) = 7.72$, $p < 0.001$, respectively. Post-hoc analyses revealed that chemistry majors' reported higher maintained feeling- (M_1 = 3.73, SD_1 = 0.94) and value-related interest (M_2 = 3.95, SD_2 = 0.91) than non-science majors (M_1 = 2.93, SD_1 = 0.81, M_2 = 3.38, SD_2 = 0.79). All differences reported were significant at $p < 0.01$. The corresponding effect sizes (Cohen's *d*) are considered medium to large (Cohen, 1992).

A sub-sample used in the cross-validation study was also evaluated for differences in self-efficacy and interest by major. Participants did not differ in self-efficacy, at either pre or post-semester. Participants did, however, report different levels of interest (feeling and value) based on major at the start of the semester, $F(3,174) = 9.30$, $p < 0.001$, and $F(3, 174) = 4.59$, $p = 0.004$, respectively. The same was true for maintained interest (feeling and value) at the end of the semester, $F(3,174) = 6.53$, $p < 0.001$, and $F(3, 293) = 4.68$, $p = 0.004$, respectively. Based on our prior results reported above, we hypothesized that chemistry majors would report higher levels of interest than non-science majors. Hence, we performed planned contrasts to test this hypothesis. Results indicate that chemistry majors showed significantly more initial feeling-related interest, $t(171) = 4.56$, $p < .001$, and value-related interest, $t(171) = 3.66$, $p < .001$, than non-science majors. The same was true for maintained-feeling, $t(171) = 3.67$, $p < .001$, and maintained-value interests, $t(171) = 2.93$, $p = .004$ (see Appendix).

Table 6. *Mean scores and differences between chemistry majors and non-science majors on interest and self-efficacy scales for the initial sample*

Scale	Chemistry majors $n = 57$ Mean score (SD)	Non-science majors $n = 136$ Mean score (SD)	Mean difference[a] (effect size[b])	*p*-value
Intitial interest (feeling)	4.14 (0.84)	3.19 (0.74)	.95 (1.20)	< 0.001
Initial interest (value)	4.38 (0.77)	3.91 (0.62)	.47 (0.67)	< 0.001
Maintained interest (feeling)	3.73 (0.94)	2.93 (0.81)	.80 (0.91)	< 0.001
Maintained interest (value)	3.95 (0.91)	3.38 (0.79)	.57 (0.67)	< 0.001
Self-efficacy (time 1)	3.49 (0.62)	3.17 (0.59)	.32 (0.53)	0.005

[a]Based on Tukey's post-hoc tests [b]Effect size represented by *Cohen's d* – small (0.20), medium (0.50), large (0.80) (Cohen, 1992)

These data support the findings from the study by Barbera et al. (2008) concerning personal interest among chemistry majors versus other majors. In addition, these results expand upon the findings reported by Uzuntiryaki & Ayden (2009) whereby "[chemistry] majors scored higher than non-majors [on self-efficacy for cognitive skills]; however, they did not appear significant." Most importantly, these data demonstrate the ability of the

modified items and revised scales to discriminate between populations of students who would be expected to score differently on self-efficacy and interest in chemistry.

Pre and post-semester comparisons of mean self-efficacy scores were analyzed by major (Table 7). These results indicate that all students with a declared major (n = 287) reported higher self-efficacy at the end of the semester than at the beginning of the semester, regardless of their major. All differences were significant at $p < 0.05$. Chemistry majors reported improved self-efficacy at the end of the semester (M_2 = 3.70) compared to the start of the semester (M_1 = 3.18), $t(56) = 5.39$, $p < .001$. The same was true for other science majors (M_1 = 2.95, M_2 = 3.62), $t(93) = 10.03$, $p < .001$; and non-science majors (M_1 = 2.78, M_2 = 3.45), $t(135) = 12.49$, $p < .001$. The effect sizes for the differences in self-efficacy from pre to post-semester were medium to large among all three groups of majors, ranging from d = .71 for chemistry majors to d = 1.0 for other science and non-science majors (Cohen, 1992). These results expand upon those reported in Table 5 by suggesting that students from all majors reported improved self-efficacy in chemistry after a semester of instruction.

Table 7. *Mean values of self-efficacy scores by major at time 1 and time 2 from the initial sample (n = 287)*

Major	Mean (SD) time 1	Mean (SD) time 2	Mean difference[a] (effect size[b])	*p*-value
Chemistry	3.17 (0.69)	3.70 (0.79)	0.53 (0.71)	< 0.001
Other science	2.95 (0.66)	3.62 (0.61)	0.67 (1.00)	< 0.001
Non-science	2.78 (0.65)	3.44 (0.60)	0.66 (1.00)	< 0.001

[a]Based on paired samples *t*-test [b]Effect size represented by *Cohen's d* – small (0.20), medium (0.50), large (0.80) (Cohen, 1992)

Discussion

Self-efficacy, interest, and effort beliefs are salient factors associated with academic motivation and are supported by a strong foundation of research (Weiner, 1985; Bandura, 1986; Renninger, 2000; Blackwell *et al.*, 2007). A chemistry-specific set of scales designed to measure interest, self-efficacy, and effort beliefs were administered to a sample of first-semester general chemistry students. The major goal of this study was to establish evidence of validity and reliability for scores from the scales such that they could be used in future studies regarding the impact of various teaching practices on these motivational factors and their relation to course performance. Demonstrating validity and reliability for data generated by a scale or instrument is paramount following any alterations to items or use in a new setting. Absence of such evidence renders the interpretations of scale scores invalid and can lead to misinformed decision-making (Arjoon *et al.*, 2013). In this study, we have presented evidence to support the internal consistency as well as the response process and structural validity of the modified items and revised scales.

Validity evidence based on response processes pertains to the agreement between the construct being measured and the actual processes respondents engage in when they generate an answer (AERA, APA, NCME, 1999). In the case of the interest scales, participants routinely described the meaning of the value-related items differently from the feeling-related items. These results, in conjunction with previous findings by Linnenbrink *et al.* (2012), led us to split the larger interest scales into two smaller subscales, which resulted in improved fit of the CFA models. Our revisions to the self-efficacy and effort beliefs scales were guided

more by the way students interpreted the items. When participants assign varied interpretations to a particular item, the meaning of that item is obscured and integrity of the score and construct associated with that item is compromised. Items that were found to illicit ambiguous or frequently incongruent responses from participants were flagged as problematic. Those items that also demonstrated lack of fit or redundancy in the CFA models were removed from the corresponding scale. Overall, our interview results for the items retained in each scale suggest that students consistently and adequately understood the meaning of the items.

To demonstrate functionality based on the internal structure of the individual scales, 1-factor CFAs were conducted to examine the degree to which the data fit each hypothesized model. Confirmatory factor analysis allows the researcher to test whether a proposed grouping of items, and the scores associated with them, appropriately describe a latent variable (Brown, 2006). Our preliminary psychometric evaluation of each scale revealed that the model fits were less than satisfactory. Revisions to each scale were informed by our qualitative studies with students from the target population as well as from the modification indices of the initial 1-factor CFA studies. Following revision of the effort beliefs and self-efficacy scales, and splitting of the interest scale into two factors, the model fit for each scale improved dramatically (see Tables 1 and 2). Nearly all fit indices improved to values considered acceptable for the CFAs conducted on data from both time points. The only exceptions to this were the RMSEA values for self-efficacy, and the chi-square values for initial interest and self-efficacy. As stated previously, the model chi-square test is a "badness of fit" index, where a significant result is not desired. Chi-square tests are sensitive to sample size and often, negligible deviations in fit produce significant results with large samples (Brown, 2006). The sample sizes (n = 373, n = 294) in our models would be considered large based on a recommended subject to indicator ratio of 1:10 (Bentler and Chou, 1987). Thus, it is not surprising that all of our preliminary models and two of our revised models failed to produce non-significant chi-square values. However, consistent with most studies involving factor analysis, we used approximate fit indices as alternative indicators of adequate model fit. The model describing the self-efficacy scale was the only one that remained problematic after revision, due to the significant chi-square value ($p < 0.05$) and high RMSEA value (0.10). Like the model chi-square test, the RMSEA value is considered a "badness of fit" index, where lower values are desired (Kline, 2011). Kline (2011) points out that values $\geq$ 0.10 might signal a "serious problem" with the model. Consequently, we urge readers to interpret our results for this model with caution, but also to consider that RMSEA values tend to be inflated in models with a small number of indicators (Breivik and Olsson, 2001).

As the revised scales were derived from qualitative and quantitative results with a single population, additional evidence for the structural validity of each revised scale was established using two additional samples. These cross-validation studies were comprised of data collected from students in the same course, at the same institution, during the year following the initial study. The second sample was obtained from students in the same course at a different institution. Supporting evidence for the structural validity of the revised scales was provided by obtaining similar fit indices with both student samples (Kline, 2011). Across all three populations the fit indices remained consistent. With the exception of the RMSEA values for the effort beliefs scale from sample 1 and the self-efficacy scale from both samples, all other indices were within recommended ranges. As the cross-validation samples sizes were larger, all of the SB-scaled chi-squared statistics were significant, however, this is not seen as a threat to the structural validity as this test is highly dependent on sample size (Brown, 2006). The maintained interest scale did have a non-significant chi-square value, however, data for this one scale was derived from a much smaller (n = 175) post-semester population.

To further support the functionality of the interest and self-efficacy scales in a college chemistry setting, we evaluated the extent to which chemistry majors differed from other majors. Based on prior studies, we operated under the assumption that declared chemistry majors would have higher self-efficacy and interest toward chemistry (Barbera *et al.*, 2008; Uzuntiryaki and Aydin, 2009). Our initial data show that chemistry majors began the semester with higher self-efficacy and interest than non-science majors. However, the gap in self-efficacy scores closed by the end of the semester, indicating no significant difference based on major. For the cross-validation sub-sample, the difference in interest scores between chemistry majors and non-science majors was retained, but there was no significant difference in self-efficacy. It is possible that the lower sample size (n = 175), and hence, lower number of chemistry majors (n = 23) compared to the initial sample could be a cause of this discrepancy. Results from the maintained interest scale suggest that chemistry majors continued to have higher levels of interest than non-science majors through the end of the semester. This is certainly plausible and expected, given that enduring interest in a particular subject area has been shown to predict major choice and number of courses taken in that subject area (Harackiewicz *et al.*, 2000; Harackiewicz *et al.*, 2008). Taken together, these results suggest that the self-efficacy and interest scales can discriminate between groups for whom it is plausible to expect differences in confidence and interest toward chemistry, providing further validity evidence of the scales (Standard 1.14, AERA, APA, NCME, 1999).

The internal consistency of scales is an estimate of reliability that relates to how well items within a scale describe the same construct (Henson, 2001). Cronbach's alpha is the most commonly reported value of internal consistency, and a cutoff value of 0.70 is often used to indicate moderate internal consistency among items used in classroom rating scales (Murphy and Davidshofer, 2005). Following item modifications and scale revisions, all of our scales had reliability estimates ≥ 0.77. These results suggest that the items belonging to each scale are consistent with other items in the same scale in describing the specific construct. Consistently high alpha values were also obtained in the cross-validation studies, further supporting the internal consistency of the revised scales with the target population.

Limitations

While the results from our psychometric evaluation of these revised scales suggest that they function well among general chemistry students, we acknowledge several limitations to the study. First, our sample size from time 1 to time 2 in the initial study dropped by 79 participants (21%). While the time 2 sample size (n = 294) remained large enough for factor analysis, the missing participants could represent an important subset of the population (e.g., students who dropped the course). However, as with any study involving multiple collections of data from a single sample, there is always a risk of attrition. While our cross-validation studies provide positive support for the generalizability of the revised scales, we encourage researchers from other institutions to use and further evaluate this set of scales, so that educators can have a more complete understanding of the psychometric properties and generalizability of these scales. Finally, we acknowledge that the meaning attributed to items in the self-efficacy scale may be different among students in the same population. We excluded two items from the scale (items 3 and 8) due to ambiguity and a lack of consensus on the meaning among interview participants. However, these were items that had the most frequently incongruent responses. Participants in our sample did not necessarily assign the exact same meaning to the remaining items. For instance, item 4 reads, *How well can you describe the properties of elements by using the periodic table?* A student with a strong background in chemistry might, for example, interpret "properties of elements" as electronegativity, ionization energy, and bonding tendencies of elements. On the other hand, a student with a weak background in chemistry may view "properties of elements" as simply

the number of protons, neutrons, and electrons in a given element. Thus, students with a strong chemistry background may, in fact, underestimate their ability because they have a deeper understanding of the theories and facets associated with certain tasks in chemistry. Students with less understanding of a chemistry task may inflate their appraisal of their ability due to an oversimplification of the task. In addition, students who are from non-English language backgrounds may interpret self-efficacy items differently than native English speakers (Lee and Fradd, 1998). Self-efficacy items involving tasks that are not completely objective and defined will always leave room for loose interpretations. However, items that are too narrowly focused and specific will lose generalizability and require the instrument to be long and arduous in order to cover the set of topics in a given course. On the other hand, Pajares (1996) cautions against using measures too general by stating that, "omnibus tests…transform self-efficacy into a generalized personality trait rather than a context-specific judgment." The balance between specificity and generality with efficacy beliefs is a difficult aspect to fully resolve, a sentiment shared by other researchers as well (Tschannen-Moran and Hoy, 2001). We feel that our qualitative data, while limited, offers some useful insight into items from the self-efficacy scale, which would be most imprecisely interpreted by students in our target population. Although some of the items retained in the scale could be also interpreted in several ways, at varying degrees of understanding; our interview data demonstrates that most students had a consistent grasp of what the task meant. Researchers concerned with the interpretability of this and other self-efficacy scales in chemistry could extend upon these findings by conducting interviews with different groups of students. Students with diverse chemistry backgrounds and experiences, as well as those whose first language is not English may provide perspectives on items that are not obvious to the developer and users of the scale. This depth of information should be regarded as absolutely vital for the validity of any inferences drawn from scores from the scale.

Implications and Future Research

Our interest when designing this study stemmed from our desire to study student motivation in chemistry. Due to the limited work in this area, our first step was to evaluate a set of modified items and revised scales, so that various aspects of motivation could be measured for a chemistry-specific population. Time constraints can often prevent instructors from administering lengthy scales or instruments, therefore, we tried to compile scales that would provide a balance between useful data and classroom administration time. Furthermore, student participation and completion rates tend to be lower with longer, more time-consuming instruments (Lichtenstein *et al.*, 2008; Heredia and Lewis, 2012). Additionally, it is crucial that each scale is actually measuring what it's developers have purported it to be measuring. Therefore, it is incumbent upon researchers to thoroughly examine relevant psychometric evidence of such scales prior to, or as part of, their use. With the present work, we show that each of our revised scales to measure self-efficacy, interest, and effort beliefs demonstrate acceptable psychometric properties for use in a general chemistry setting. Additionally, our revised scales, which measure three well-defined latent traits, are comprised of a small number of items. Our revised scales consist of 7 initial interest, 8 maintained interest, 6 effort beliefs, and 6 self-efficacy items. When used together in future studies, this equates to 19 (with initial interest) or 20 (with maintained interest) total items to address three latent traits. By comparison, the CAEQ (measuring three distinct latent traits), and the CSCI (measuring five types of self-concept) are comprised of 69 and 40 items, respectively (Dalgety *et al.*, 2003; Bauer, 2005).

Our next step is to now utilize these scales to study the impact of teaching practice on students' self-efficacy, interest, and effort beliefs. Many studies are suggestive of the powerful influence of self-efficacy on performance (Zimmerman *et al.*, 1992; Pajares, 1996;

Zusho *et al.*, 2003). To further corroborate these findings, and expand upon them in the college chemistry setting, studies involving measures of self-efficacy and performance together are needed. Even less explored are the relationships among interest, effort beliefs, and performance. There is a particular lack of research involving these latent traits in college level sciences. In future studies, we plan to investigate how different practices affect these motivational factors as well as how these factors affect each other and ultimately students' performance. Therefore, follow-up studies will use *a priori* path analysis models to evaluate the correlation between scales and their mediation of course performance (Xu *et al.*, 2013).

In addition to our ongoing studies, we offer a few avenues for extension of the current study. A large-scale study could further utilize the power of structural equation modeling through invariance (or measurement equivalence) analysis. Invariance analysis allows the researcher to test whether a proposed model is equivalent across different groups of participants. One form of testing for invariance is to use a multi-group CFA, whereby the researcher is able to test for the equivalence of the measurement and structural solution (Brown, 2006). Put simply, invariance analysis can inform the researcher as to whether the same trait is being measured across different groups (race, gender, major), which is an important consideration for the validity of an instrument (Hutchinson *et al.*, 2008).

Instructors who are interested in gauging the motivational atmosphere of their classes might find our chemistry-specific scales useful. Students' beliefs about motivation and effort precede and govern their actions in the course. At the start of a semester, an instructor might want to have knowledge of his or her students' interest and confidence toward chemistry for the purpose of tailoring certain aspects of the course to their group of students. Due to the brevity of the scales, data could be collected at multiple time points throughout the semester with minimal time commitment. This could be especially useful if an innovative instructional strategy were to be implemented. The instructor could evaluate the impact of their instructional strategy on dimensions beyond course performance measures. This would be informative as performance measures alone tell instructors nothing about a student's motivational or affective disposition toward the course, which are vital components of a student's academic success (Zusho *et al.*, 2003). We encourage educators who employ novel instructional strategies to consider measuring the motivational and affective processes of their students, in order to add to the current understanding of the impacts of these strategies. As stated in the 2012 DBER report, "the interplay between faculty behavior [i.e., teaching strategies] and student affect merits further exploration." We feel that the present work aids instructors in this exploration by providing measurement tools adapted for college chemistry, founded on prevailing theories from educational psychology, and subjected to the rigor of a thorough psychometric evaluation.

References

AERA, APA, NCME (1999), *Standards for educational and psychological testing*, Washington, DC: American Educational Research Association.

Ames, C., (1992), Classrooms - Goals, Structures, and Student Motivation, *Journal of educational Psychology,* **84**(3), 261-271.

Anderman, E. M. and Young, A. J., (1994), Motivation and strategy use in science: Individual differences and classroom effects, *Journal of research in science teaching,* **31**(8), 811-831.

Andrew, S., (1998), Self‐efficacy as a predictor of academic performance in science, *Journal of advanced nursing,* **27**(3), 596-603.

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, *Journal of Chemical Education,* **90**(5), 536-545.

Bandura, A., (1977), Self-efficacy: toward a unifying theory of behavioral change, *Psychological review,* **84**(2), 191-215.

Bandura, A., (1986), *Social foundations of thought and action: A social cognitive theory*, Englewood Cliffs, NJ: Prentice Hall.

Bandura, A., Blanchard, E. B. and Ritter, B., (1969), Relative efficacy of desensitization and modeling approaches for inducing behavioral, affective, and attitudinal changes, *Journal of Personality and Social Psychology,* **13**(3), 173.

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

Barbera, J. and VandenPlas, J. R. (2011). All Assessment Materials Are Not Created Equal: The Myths about Instrument Development, Validity, and Reliability, Investigating Classroom Myths through Research on Teaching and Learning, American Chemical Society, **1074**, 177-193.

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

Bentler, P. M., (1990), Comparative fit indexes in structural models, *Psychological bulletin,* **107**(2), 238-246.

Bentler, P. M. and Chou, C.-P., (1987), Practical issues in structural modeling, *Sociological Methods & Research,* **16**(1), 78-117.

Betz, N. E. and Hackett, G., (1983), The Relationship of mathematics self-efficacy expectations to the selection of science-based college majors, *Journal of Vocational Behavior,* **23**(3), 329-345.

Blackwell, L. (2002). Psychological mediators of student achievement during the transition to junior high school: The role of implicit theories Unpublished Doctoral Dissertation, Columbia University.

Blackwell, L. S., Trzesniewski, K. H. and Dweck, C. S., (2007), Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention, *Child development,* **78**(1), 246-263.

Bong, M. and Clark, R. E., (1999), Comparison between self-concept and self-efficacy in academic motivation research, *Educational psychologist,* **34**(3), 139-153.

Breivik, E. and Olsson, U. H., (2001), Adding variables to improve fit: The effect of model size on fit assessment in LISREL, *Structural equation modeling: Present and future*, 169-194.

Britner, S. L. and Pajares, F., (2001), Self-efficacy beliefs, motivation, race, and gender in middle school science, *Journal of Women and Minorities in Science and Engineering,* **7**(4), 15.

Brophy, J. E., (2010), *Motivating students to learn*, New York: Routledge.

Brown, C. E., Henry, M. L., Barbera, J. and Hyslop, R. M., (2012), A Bridge between Two Cultures: Uncovering the Chemistry Concepts Relevant to the Nursing Clinical Practice, *Journal of Chemical Education,* **89**(9), 1114-1121.

Brown, T. A., (2006), *Confirmatory factor analysis for applied research*, New York: Guilford Press.

Browne, M. W. and Cudeck, R., (1992), Alternative ways of assessing model fit, *Sociological Methods & Research,* **21**(2), 230-258.

Chou, C.-P., Bentler, P. M. and Satorra, A., (1991), Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study, *British Journal of Mathematical and Statistical Psychology,* **44**(2), 347-357.

Cohen, J., (1992), Statistical power analysis, *Current directions in psychological science,* **1**(3), 98-101.

Council, N. R. (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering. Committee on the Status, Contributions, and Future Direction of Discipline-Based Education Research. Board on Science Education, Division of Behavioral and Social Sciences and Education. S. R. Singer, N. R. Nielsen and H. A. Schweingruber. Washington D.C.

Creswell, J. W., (2013), *Qualitative inquiry and research design: Choosing among five approaches*, 3rd, Thousand Oaks, CA: Sage.

Dalgety, J. and Coll, R. K., (2006), Exploring first-year science students' chemistry self-efficacy, *International Journal of Science and Mathematics Education,* **4**(1), 97-116.

Dalgety, J., Coll, R. K. and Jones, A., (2003), Development of Chemistry Attitudes and Experiences Questionnaire (CAEQ), *Journal of research in science teaching,* **40**(7), 649-668.

Dalgety, J. and Salter, D., (2002), The development of the chemistry attitudes and experiences questionnaire (CAEQ), *Chem. Educ. Res. Pract.,* **3**(1), 19-32.

Dawes, J., (2008), Do data characteristics change according to the number of scale points used? An experiment using 5 point, 7 point and 10 point scales, *International Journal of Market Research,* **51**(1), 61-77.

Dweck, C. S., (1986), Motivational Processes Affecting Learning, *American psychologist,* **41**(10), 1040-1048.

Dweck, C. S., (2000), *Self-theories: Their role in motivation, personality, and development*, Psychology Press.

Dweck, C. S. (2012). Implicit Theories, Handbook of Theories of Social Psychology, P. V. Lange, A. Kruglanski and T. Higgins, London, Sage, **2**, 43-61.

Dweck, C. S. and Leggett, E. L., (1988), A social cognitive approach to motivation and personality, *Psychological review,* **95**(2), 256-273.

Dweck, C. S. and Sorich, L., (1999), Mastery-oriented thinking, *Coping*, 232-251.

Eccles, J. S. and Wigfield, A., (2002), Motivational beliefs, values, and goals, *Annual review of psychology,* **53**(1), 109-132.

Elliot, A. J., (1999), Approach and avoidance motivation and achievement goals, *Educational psychologist,* **34**(3), 169-189.

Goodman, S., Jaffer, T., Keresztesi, M., Mamdani, F., Mokgatle, D., Musariri, M., Pires, J. and Schlechter, A., (2011), An investigation of the relationship between students' motivation and academic performance as mediated by effort, *South African Journal of Psychology,* **41**(3), 373-385.

Gore, P. A., (2006), Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies, *Journal of Career Assessment,* **14**(1), 92-115.

Guay, F., Ratelle, C. F. and Chanal, J., (2008), Optimal learning in optimal contexts: The role of self-determination in education, *Canadian Psychology/Psychologie canadienne,* **49**(3), 233-240.

Gungor, A., Eryılmaz, A. and Fakıoglu, T., (2007), The relationship of freshmen's physics achievement and their related affective characteristics, *Journal of research in science teaching,* **44**(8), 1036-1056.

Hancock, G. R., (1999), A sequential Scheffé ‐ type respecification procedure for controlling type I error in exploratory structural equation model modification, *Structural Equation Modeling: A Multidisciplinary Journal,* **6**(2), 158-168.

Hansford, B. C. and Hattie, J. A., (1982), The relationship between self and achievement/performance measures, *Review of educational research,* **52**(1), 123-142.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M. and Elliot, A. J., (2000), Short-term and long-term consequences of achievement goals: Predicting interest and performance over time, *Journal of educational Psychology,* **92**(2), 316-330.

Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L. and Tauer, J. M., (2008), The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance, *Journal of educational Psychology,* **100**(1), 105-122.

Henson, R. K., (2001), Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha, *Measurement and evaluation in counseling and development,* **34**, 177-189.

Heredia, K. and Lewis, J. E., (2012), A Psychometric Evaluation of the Colorado Learning Attitudes about Science Survey for Use in Chemistry, *Journal of Chemical Education,* **89**(4), 436-441.

Hidi, S., (1990), Interest and its contribution as a mental resource for learning, *Review of educational research,* **60**(4), 549-571.

Hidi, S. and Baird, W., (1986), Interestingness- A neglected variable in discourse processing, *Cognitive Science,* **10**(2), 179-194.

Hidi, S. and Harackiewicz, J. M., (2000), Motivating the academically unmotivated: A critical issue for the 21st century, *Review of educational research,* **70**(2), 151-179.

Hidi, S. and Renninger, K. A., (2006), The four-phase model of interest development, *Educational psychologist,* **41**(2), 111-127.

Hong, Y. Y., Coleman, J., Chan, G., Wong, R. Y., Chiu, C. Y., Hansen, I. G., Lee, S. I., Tong, Y. Y. and Fu, H. Y., (2004), Predicting intergroup bias: The interactive effects of implicit theory and social identity, *Personality and Social Psychology Bulletin,* **30**(8), 1035-1047.

Hooper, D., Coughlan, J. and Mullen, M., (2008), Structural equation modelling: guidelines for determining model fit, *Electronic Journal of Business Research Methods,* **6**(1), 53-60.

Hu, L. T. and Bentler, P. M. (1995). Evaluating model fit, Structural equation modeling: Concepts, issues, and applications R. H. Hoyle, London, Sage, 76-99.

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

Huck, S., (2012), *Reading Statistics and Research*, 6th, Boston, MA: Pearson.

Hulleman, C. S., Durik, A. M., Schweigert, S. B. and Harackiewicz, J. M., (2008), Task values, achievement goals, and interest: An integrative analysis, *Journal of educational Psychology,* **100**(2), 398-416.

Hutchinson, S. R., Raymond, K. J. and Black, K. R., (2008), Factorial invariance of a campus climate measure across race, gender, and student classification, *Journal of Diversity in Higher Education,* **1**(4), 235-250.

Jones, B. D., Wilkins, J. L. M., Long, M. H. and Wang, F. H., (2012), Testing a motivational model of achievement: How students' mathematical beliefs and interests are related to their achievement, *European Journal of Psychology of Education,* **27**(1), 1-20.

Jöreskog, K. G. and Sörbom, D., (2006), LISREL 8.80 for Windows, *Computer Software. Lincolnwood, IL: Scientific Software International, Inc*,

Kline, R. B., (2011), *Principles and practice of structural equation modeling*, 3rd, New York: Guilford Press.

Knafl, K., Deatrick, J., Gallo, A., Holcombe, G., Bakitas, M., Dixon, J. and Grey, M., (2007), The analysis and interpretation of cognitive interviews for instrument development, *Research in Nursing & Health,* **30**(2), 224-234.

Lent, R. W., Brown, S. D. and Hackett, G., (1994), Toward a unifying social cognitive theory of career and academic interest, choice, and performance, *Journal of Vocational Behavior,* **45**(1), 79-122.

Lent, R. W., Brown, S. D. and Larkin, K. C., (1984), Relation of self-efficacy expectations to academic achievement and persistence, *Journal of counseling psychology,* **31**(3), 356.

Lent, R. W. and Hackett, G., (1987), Career self-efficacy: Empirical status and future directions, *Journal of Vocational Behavior,* **30**(3), 347-382.

Lewis, S. E., Shaw, J. L., Heitz, J. O. and Webster, G. H., (2009), Attitude Counts: Self-Concept and Success in General Chemistry, *Journal of Chemical Education,* **86**(6), 744.

Lichtenstein, M. J., Owen, S. V., Blalock, C. L., Liu, Y., Ramirez, K. A., Pruski, L. A., Marshall, C. E. and Toepperwein, M. A., (2008), Psychometric reevaluation of the scientific attitude inventory‐revised (SAI‐II), *Journal of research in science teaching,* **45**(5), 600-616.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A. and Harackiewicz, J. M., (2010), Measuring Situational Interest in Academic Domains, *Educational and psychological measurement,* **70**(4), 647-671.

Lopez, F. G. and Lent, R. W., (1992), Sources of mathematics self‐efficacy in high school students, *The Career Development Quarterly,* **41**(1), 3-12.

McCoach, D. B. and Siegle, D., (2003), Factors that differentiate underachieving gifted students from high-achieving gifted students, *Gifted Child Quarterly,* **47**(2), 144-154.

Multon, K. D., Brown, S. D. and Lent, R. W., (1991), Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation, *Journal of counseling psychology,* **38**(1), 30.

Murphy, K. R. and Davidshofer, C. O., (2005), *Psychological testing: Principles and applications*, New Jersey: Prentice Hall.

Nakhleh, M. B., (1992), Why some students don't learn chemistry: Chemical misconceptions, *Journal of Chemical Education,* **69**(3), 191-196.

Nicholls, J. G., Cheung, P. C., Lauer, J. and Patashnick, M., (1989), Individual differences in academic motivation: Perceived ability, goals, beliefs, and values, *Learning and Individual differences,* **1**(1), 63-84.

Pajares, F., (1996), Self-efficacy beliefs in academic settings, *Review of educational research,* **66**(4), 543-578.

Pajares, F. and Kranzler, J., (1995), Self-efficacy beliefs and general mental ability in mathematical problem-solving, *Contemporary educational psychology,* **20**(4), 426-443.

Pajares, F. and Miller, M. D., (1994), Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis, *Journal of educational Psychology,* **86**(2), 193-203.

Parker, P. D., Marsh, H. W., Ciarrochi, J., Marshall, S. and Abduljabbar, A. S., (2014), Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes, *Educational Psychology,* **34**(1), 29-48.

Pintrich, P. R., (2003), A motivational science perspective on the role of student motivation in learning and teaching contexts, *Journal of educational Psychology,* **95**(4), 667-686.

Pintrich, P. R. and De Groot, E. V., (1990), Motivational and self-regulated learning components of classroom academic performance, *Journal of educational Psychology,* **82**(1), 33-40.

Raker, J. R., Emenike, M. E. and Holme, T. A., (2013), Using Structural Equation Modeling To Understand Chemistry Faculty Familiarity of Assessment Terminology: Results from a National Survey, *Journal of Chemical Education,* **90**(8), 981-987.

Renninger, K. (2000). Individual interest and its implications for understanding intrinsic motivation, Intrinsic and extrinsic motivation: The search for optimal motivation and performance, C. Sansone and J. Harackiewicz, San Diego, CA, Academic Press, 373-404.

Renninger, K. and Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study, Development of achievement motivation, A. Wigfield and J. Eccles, San Diego, CA, Academic Press, 175-191.

Satorra, A. and Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis, Latent variables analysis: Applications for developmental research, A. von Eye and C. C. Clogg, Thousand Oaks, CA, Sage Publications, Inc, 399-419.

Schermelleh-Engel, K., Moosbrugger, H. and Müller, H., (2003), Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures, *Methods of psychological research online,* **8**(2), 23-74.

Schiefele, U., (1991), Interest, learning, and motivation, *Educational psychologist,* **26**(3-4), 299-323.

Schiefele, U., (1999), Interest and learning from text, *Scientific studies of reading,* **3**(3), 257-279.

Schunk, D. and Pajares, F. (2002). The development of academic self-efficacy, Development of achievement motivation, A. Wigfield and J. Eccles, San Diego, CA, Academic Press, 16-29.

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

Singh, K., Granville, M. and Dika, S., (2002), Mathematics and science achievement: Effects of motivation, interest, and academic engagement, *Journal of Educational Research,* **95**(6), 323-332.

Sorich, L. and Dweck, C. S. (1997). Reliability data for new scales measuring students' beliefs about effort and responses to failure. Unpublished raw data, Columbia University.

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

Stipek, D. and Gralinski, J. H., (1996), Children's beliefs about intelligence and school performance, *Journal of educational Psychology,* **88**(3), 397-407.

Taasoobshirazi, G. and Glynn, S. M., (2009), College Students Solving Chemistry Problems: A Theoretical Model of Expertise, *Journal of research in science teaching,* **46**(10), 1070-1089.

Tai, R. H., Sadler, P. M. and Loehr, J. F., (2005), Factors influencing success in introductory college chemistry, *Journal of research in science teaching,* **42**(9), 987-1012.

Tschannen-Moran, M. and Hoy, A. W., (2001), Teacher efficacy: capturing an elusive construct, *Teaching and Teacher Education,* **17**(7), 783-805.

Tuan, H. L., Chin, C. C. and Shieh, S. H., (2005), The development of a questionnaire to measure students' motivation towards science learning, *International Journal of Science Education,* **27**(6), 639-654.

Tucker, L. R. and Lewis, C., (1973), A reliability coefficient for maximum likelihood factor analysis, *Psychometrika,* **38**(1), 1-10.

Uitto, A., Juuti, K., Lavonen, J. and Meisalo, V., (2006), Students' interest in biology and their out-of-school experiences, *Journal of Biological Education,* **40**(3), 124-129.

Uzuntiryaki, E. and Aydin, Y. C., (2009), Development and Validation of Chemistry Self-Efficacy Scale for College Students, *Research in Science Education,* **39**(4), 539-551.

Villafane, 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, *Chemistry Education Research and Practice,* **15**(2), 114-127.

Weiner, B., (1985), An attributional theory of achievement-motivation and emotion, *Psychological review,* **92**(4), 548-573.

Wigfield, A. and Eccles, J. S. (2002). Introduction, Development of achievement motivation, A. Wigfield and J. S. Eccles, San Diego, CA, Academic Press, 1-10.

Wren, D. and Barbera, J., (2013), Gathering Evidence for Validity during the Design, Development, and Qualitative Evaluation of Thermochemistry Concept Inventory Items, *Journal of Chemical Education,* **90**(12), 1590-1601.

Xu, X. and Lewis, J. E., (2011), Refinement of a Chemistry Attitude Measure for College Students, *Journal of Chemical Education,* **88**(5), 561-568.

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

Zeldin, A. L., Britner, S. L. and Pajares, F., (2008), A comparative study of the self‐efficacy beliefs of successful men and women in mathematics, science, and technology careers, *Journal of research in science teaching,* **45**(9), 1036-1058.

Zimmerman, B. J., (2000), Self-efficacy: An essential motive to learn, *Contemporary educational psychology,* **25**(1), 82-91.

Zimmerman, B. J., Bandura, A. and Martinez-Pons, M., (1992), Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting, *American educational research journal,* **29**(3), 663-676.

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

Appendix

Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry

Brent Ferrell and Jack Barbera

Department of Chemistry and Biochemistry, University of Northern Colorado, Greeley, CO 80639
Email: jack.barbera@unco.edu

Scales and Demographic Items

Initial Interest Scale

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
A	B	C	D	E

II 1. I am fascinated by chemistry.	A	B	C	D	E
II 2. I chose to take general chemistry because I'm really interested in the topic.	A	B	C	D	E
II 3. I am really excited about taking this class	A	B	C	D	E
II 4. I am really looking forward to learning more about chemistry.	A	B	C	D	E
II 5. I think the field of chemistry is an important discipline.	A	B	C	D	E
II 6. I think that what we will study in General Chemistry will be important for me to know.	A	B	C	D	E
II 7. I think that what we will study in General Chemistry will be worthwhile for me to know.	A	B	C	D	E

Maintained Interest Scale

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
A	B	C	D	E

MI 1. What we are learning in chemistry class this semester is fascinating to me. A B C D E

MI 2. This semester, I really enjoy the chemistry material we cover in class. A B C D E

MI 3. I am excited about what we are learning in chemistry class this semester. A B C D E

MI R4. To be honest, I don't find the chemistry material we cover in class interesting. A B C D E

MI 5. What we are studying in chemistry class is useful for me to know. A B C D E

MI 6. The things we are studying in chemistry this semester are important to me. A B C D E

MI 7. What we are learning in chemistry this semester is important for my future goals. A B C D E

MI 8. What we are learning in chemistry this semester can be applied to real life. A B C D E

Self-Efficacy Scale

Very Poorly	Poorly	Average	Well	Very Well
A	B	C	D	E

SE 1. To what extent can you explain chemical laws and theories?	A	B	C	D	E
SE 2. How well can you choose an appropriate formula to solve a chemistry problem?	A	B	C	D	E
SE 3. How well can you describe the structure of an atom?	A	B	C	D	E
SE 4. How well can you describe the properties of elements by using the periodic table?	A	B	C	D	E
SE 5. How well can you read the formulas of elements and compounds?	A	B	C	D	E
SE 6. How well can you interpret chemical equations?	A	B	C	D	E
SE 7. How well can you interpret graphs/charts related to chemistry?	A	B	C	D	E
SE 8. How well can you solve chemistry problems?	A	B	C	D	E

Effort Beliefs Scale

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
A	B	C	D	E

EB 1R*. To tell the truth, when I work hard at chemistry, it makes me feel like I'm not very smart	A	B	C	D	E
EB 2R*. It doesn't matter how hard you work if you're not smart in chemistry, you won't do well in it.	A	B	C	D	E
EB 3R*. If you're not good at chemistry, working hard won't make you good at it.	A	B	C	D	E
EB 4R*. If chemistry is hard for someone, it means that he or she probably won't be able to do really well at it.	A	B	C	D	E
EB 5R*. If you're not doing well at chemistry, it's better to try something easier.	A	B	C	D	E
EB 6. When chemistry is hard, it just makes me want to work more on it, not less.	A	B	C	D	E
EB 7. If you don't work hard at chemistry and put in a lot of effort, you probably won't do well.	A	B	C	D	E
EB 8. The harder you work at chemistry, the better you will be at it.	A	B	C	D	E
EB 9. If a chemistry assignment is hard, it means I'll probably learn a lot doing it.	A	B	C	D	E

*R indicates item must be reverse-coded prior to analysis

Demographics

D1. Gender : **A**- Male **B** - Female

D2. Age : **A** - <18 **B** – 18-20 **C** – 21-23 **D** – 24-25 **E** - >25

D3. How many years have you been in college?

A – This is my first semester **B** – 1 yr. **C** – 2 yrs. **D** – 3 yrs. **E** - >3 yrs.

D4. Is this your first chemistry class in college? **A** - Yes **B** - No

D5. How long ago did you take high school chemistry?

A – I did not take chemistry in high school

B – 1 yr. ago **C** – 2 yrs. ago **D** – 3 yrs. ago **E -** > 3 yrs. ago

D6. What is your declared major?

A – Chemistry (including Forensics, Biochemistry, Teaching, or Pre-Health)

B – Other Science (Biology, Physics, or Mathematics)

C – Other (including Sports & Exercise Science, Nursing, Earth Science, Statistics)

D – Undeclared

D7. Would you be willing to participate in a 30-minute interview regarding your interest, effort beliefs, and self-efficacy about chemistry? These interviews help us to further understand your responses and how the items make sense to you so we can make improvements to the questionnaire.

A - YES B - NO

Table 1. Time 1 item-level descriptive statistics for initial sample (n = 373)

Item	Mean	Std. Dev.	Skew	Kurtosis
II1	3.59	0.95	-0.31	-0.33
II2	3.09	1.07	0.06	-0.68
II3	3.40	0.94	-0.10	-0.23
II4	3.81	0.86	-0.67	0.46
II5	3.97	0.79	-0.73	1.14
II6	4.11	0.88	-1.20	1.82
II7	4.09	0.81	-0.98	1.51
EB1R	3.21	1.15	-0.24	-0.77
EB2R	3.96	0.92	-0.86	0.45
EB3R	4.19	0.79	-1.15	2.07
EB4R	3.92	0.87	-0.90	0.76
EB5R	3.87	0.78	-0.44	-0.04
EB6	3.48	1.03	-0.52	-0.34
EB7	4.07	0.80	-1.10	1.98
EB8	4.22	0.76	-1.36	3.54
EB9	3.57	0.84	-0.25	-0.10
SE1	2.47	0.86	-0.04	-0.53
SE2	2.63	0.93	-0.04	-0.53
SE3	3.14	1.04	-0.01	-0.53
SE4	3.16	0.95	-0.15	-0.45
SE5	3.12	0.95	-0.04	-0.21
SE6	2.91	0.92	-0.08	-0.23
SE7	3.16	0.86	-0.14	0.16
SE8	2.94	0.89	-0.26	-0.14

II – Initial interest, EB – Effort beliefs, SE – Self-efficacy, R – indicates item has been reverse-coded

Table 2. Time 2 descriptive statistics for maintained interest items with initial sample (n = 294)

Item	Mean	Std. Dev.	Skew	Kurtosis
MI 1	3.32	.97	-.26	-.16
MI 2	3.23	.96	-.11	-.36
MI 3	3.15	.93	-.02	-.16
MI R4	3.20	1.15	-.24	-.96
MI 5	3.73	.96	-.50	-.23
MI 6	3.37	.97	-.18	-.29
MI 7	3.75	1.07	-.67	-.28
MI 8	3.56	.94	-.23	-.29

MFeel – Maintained interest (feeling), MVal – Maintained interest (value),
R – indicates item has been reverse-coded

Table 3. Mean scores and differences between chemistry majors and non-science majors on interest scales for the cross-validation study sub-sample

Scale	Chemistry majors n = 23 Mean score (SD)	Non-science majors n = 109 Mean score (SD)	Mean difference[a] (effect size[b])	p-value
Initial interest (feeling)	3.88 (0.75)	2.98 (0.83)	0.90 (0.70)	< 0.001
Initial interest (value)	4.50 (0.52)	3.67 (1.07)	0.83 (0.56)	<0.001
Maintained interest (feeling)	3.80 (0.85)	2.99 (0.92)	0.81 (0.56)	<0.001
Maintained interest (value)	3.85 (0.80)	3.30 (0.82)	0.55 (0.45)	0.004

[a]Based on planned contrasts [b]Effect size represented by *Cohen's d* – small (0.20), medium (0.50), large (0.80) (Cohen, 1992)

CFA model diagrams

The figures below show the revised models for each scale for each time point. In the following figures, indicators are represented with a boxed border and latent variables are represented with an oval border. Error terms are represented by arrows pointing toward the indicators. The factor loadings are the numbers between the latent variables and the indicators. Correlations between latent variables are indicated by the number next to the double arrows in between two factors.

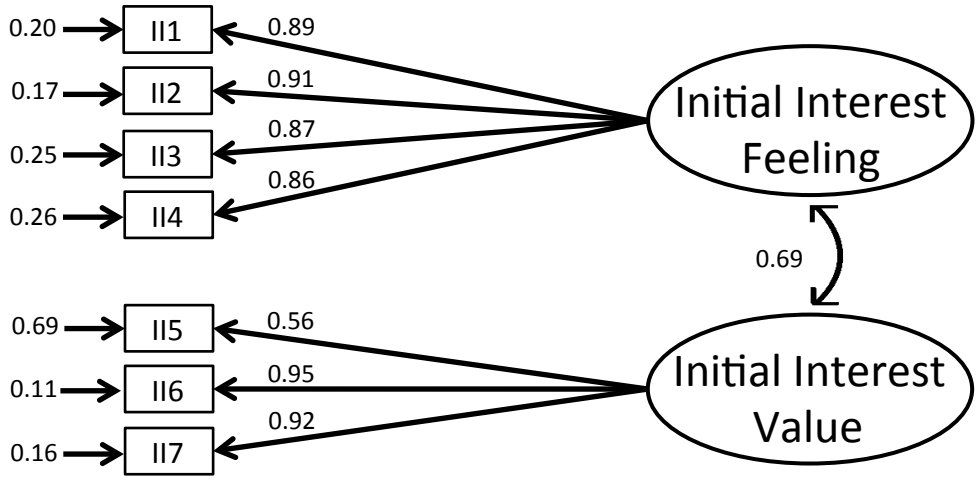

Figure 1. Time 1 initial interest CFA model for initial sample (n = 373)

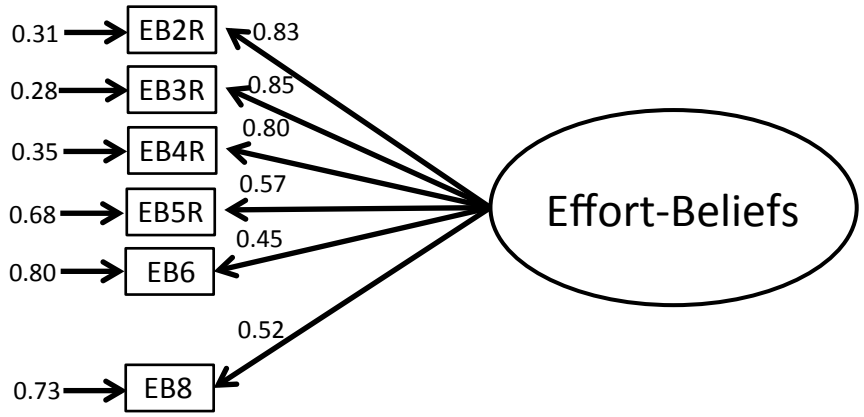

Figure 2. Time 1 effort beliefs CFA model for initial smaple (n = 373)

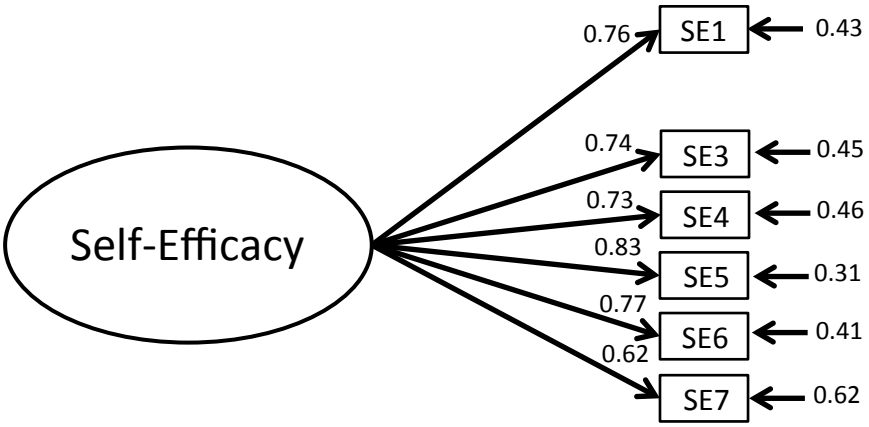

Figure 3. Time 1 self-efficacy CFA model for initial sample (n = 373)

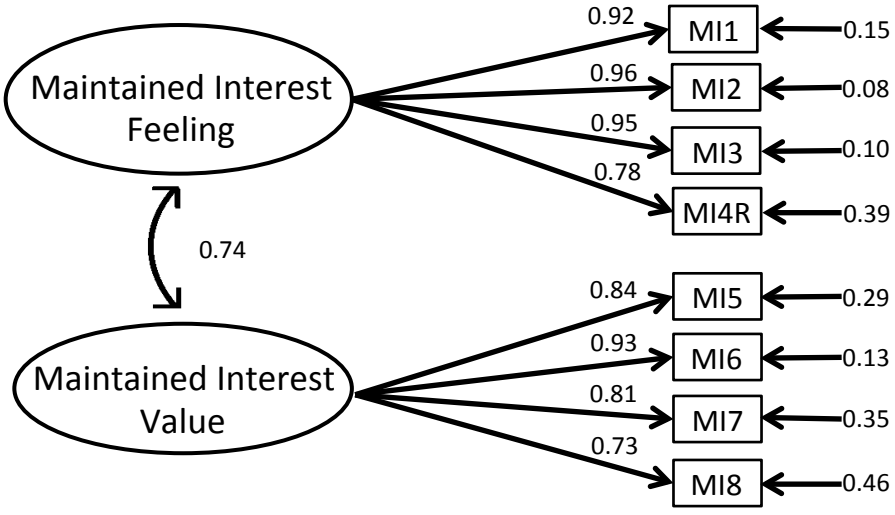

Figure 4. Time 2 maintained interest CFA model for initial sample (n = 294)

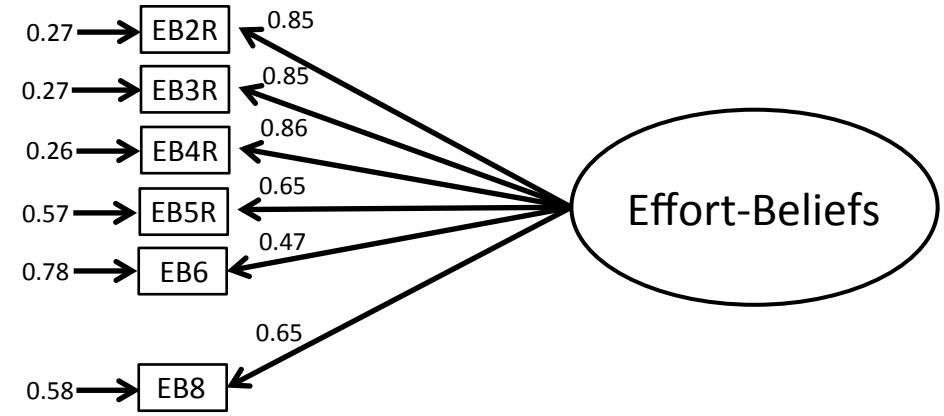

Figure 5. Time 2 effort beliefs CFA model for initial sample (n = 294)

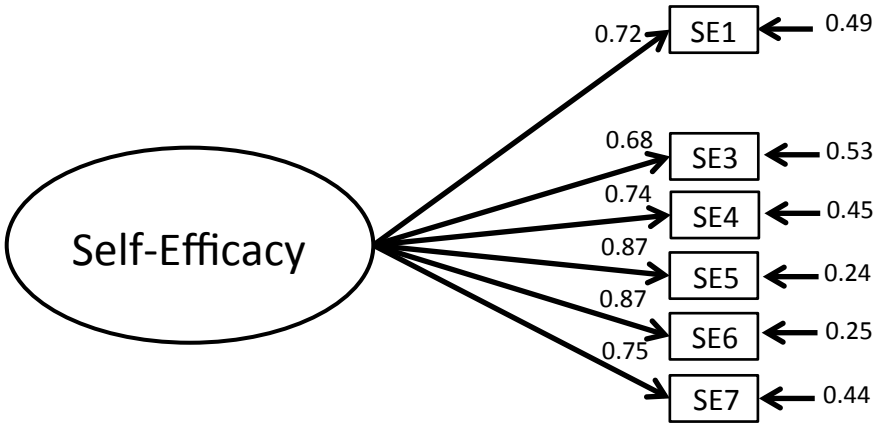

Figure 6. Time 2 self-efficacy CFA model for initial sample (n = 294)