SELF-EFFICACY STATE EXPERIENCES IN INTRODUCTORY PHYSICS: WITH IMPLICATIONS FOR GENDER IN PHYSICS

By

Jayson Micheal Nissen

Bachelor of Science, Montana State University, 2010

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Physics)

The Graduate School

The University of Maine

May 2016

Advisory Committee:

Jonathan T. Shemwell, PhD., Assistant Professor of Education, Advisor

MacKenzie R. Stetzer, PhD., Assistant Professor of Physics

Michael C. Wittmann, PhD., Professor of Physics

Robert W. Meulenberg, PhD., Associate Professor of Physics

Craig A. Mason, PhD., Professor of Education

DISSERTATION

ACCEPTANCE STATEMENT

On behalf of the Graduate Committee for Jayson Micheal Nissen, I affirm that this manuscript is the final and accepted dissertation. Signatures of all committee members are on file with the Graduate School at the University of Maine, 42 Stodder Hall, Orono, Maine.

Submitted for graduation in December 2015

12/21/15

Jonathan T. Shemwell, PhD., Assistant Professor of Education (Date)

LIBRARY RIGHTS STATEMENT

SELF-EFFICACY STATE EXPERIENCES IN INTRODUCTORY PHYSICS: WITH IMPLICATIONS FOR GENDER IN PHYSICS

By Jayson Micheal Nissen

Dissertation Advisor: Jonathan T. Shemwell

An Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Physics)
May 2016

Few undergraduates choose physics as a major, and among those who do very few are women. One potential contributor to this problem is the impact that physics instruction seems to have on students' self-efficacy, which is student's thoughts and feelings about their capabilities to succeed as learners in physics. Self-efficacy plays an important role in student achievement in academics both in general and for students pursuing STEM degrees. Conversely, research has shown that the self-efficacy of both men and women tends to be reduced after taking traditional and research-based physics courses. Moreover, self-efficacy tends to be reduced further for women than for men. Whether the negative shifts in self-efficacy in physics are caused by physics instruction remains unclear. It may be that the negative shift in self-efficacy reflects a broader trend in university education that has little to do with physics per se. I investigated this and other alternative explanations for negative shifts in self-efficacy in physics courses using an in-the-moment measurement technique called the Experience Sampling Method. The technique allowed me to collect students' day-to-day feelings of self-efficacy, which I called states, and to compare students'

self-efficacy states in physics to those in other STEM courses. I found that students experienced much lower self-efficacy states in physics than in their other STEM courses. Moreover, this difference largely affected women who experienced physics, and only physics, with much lower self-efficacy states than men. Given that experiences are an established sources of self-efficacy beliefs and women also had much more negative shifts in their self-efficacy beliefs I concluded that the experience of physics instruction was probably a causal factor in women's reduced self-efficacy. Further analysis found that the gender difference in self-efficacy states was more than twice that predicted by students' pre-course achievement, attitudes and beliefs. Thus I tentatively concluded that the negative impact on women's self-efficacy resulted from inequities in the physics-learning environment rather than preexisting gender differences. I present evidence that the physics course I investigated was similar to other research-based physics courses and tentatively I concluded that physics instruction in general is detrimental to women's self-efficacy.

SELF-EFFICACY STATE EXPERIENCES IN INTRODUCTORY PHYSICS: WITH IMPLICATIONS FOR GENDER IN PHYSICS

By Jayson Micheal Nissen

Dissertation Advisor: Jonathan T. Shemwell

A Lay Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Physics)
May 2016

Keywords: physics, self-efficacy, gender, experience sampling method, equity, research-based instruction

Few undergraduates choose physics as a major, and among those who do very few are women. One potential contributor to this problem is the impact that physics instruction seems to have on students' self-efficacy, which is student's thoughts and feelings about their capabilities to succeed as learners in physics. Self-efficacy plays an important role in student achievement in academics both in general and for students pursuing STEM degrees. Conversely, research has shown that the self-efficacy of both men and women tends to be reduced after taking traditional and research-based physics courses. Moreover, self-efficacy tends to be reduced further for women than for men. Whether the negative shifts in self-efficacy in physics are caused by physics instruction remains unclear. It may be that the negative shift in self-efficacy reflects a broader trend in university education that has little to do with physics per se. I investigated this and other alternative explanations for negative shifts in self-efficacy in physics courses using an in-the-moment measurement technique called

the Experience Sampling Method. The technique allowed me to collect students' day-to-day feelings of self-efficacy, which I called states, and to compare students' self-efficacy states in physics to those in other STEM courses. I found that students experienced much lower self-efficacy states in physics than in their other STEM courses. Moreover, this difference largely affected women who experienced physics, and only physics, with much lower self-efficacy states than men. Given that experiences are an established sources of self-efficacy beliefs and women also had much more negative shifts in their self-efficacy beliefs I concluded that the experience of physics instruction was probably a causal factor in women's reduced self-efficacy. Further analysis found that the gender difference in self-efficacy states was more than twice that predicted by students' pre-course achievement, attitudes and beliefs. Thus I tentatively concluded that the negative impact on women's self-efficacy resulted from inequities in the physics-learning environment rather than preexisting gender differences. I present evidence that the physics course I investigated was similar to other research-based physics courses and tentatively I concluded that physics instruction in general is detrimental to women's self-efficacy.

DEDICATION

To the woods and waters of Maine and the people that shared them with me.

ACKNOWLEDGEMENTS

I would like to thank my wife for both the moral and intellectual support she has provided me during my work on this research. I would also like to thank my good friend Kevin Roberge. Many of the ideas in this dissertation were worked out on our many runs and bike rides through the university forest.

All of the members of the Physics Education Research Laboratory were also critical for me to reach this point both in terms of their moral support as my friends and their intellectual support as my colleagues. I am grateful for the feedback and support that my committee members have provided me.

I am also indebted to Jennifer Schmidt for replying to my cold call at the beginning of this research. Her enthusiasm and support for my use of the ESM was essential for me to figure out the nuances of the methodology that allowed me reach this point.

I also want to acknowledge my advisor Jon Shemwell. Who has supported me in pursuing these unorthodox questions I have posed, and helped me to develop the skills and knowledge to pursue them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
INTRODUCTION

A central goal in the development of research-based teaching practices in physics is to support students in developing expert attitudes and beliefs about learning and doing physics (Redish et al., 1998; Brewe et al., 2009; Adams et al., 2006; Etkina and Van Heuvelen, 2007). Within the broad category of attitudes and beliefs are students' beliefs in their ability to succeed in learning and doing physics, what scholars (Bandura, 1997) call physics self-efficacy. Self-efficacy has received recent attention by physics education researchers because of the important role it plays in students' agency and success in learning (Kost-Smith, 2011; Sawtelle et al., 2010, 2012a,b). The benefits of self-efficacy have been identified in cognitive tasks where increasing self-efficacy causes improved cognitive performance (Bouffard-Bouchard, 1990). Self-efficacy is also related to long-term outcomes. Self-efficacy in a content domain correlates with taking harder courses in that domain (Betz and Hackett, 1983) and increased academic success (Zajacova et al., 2005; Williams and Williams, 2010; Pietsch et al., 2003). In college, self-efficacy is correlated with the major students choose (Betz and Hackett, 1983; Marra and Bogue, 2009) and increasing students' self-efficacy increases their long-term interest in pursuing STEM degrees (Luzzo et al., 1999). The relationship between self-efficacy and desirable outcomes also extends to introductory physics courses, which are the focus of this study. S'elf-efficacy is correlated with performance on course exams (Kost-Smith, 2011) and with passing introductory physics courses (Sawtelle et al., 2012a). Therefore it stands to reason that increasing physics self-efficacy is an important educational outcome that contributes to many other desirable outcomes.

Self-efficacy belongs to the broader category of educational outcomes consisting of positive affect toward STEM domains. The study of affect includes students' attitudes, beliefs, values, feelings and emotions. Declining STEM enrollments and increasingly negative attitudes toward STEM suggest that positive affect is in short supply within STEM instruction (Osborne et al., 2003). One aspect of affect in STEM learning, students' attitudes toward learning and doing science, consistently decreases from pre to post instruction in STEM courses including physics (Adams et al., 2006), chemistry (Adams et al., 2008) and biology (Semsar et al., 2011). This decrease is problematic given that developing students' attitudes towards doing and learning science is a core affective goal of science education (Brewe et al., 2009) and these attitudes are an important predictor of student success in STEM fields (Redish et al., 1998; Perkins et al., 2004; Madsen et al., 2015). The positive relationship between attitudes and student success also extend to other affective constructs that predict student achievement in STEM, including self-efficacy (Multon et al., 1991; Lee, 2009; Marra and Bogue, 2009), interest (Koller et al., 2001) and motivation (Singh et al., 2002; Mujtaba and Reiss, 2013). Given that college is where students ultimately commit to STEM majors and many STEM majors leave their degrees in spite of being fully capable of success (Seymour and Hewitt, 1997) affective traits are important outcomes of introductory STEM courses. As Redish et al. (1998) pointed out, the transition to college courses introduces an ideal opportunity for developing students' attitudes and beliefs about doing and learning science. Yet, the negative shifts in students' affective traits in STEM courses indicate that this opportunity is not being capitalized on.

Affect and cognition are not independent, instead they are interrelated (Pintrich et al., 1993). These interrelationships have been demonstrated both in terms of short-term and long-term outcomes for self-efficacy. In a controlled experiment Bouffard-Bouchard (1990) found that increasing students' self-efficacy caused students to set higher standards for themselves, use more efficient strategies and achieve greater intellectual performance. This causal relationship between self-efficacy and performance is likely a reciprocal causal relationship that goes in both directions (Williams and Williams, 2010). Williams and Williams (2010) found that greater self-efficacy leads to greater performance that, in a self-reinforcing cyclic process, then leads to increased self-efficacy. In addition to these direct benefits for learning and performance, Luzzo et al. (1999) used an experiment to demonstrate that increasing students' self-efficacy for science and mathematics increased their interest in pursuing science and mathematics careers.

My purpose in this research follows from the clash between the importance of self-efficacy in introductory physics and the absence of physics instruction that supports students in developing self-efficacy. The research I previously described indicates that not only is self-efficacy a predictor of important student outcomes it is also a causal agent in students' cognition, performance and interest. Yet, students' self-efficacy tends to decrease from pre to post instruction in introductory physics courses (Sawtelle et al., 2010; Lindstrom and Sharma, 2011; Kost-Smith, 2011) or at best not change at all (Sawtelle et al., 2010; Cavallo et al., 2004). My purpose in the investigation presented in this dissertation was to investigate the extent to which the physics-learning environment was a cause of the shifts in students' self-efficacy. If the physics-learning environment is a cause a negative shift in self-efficacy then physics educators, researchers and policy makers will be confronted with the need to develop physics instruction that addresses this detrimental impact on students. Supporting students in developing self-efficacy is important because it supports students in doing better both in the short term through increased cognitive performance, more efficient strategy use and setting higher goals and in the long term by supporting student course performance and supporting students in choosing STEM majors.

To approach this investigation I distinguished between two different measures of self-efficacy, states and traits. Self-efficacy traits are the general pattern of self-efficacy that a student tends to experience in physics. Self-efficacy states are the in-the-moment experience of self-efficacy in the activity at hand. For example, a physics graduate student likely has relatively high self-efficacy traits in physics. Therefore, that graduate student will tend to experience relatively high self-efficacy states while learning physics content, but they will also likely experience very low self-efficacy states at times. I labeled the measure of self-efficacy that retrospective surveys provide as traits because they measured students' general tendency to feel efficacious in physics activities and they were relatively stable, tending to change only small amounts over long periods of time. I developed a direct measure of students' subjective feelings of capability that arose in the process of learning physics. I labeled this direct measure self-efficacy states because it varied large amounts over short time frames and was specific to the activity at hand.

In the rest of this chapter I motivate my investigation of self-efficacy from what is known about self-efficacy in introductory physics courses. In reviewing the prior research I focus on what is known about gender differences in self-efficacy since gender differences in self-efficacy exist both in general and in physics. These gender differences in self-efficacy motivated me to focus on gender as a central variable in my investigation. I situate the known gender differences in physics self-efficacy within the known gender differences for a broad array of student outcomes for: conceptual learning, attitudes, experience and the under-representation of women in physics. The gender differences in experience in physics have largely been measured using retrospective surveys or retrospective interviews administered distantly from the actual experience. This distance between experience and measurement motivated my development of a direct measure of self-efficacy states because the passage of time may introduce inaccuracies in student retrospections on their experiences.

Additionally, because my self-efficacy state construct was novel, I also motivate the need for this measure from the broader literature on self-efficacy. To motivate the validity and utility of the self-efficacy state construct, I review the literature on the experience sampling method (ESM), which I adapted to collect self-efficacy states. At the conclusion of this chapter I provide an overview of the structure of the my dissertation that follows from the purpose, motivation and goals that I developed based on the extant research that I review in this chapter.

1.1 Self-efficacy in physics

A major problem is that physics frequently undermines or at best fails to support students' self-efficacy. Decreases in students' self-efficacy has been measured across a wide range of introductory physics courses for majors and for non-majors. In the largest investigation of self-efficacy in introductory physics, Kost-Smith (2011) found consistently negative shifts in self-efficacy traits for 2,155 students in 6 semesters of physics instruction, 3 first semester and 3 second semester courses. The courses Kost-Smith (2011) investigated implemented research-based teaching practices and were calculus-based. Sawtelle et al. (2010) found negative shifts in self-efficacy traits for 175 students in 3 semesters of a traditional lecture-based physics course and neutral shifts in self-efficacy traits for 70 students in 3 semesters of a research-based physics course. Both of the courses that Sawtelle et al. (2010) investigated were calculus-based. Cavallo et al. (2004) found neutral shifts in self-efficacy traits for 196 students in an inquiry-based physics course using algebra-based instruction. They measured self-efficacy across a full year of instruction. Lindstrom and Sharma (2011) found negative shifts for 203 students in two first semester physics courses and neutral shifts for 124 students in one second semester physics course. These courses implemented some research-based teaching practices and were algebra-based.

I found no studies reporting positive shifts in physics self-efficacy traits. It is difficult to compare the results in these studies because they investigated different types of courses with different instruments and some combined first and second semesters of instruction while others did not. Nonetheless, the fact that none of these studies found positive shifts in students' self-efficacy suggests that physics is failing to support the development of self-efficacy. Furthermore, the lecture-based and research-based instruction that negatively impacted students self-efficacy traits (Kost-Smith, 2011; Sawtelle et al., 2010) are two of the most prevalent pedagogies for calculus-based physics instruction. These findings indicate that common forms of physics instruction are harmful to students' belief in their ability to learn and do physics. My central purpose to link the shifts in students' self-efficacy to the physics-learning environment was motivated by these consistent negative shifts in students' self-efficacy in physics courses.

The research on self-efficacy in physics has focused solely on physics and has not also measured self-efficacy in other STEM domains; therefore, it has not ruled out the possibility of a general trend of decreases in self-efficacy in introductory STEM courses. Thus, one possible alternative explanation for the negative shift in physics self-efficacy is that it is an epiphenomenon of a general trend throughout STEM education. There have been several investigations of self-efficacy in other STEM domains and I review these studies in Chapter 4. Similar to the studies of self-efficacy in physics, studies in other domains have studied a wide range of courses (research-based, lecture-based, for majors, for non-majors, etc.) with a wide range of instruments and with varying levels of reporting on their findings. This variability makes it difficult to directly compare between these studies. Direct comparisons between STEM domains are much easier to make in terms of the shifts in students' attitudes about learning and doing physics since the Colorado Learning Attitudes about Science Survey (CLASS) has been adapted for, and used in, multiple STEM domains. While the CLASS reports on attitudes, not self-efficacy, the two are interrelated since self-efficacy falls within the broader scope of attitudes. Thus, a general negative trend in self-efficacy across STEM domains is supported by the decreases in students' attitudes measured in many introductory STEM courses including physics (Adams et al., 2006), chemistry (Adams et al., 2008) and biology (Semsar et al., 2011). If there were a general trend, however, this would not appear to be overly strong given that some of the studies in physics found neutral shifts in students' self-efficacy.

Locating a cause of the shift in self-efficacy traits within the physics-learning environment follows from an assumption that self-efficacy traits in a domain develop from the experiences that students have in that domain. This relationship between experience and self-efficacy traits comes directly from the theoretical basis of self-efficacy (Bandura, 1997). For example, one of the most important sources of self-efficacy traits are *mastery experiences*, experiences of accomplishment and success that result in the elevation of an individual's assessment of his or her abilities in a specific domain (Bandura, 1997; Zeldin et al., 2008; Zeldin and Pajares, 2000; Betz and Schifano, 2000; Luzzo et al., 1999). The experimental study of Luzzo et al. (1999) demonstrated that experiences are causal relationships in the development of self-efficacy traits. Their design used four treatment conditions and the students in the conditions that received mastery experiences developed and maintained higher self-efficacy than those in the control group and than those that received the less powerful vicarious experiences.

Given the importance of mastery experiences in the development of self-efficacy traits, it seems reasonable that increasing the effectiveness of instruction would improve self-efficacy development by providing more opportunities for mastery experiences. Subsequently, I would expect courses implementing research-based teaching practices with documented increased conceptual learning to support students in developing self-efficacy traits. Unfortunately, courses with relatively high levels of conceptual learning do not support students in developing self-efficacy traits (Kost-Smith, 2011; Sawtelle et al., 2010), though in the research-based courses investigated by Sawtelle et al. (2010) students' self-efficacy traits did not decrease. This consistent lack of positive shifts in physics self-efficacy traits raises the questions of what self-efficacy states are students experiencing during physics instruction and how do those states compare to those experienced in similar activities, such as other STEM courses?

1.2 Gender differences in physics self-efficacy

Female students tend to start physics instruction with lower self-efficacy traits than male students and tend to have larger negative shifts in their self-efficacy traits resulting in the gender differences tending to increase (Kost-Smith, 2011; Sawtelle et al., 2010; Lindstrom and Sharma, 2011; Cavallo et al., 2004). This difference in shifts in self-efficacy traits indicates that there is likely a difference in the self-efficacy states men and women experience during physics instruction since the development of self-efficacy is rooted in experience (Bandura, 1997; Bouffard-Bouchard, 1990; Luzzo et al., 1999). It is possible, however, that these gender differences in self-efficacy traits do not reflect gender differences in experiences since retrospective measures do not always accurately reflect actual gender differences in experience (Goetz et al., 2013). They named this difference between retrospective measures and experiences a state-trait gap. Thus, a possible explanation of the gender differences in self-efficacy traits is that it arises due to differences in retrospection rather than differences in experience. Furthermore, if there are gender differences in the self-efficacy states students experience in physics it is possible that these differences are an epiphenomenon of larger trend in introductory STEM courses. Since prior studies of self-efficacy in physics (Kost-Smith, 2011; Sawtelle et al., 2010; Lindstrom and Sharma, 2011; Cavallo et al., 2004) have only investigated physics instruction they cannot rule out the possibility of gender differences in self-efficacy traits being common throughout introductory STEM courses. These two alternative explanations (epiphenomenon and state-trait gap) for the gender difference in physics self-efficacy traits must be ruled out before a cause of a larger negative shift in women's physics self-efficacy traits can be squarely placed within the physics-learning environment.

The gender differences in self-efficacy traits in physics are part of a larger trend of gender differences in affect and achievement within introductory physics courses. In introductory physics courses women tend to start instruction with lower levels of both conceptual knowledge (Madsen et al., 2013; Kost et al., 2009a) and attitudes about learning and doing physics (Kost et al., 2009a). Therefore it is possible that any gender differences in the self-efficacy states that students experienced while learning physics are mostly a result of the traits (conceptual knowledge, attitudes and self-efficacy traits) that students started instruction with. If this is the case then the physics instruction could be said to fairly treat men and women (Rodriguez et al., 2012) since it did not have an effect on either group beyond what was predicted by pre-existing differences. I further discuss equity and fairness in Chapter 6. If the gender differences in the self-efficacy states that students experienced were much larger than those predicted by students' pre-course traits then this would clearly be inequitable since the physics-learning environment negatively impacted women more than men. If college physics instruction is inequitable then much of the onus for addressing gender differences in self-efficacy states and traits is located within the environment of college physics instruction and it would therefore be the responsibility of the educators, researchers and administrators at the college level to address this inequity.

1.3 Directly measuring self-efficacy in the process of learning physics

One significant challenge for investigating the relationship between students' experiences in physics and their development of physics self-efficacy traits is that there are very few methods for measuring self-efficacy states in a direct way. Most studies investigating the development of self-efficacy have indirectly measured experiences that support the development of self-efficacy using retrospective measures, primarily in the form of quantitative surveys and to a lesser extent interviews (Sawtelle et al., 2012a). Because retrospective measures are generally reported long after the experience, they primarily access portions of that experience that students were consciously aware of. Thus, retrospective measures may miss many important features of experience that support self-efficacy trait development that students were not consciously aware of. An example of these subtle aspects of experience is receiving non-verbal feedback from other students in the form of laughter, smiling or nodding, which may support students' self-efficacy trait development (Sawtelle et al., 2012a). Furthermore, retrospective survey measures are inherently biased by the passage of time. Goetz et al. (2013) reported on this bias in the form of a gap between experience and retrospections on experience, referred to as a state-trait gap, for anxiety in mathematics. They concluded that two populations of students could have similar experiences, but who reflect on those experiences differently due to differences in their attitudes about the domain of that experience. Given that there are pre-existing gender differences in students' attitudes about learning and doing physics (Kost et al., 2009a) it is possible that gender differences in retrospective self-efficacy trait measures do not actually represent gender differences in self-efficacy state experiences.

I approached measuring self-efficacy from two distinct directions, states and traits. I separated self-efficacy between states and traits in response to the con-

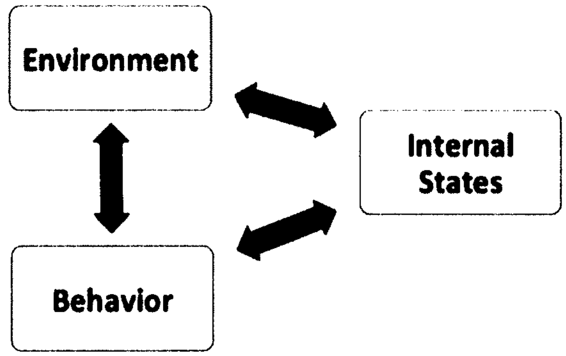


Figure 1.1: The three majors classes of determinants. These classes are proposed by Social Cognitive Theory. The arrows represent the reciprocal causal relationships that exist between each of the classes.

tradiction of measuring self-efficacy, which is "a dynamic fluctuating property, not a static trait" (Bandura, 1997, p. 480), with surveys administered only once or twice a semester that treat self-efficacy as a relatively static trait. Drawing on Bandura's definition of self-efficacy as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, p. 3) I defined self-efficacy states as being *dynamic judgments* of one's ability to organize and execute the courses of action required to produce given attainments *in the activity at hand.* The concept of states in this definition bears unpacking. Bandura (1986, 1997) proposed that internal states are one of the three major classes of determinants in human agency, along with behavior and environment, as shown in Figure 1.1. According to Bandura (1997), states arise within the individual, have a complex latent structure (consisting of affect, cognition and biological events) and are dynamically responsive to both the perceived environment and the individual's behavior. Self-efficacy states lie within this latent structure, but self-efficacy is likely a central causal force that affects the other aspects of the internal states, behavior and perceived environment.

Traits are the relatively stable and automatic patterns of internal states, including thoughts, feelings, and behaviors that habitually occur in different circumstances and contexts (Jackson et al., 2012). As Jackson et al. (2012, p. 746) point out, "From this definition it is important to note that personality traits reflect more than just behaviors; traits also reflect thoughts (cognitions) and feelings (emotions, affect)." This definition of traits is consistent with the description of self-efficacy as context dependent, tending to be very stable and resulting in habitual patterns of behavior (Bandura, 1997, 2006). Therefore, I defined self-efficacy traits as the *relatively stable pattern of judgments* of one's ability to organize and execute the courses of action required to produce given attainments *in a domain.* Following from these definitions I propose thinking of traits as representing the patterns that arise between the three major classes of determinants: internal states, environment and behavior.

One distinction between states and traits is the emphasis on who is responsible for them. Traits by definition belong to the individual and reside within the individual. In contrast, states reside within the individual but arise in responses to both the individual and the environment. In education settings this distinction could be critical for empowering educators to take action to change the status quo given that states emphasize the role of the environment and the learning environment is, within limits, the responsibility of the educator.

1.4 Measuring self-efficacy states

I measured self-efficacy states using the Experience Sampling Method (ESM). Participants in the ESM report their affect as it is happening, or immediately after, by answering a brief survey when signaled. Signals are either random throughout a person's day or situation-dependent. For example, Goetz et al. (2013) used situation dependent signaling to collect data only during mathematics courses. The surveys are designed to measure several aspects of experience including environmental characteristics (what participants were doing, where they were, who they were with), thoughts and feelings. The surveys are very short, typically taking one to three minutes to fill out, and they are meant to reflect feelings and emotions occurring at the time of the signal. Experiences are normally sampled throughout a person's day so that experiences of interest, such as feelings of self-efficacy in physics class, can be compared to other daily experiences. Data collection normally continues for several days so that target and background experiences are sampled multiple times for each person. The scale of ESM studies can vary greatly in terms of the number of participants, the number of samples collected and the length of data collection. For example, ESM studies vary from focusing on a single person for one week (Hektner et al., 2007) to the Sloan Survey of Youth and Social Development (SSYSD), which collected data from a total of 4,816 students at 33 public secondary schools over five years.

Because the ESM can collect experiences throughout daily life, it is particularly useful in identifying the environmental characteristics that correspond with affective components of experience. Shernoff et al. (2003) used data from the SSYSD to show that students were much more likely to have high levels of engagement in collaborative group work than in lecture. Lecture, however, made up the majority of activities students engaged in and collaborative group work was much less common. Separate studies by Johnson (2009) and Rathunde (2003) utilized similar methods to compare experiences of students attending non-traditional middle and high schools to those in the traditional public schools investigated in the SSYSD. They found that students in the non-traditional schools experienced high levels of engagement far more often than those in the traditional public schools in part because they spent more time in collaborative group work and much less time in passive activities such as lecture and watching media.

Despite the usefulness and availability of the ESM, very few studies have investigated the contextual dependencies of college learning. One exception is Goetz and colleagues' (Goetz et al., 2009) investigation of Control-Value Theory with 50 college freshmen over a one week period. In their review of and guide to using the ESM Hektner et al. (2007) pointed out only two studies using the ESM in college learning compared to many uses in secondary education. As Pekrun and Stephens (2010) pointed out the college classroom and the secondary classroom have different goals, different populations of students and different expectations that require both environments to be independently investigated. Further Pekrun and Stephens (2010) pointed out that very little research has been done on students affective experiences in college learning environments. The lack of studies in college settings indicates that there is a need to investigate affective experience with more direct measures like the ESM in college level learning.

The ESM has also been used to look at how emotions interact and how this interaction varies across contexts. Several studies have (Shernoff, D., Csikszentmihaly, M., Schneider, B., Shernoff, 2003; Johnson, 2009; Rathunde, 2003) used the theoretical framework of Flow Theory and focused on affective engagement, which they defined as a combination of concentration, interest and enjoyment. Goetz et al. (2009) used the ESM to investigate the role of the emotions of control and value in predicting positive and negative affect in educational settings to test Pekrun's (Pekrun, 2006) Control-Value Theory. In one of the few ESM studies to investigate college students' experiences, Goetz et al. (2009) collected data from 50 freshman university students throughout their daily lives over one week. They found that school was experienced with much less enjoyment and contentment than non-school activities and that much of this difference was explained by the control and value that students experienced in school activities. They also found that the control-value interaction was a statistically significant predictor of the positive affect that students experienced. This interaction indicated that together high states of control and value lead to greater positive affect than would be expected from only the independent contributions of control and value. These sorts of findings illustrate the potential usefulness of the ESM for measuring the environmental and affective characteristics of students' experiences. Despite these examples of the ESM's utility in developing affective theories, in my review of previous studies either using the ESM or investigating self-efficacy I found only one study that used the ESM to investigate self-efficacy states.

Schmidt and Shumow (2012) investigated students' efficacy experience in relation to teachers' practices and beliefs in secondary-school science classrooms. They collected data from 228 students in 11 science classrooms across a five-day school week. Their study was focused solely on the science classroom and did not collect ESM data outside of the science class. This focus left open the question about how experiences in science learning compared to learning in other domains. They supplemented ESM data with self-efficacy surveys, classroom video and teacher interviews in an attempt to trace the development of science self-efficacy traits. Survey results indicated that some classrooms supported self-efficacy trait development while others undermined it. The differences in self-efficacy trait development were related to teachers' practices, such as supporting students in participating in class, and teacher's beliefs about the gendered nature of science. In one course where the teacher was more supportive and responsive to male students than to female students and viewed science as a male domain, a gender difference in science self-efficacy traits resulted such that male students' self-efficacy traits increased while female students' decreased. In contrast, the courses where the teacher did not express a belief in science being gendered or behave in gender preferential ways gender differences in science self-efficacy traits did not develop. Schmidt and Shumow (2012) did not investigate a numerical relationship between efficacy experiences and science self-efficacy trait development leaving open the question of the strength of this relationship. Instead, they illustrated this relationship with two examples: one of a student whose self-efficacy improved and one whose declined. The student who improved experienced higher self-efficacy states as the course progressed and the student who declined experienced lower self-efficacy states as the course progressed. The two example case studies provide evidence of a relationship between the self-efficacy states that students experience and the self-efficacy traits that they develop. Furthermore, the relationships between teachers' beliefs and actions and the self-efficacy traits that students develop demonstrate the importance of the environment in shaping students' self-efficacy states and traits.

1.5 Structure of this dissertation

I developed four design goals to support my research in achieving my central purpose of investigating if the physics-learning environment was a cause the negative shift in students' self-efficacy traits. I describe these goals along with the design and methods used in this investigation in detail in Chapter 2 and I use them here to layout the structure of the rest of the dissertation. A central focus and contribution of this study was my adaptation of the ESM as a novel approach to measuring self-efficacy states. Because this approach has not been used in other studies I checked the internal and external validity of the self-efficacy state measure, design goal three, throughout the dissertation. I present my development of the self-efficacy state measure and the bulk of the investigation into its validity in Chapter 3. I present the quantitative analysis I used to verify the internal validity of the self-efficacy state measure. The first and second design goals were to determine whether negative effects on self-efficacy traits can be detected as a depression of self-efficacy states that occurred within the physics-learning environment and to investigate likely alternative causes of depressed self-efficacy states in physics. The most prominent alternative explanation was that depressed self-efficacy states were not unique to physics but were instead part of a broader trend in college learning. These purposes were explored in Chapters 4, 5 and 6. In Chapter 4 I present my investigation of the differences in self-efficacy states in physics compared to in other STEM courses for all students. In Chapter 5 I present my investigation of the role of gender in students' self-efficacy states in physics compared to in other STEM courses. This comparison allowed me to investigate the extent to which gender differences were an epiphenomenon of a general trend in STEM courses or alternatively the extent to which gender differences in traits did not reflect gender differences in experience but were instead a result of differences in retrospection.The format of Chapter 5 varies from the other chapters because it has been accepted for publication in *Physical Review Special Topics - Physics Education Research* . In Chapter 6 I present my investigation of the role of students' pre-course traits in the self-efficacy states that they experienced in physics. The fourth design goal was to obtain evidence of the extent to which the results would generalize to physics instruction beyond my particular context. In Chapter 5 I also present my investigation into the generalizability of my findings to other physics courses, design goal four. Finally in Chapter 7 I summarize my results and draw conclusions about their likely causes, implications for teaching and directions for future research.

Chapter 2
DESIGN AND METHODS OF THE INVESTIGATION

In this chapter I lay out the four design goals that supported my investigation into the cause of the shifts in students' physics self-efficacy traits. I describe comparisons that supported achievement of these four design goals and my reasoning for collecting the specific state and trait measures to make these comparisons. Next, I describe the methods for collecting, processing and analyzing the data I used in this study. Then, I lay out who the participants in the study were and describe the physics course that I investigated. Lastly, I summarize the questions that were investigated and the methods I used to investigate them in Chapters 3 through 6. Finally, in this chapter I provide a general overview of the methods used in my investigation. The details of the methods used to answer specific research questions are presented in the various chapters that address those components of the investigation (i.e. Chapters 3 through 6).

2.1 Design

Broadly, the study used a within-subject design to investigate the relationship between students' self-efficacy states and traits while learning physics. This design allowed measuring students' self-efficacy states with the same instrument in multiple settings, which facilitated the comparison of the experience in physics to that in other STEM courses for the same students. This design also allowed investigating the relationship between self-efficacy and gender while controlling for other variables. These design features followed from the central purpose of this dissertation, which was to investigate whether the physics-learning environment is a cause of the reduction in students' self-efficacy traits that has been observed in many physics courses. To that end I drew on *Causal Modeling* by Asher (1983) to develop four design goals.

1. Determine whether negative effects on self-efficacy traits can be detected as a depression of self-efficacy states that occurred within the physics-learning environment.

2. Investigate likely alternative causes of depressed self-efficacy in physics, most prominently the explanation that depressed self-efficacy states would not be unique to physics but instead part of a broader trend in college learning.

3. Check the validity for the self-efficacy state measure.

4. Obtain evidence of the extent to which the results would generalize to physics instruction beyond my particular context.

These goals were met through the coordination of state and trait data. State data was obtained from students' in-the-moment experiences such as their self-efficacy states. Trait data consisted of students' self-efficacy traits, conceptual knowledge, attitudes and course grades. I measured the shift in students' self-efficacy traits from pre to post instruction in order to locate any effect on students' self-efficacy as happening across instruction. I reasoned that if I had only measured self-efficacy traits post instruction, and not also pre-instruction, any gender differences in those traits could not be said to be a result of physics instruction because they could have existed prior to physics instruction rather than developing as a result of physics instruction. I measured self-efficacy states in order to identify any potential effect within the physics-learning environment. Linking the effects measured by the state and trait measures required investigating the validity of the relationship between these two measures.

The major alternative explanation for negative shifts in students' self-efficacy traits was that they were not a unique effect of physics instruction but rather an epiphenomenon of a larger trend throughout introductory college STEM courses. This possibility was the motivation for collecting self-efficacy state data in STEM and non-STEM courses to compare to physics. Obviously, the possibility of an epiphenomenon would be unlikely if students experienced lower self-efficacy states in physics than in other STEM courses. Furthermore, I collected state data from throughout participants' full waking day, which included recreational, homework and off campus time. This broad data collection provided context for interpreting students' experiences in physics and allowed comparing self-efficacy states in physics to those experienced in other STEM courses. Specifically, I used the difference between self-efficacy states in school compared to non-school activities as a benchmark for very large differences.

There were three other alternative explanations for the self-efficacy trait effect that I investigated. The first was applicable to investigating the effect of physics instruction on self-efficacy both with and without taking gender into account. The remaining two are only applicable to gender differences. The first was that the shift in self-efficacy traits resulted from experiences in marginal activities, such as receiving course grades, rather than in activities central to the process of learning physics. Marginal experiences would likely not be a major cause of the effect if I found that the self-efficacy states measured in physics were lower than those measured in other school activities. The second alternative explanation was that any gender differences in the shifts in self-efficacy traits were a result of gender differences in retrospection, how students remembered their experiences, rather than gender differences in self-efficacy state experiences in physics. For example, men and women could have had similar self-efficacy state experiences in physics, but women could have focused more on the low experiences they had and men could have focused more on the high experiences they had when reporting on their self-efficacy traits. I reasoned that gender differences in self-efficacy traits would have had little to do with gender differences in retrospection alone if women had lower self-efficacy states in physics than men.

The final alternative explanation is the possibility that any gender differences in self-efficacy states that students experienced in physics were mostly a result of preexisting gender differences rather than the physics-learning environment. I investigated this possibility by using a regression model to check whether the gender differences in self-efficacy states in physics were larger than those predicted by the traits that students started physics instruction with. I reasoned that if they were larger, then this would be an indication that physics instruction was impacting men and women differently due to inequities beyond the differences in the traits students started instruction with.

Evidence for the validity of the self-efficacy state construct, design goal three, was provided both internally within the ESM state data and externally between the self-efficacy state and trait data. In order to support the internal validity of the self-efficacy state construct I treated it as a latent variable, meaning that it was constructed by combining the responses to multiple questions about students' feelings of efficacy. I used a latent variable design to measure self-efficacy states and three complementary affective states using twenty Likert-scale affective questions. I reasoned that these multiple measures would provide two forms of evidence of the internal validity of the self-efficacy state measure. The first form of evidence I investigated by checking the relationships between the component questions of the self-efficacy state construct (skill, control and success), and the second by checking the relationships between self-efficacy states and the complementary states. I checked the external validity of the self-efficacy state construct throughout the research by checking how similar the differences for self-efficacy states were with the differences for self-efficacy traits. For example, lower self-efficacy states for women than for men in physics would be consistent with larger negative shifts in women's physics self-efficacy traits than men's. Therefore, if the trait measure produced this effect and the state measure did not, at least one of the measures would be in question. Since my state measure was novel, its validity would be brought into question more than that of the trait measure.

Investigating the extent to which any effect of physics instruction on students' self-efficacy traits that I measured generalized to other physics courses placed several demands on the research design since I collected data in only one physics course. I refer to this course as the focal IE physics course because it used interactive engagement (IE) teaching methods. I describe these teaching methods and this focal IE physics course in the context section in this chapter. I investigated an IE physics course as opposed to a traditional physics course because IE instruction is more effective at supporting student conceptual learning (Hake, 1998), and gender differences in conceptual knowledge tend to be smaller after IE instruction than after traditional physics instruction (Madsen et al., 2013). Thus, I reasoned that since self-efficacy and performance are related any effects I found in an IE physics course would be a conservative measure of effects in traditional lecture-based physics instruction. This reasoning is supported by larger negative shifts in students' self-efficacy traits in the traditional physics courses investigated by Sawtelle et al. (2010) than in the IE physics courses investigated by Kost-Smith (2011).

I investigated the extent to which my findings generalized to IE physics courses overall, design goal four, by checking the representativeness of the participants, the course and the experiences measured. I checked the representativeness of the participants by comparing them to the students who did not participate using the trait measures I collected. I investigated how representative the focal IE physics course was by comparing it to IE physics courses investigated by Kost et al. (2009a)

Table 2.1: The design structure of the dissertation. This table lays out the design goals and the questions that were investigated to address those goals along with the type of data that was used in the analysis. To save space in the table I shortened self-efficacy to SE.

Design Goal	Question	Measure
	Locate the effect within instruction	
Effect across instruction	How did self-efficacy traits shift from pre to post instruction?	SE traits
Effect within instruction	How did self-efficacy states in physics compare to those in other activities?	SE states
	Validity of the self-efficacy state measure	
Structure of the self-efficacy state construct	How strong and unique were the relationships between the core components of the self-efficacy state construct?	20 Likert-scale questions
Relationships between state constructs	How closely did the relationships between the self-efficacy states and the complementary states match the expected relationships?	States
Relationships between state constructs	How consistent were the differences for self-efficacy states with the differences for the complementary states in the comparisons?	States
Relationship between self-efficacy states and traits	Were the gender differences in the shifts in physics self-efficacy traits consistent with the gender differences in self-efficacy states experienced in physics?	SE states and SE traits
	Alternative explanations	
Epiphenomenon	How different were self-efficacy states in physics than those in other STEM courses?	SE states
Marginal Experiences	How different were self-efficacy states in physics compared to in other activities?	SE states
State-trait gap	Were there gender differences in the self-efficacy states students experienced in the focal IE physics course?	SE states
Inequities	What were the relationships between pre-course traits, gender and self-efficacy states experienced in the focal IE physics course?	SE states and traits
	Representativeness	
Participants	What differences existed for the trait measures comparing those that participated and those that did not participate in the research?	Traits
Course	How different was the focal IE physics course than other IE physics courses in terms of the means and gender differences in means on the trait measures?	Traits
Experiences	How different were the self-efficacy states experienced in the components of the focal IE physics course?	SE states

and Kost-Smith (2011) to investigate the extent to which they had similar effects on a range of outcomes. In both of these comparisons I used the trait measures for self-efficacy, attitudes, conceptual knowledge and course grades because they covered a broad range of outcomes and were all used by Kost et al. (2009a); Kost-Smith (2011). I investigated the representativeness of the experiences collected by comparing students' affective states in two different weeks of the semester.

The comparisons that I made to support these four design goals are laid out in Table 2.1. In the right hand column I indicated which type of data was used in these comparisons.

2.1.1 State data collection design

In addition to self-efficacy I collected data for three complementary affective states: activation, intrinsic motivation and stress, which are defined in Table 2.2. As described earlier the complementary affective states provided validity for the self-efficacy states by comparing the expected relationships between each of the complementary states and self-efficacy states with those that actually arose. The relationship between self-efficacy and stress was expected to be negative. When self-efficacy is higher, stress should be lower because self-efficacy is a measure of personal skill and stress arises when skill does not meet the demands of the situation (Lazarus and Folkman, 1984). The relationship between self-efficacy and both activation and intrinsic motivation was expected to be positive. People are more likely to become activated when they feel efficacious (Bandura, 1997), and people are also more likely to internalize motivation for activities in which they feel efficacious (Ryan and Deci, 2000). The relationship between stress and intrinsic motivation was expected to be negative following from their respective expected relationships with self-efficacy. The relationship between stress and activation is more complicated. Moderate levels of stress are associated with greater attention and learning while very low and very high

Table 2.2: Affective state construct definitions.

Construct	Definition
Self-Efficacy	Dynamically responsive judgments of one's ability to organize and execute the courses of action required to produce given attainments in the activity at hand.
Activation	An elevated level of excitement and involvement in the task, consistent with Thayer (1996) and in contrast to a relaxing state (Thayer, 1967).
Intrinsic Motivation	A drive to engage in the activity at hand, derived from within, either because it is personally enjoyable or valuable, as opposed to extrinsic motivation, which is driven by external pressures or rewards (Deci et al., 1999; Csikszentmihalyi, 1975).
Stress	Negative feelings resulting from an individual's perception that they do not have the resources to cope with a perceived situation (Lazarus and Folkman, 1984).

levels of acute stress are associated with decreased attention and learning (Lupien et al., 2007). Therefore a positive relationship was expected between stress and activation. However, it was possible that no relationship or a negative relationship could have resulted from an extremely stressful environment. The three relationships between self-efficacy, activation and stress are complicated in that they are not consistently positive or consistently negative.

A design priority was collecting sufficient samples of student experiences to reliably compare students' self-efficacy states in physics to those in other STEM courses for male and female students. Pilot studies showed that roughly 100 samples would be necessary for both male and female students in both the focal IE physics course and the other STEM courses in order to obtain a statistically reliable comparison. Collecting sufficient samples necessitated collecting ESM data over two separate weeks, including as many students as possible and over sampling the focal IE physics course. I over sampled the focal IE physics course to ensure adequate samples in physics since students spent less time there than in the broader categories of experience, for example their non-physics STEM courses.

Having two weeks of data collection also allowed me to investigate the representativeness of the experiences collected by investigating how similar the experience was in these two different weeks of instruction. To increase the range of course material that participants experienced, I selected weeks during the first and second half of the semester with six weeks between them. In order to keep the focus of the sampling on instruction, I selected the weeks so that no tests were taken or returned during these weeks, or during the week directly before or the week directly after the data collections.

2.2 Instrumentation, procedures and processing

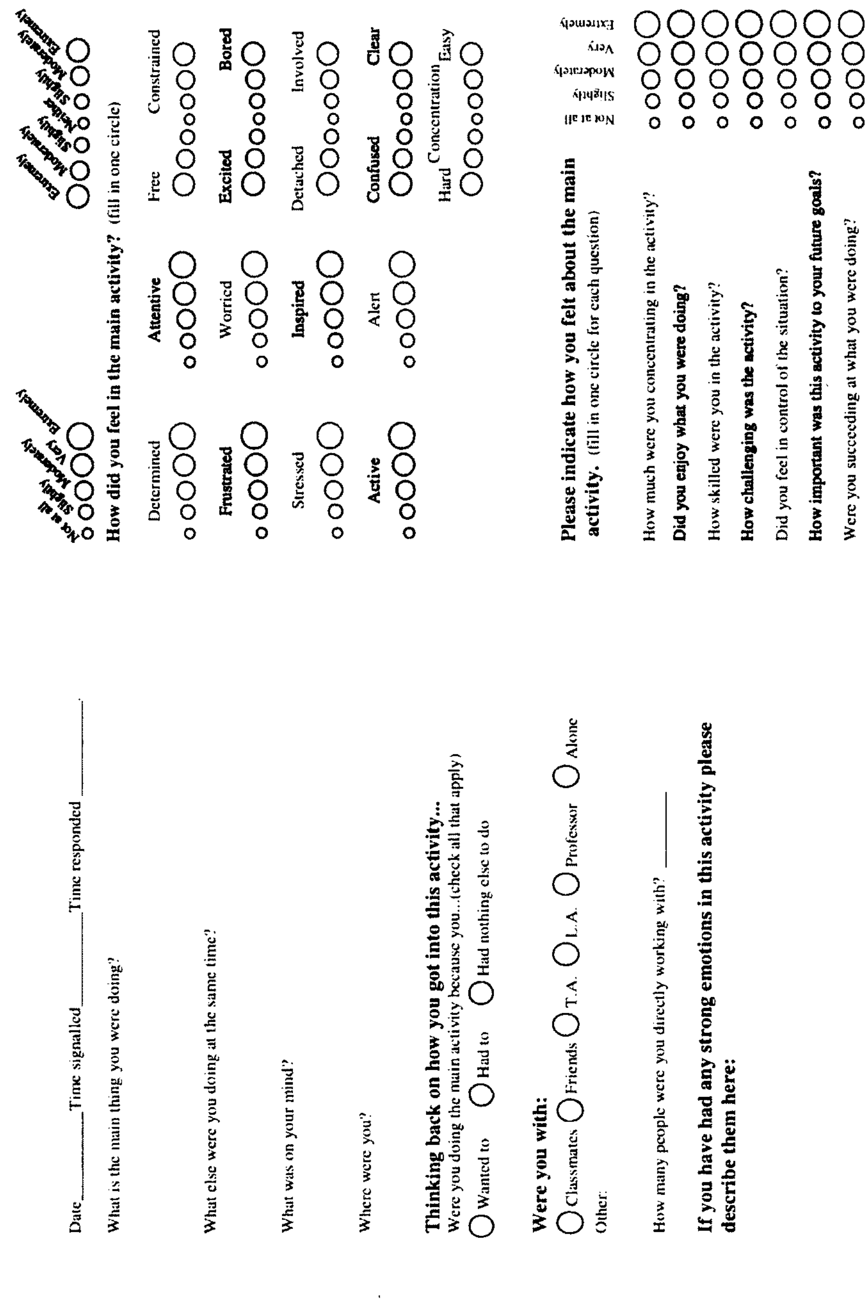
2.2.1 The experience sampling form

In general, the data collection instrument for ESM studies is a short survey that participants fill out when randomly signaled, or shortly after, about the activity they were engaged in at the moment of the signal. ESM studies typically refer to this instrument as the experience sampling form (ESF). I developed the ESF for this study modeled on those used in studies overviewed in Hektner et al. (2007). Each ESF was the single side of one standard-sized page split between the left and right halves, Figure 2.1. First the left section asked four free-response questions: (1) the main and (2) the secondary activities students were doing, (3) where they were and (4) what they were thinking about. I only used the first of these free response items, the main activity students were engaged in, for subsequent analysis because the research focused on how experience varied between different activities (e.g., different types of courses). Next, students answered a multiple-choice question assessing the autonomy they experienced in the activity, "Were you participating in the main activity because: you wanted to, you had to or you had nothing else to do? Circle all that apply." I used this question to provide validity for interpreting the intrinsic-motivation state measure as being scaled from intrinsic to extrinsic as opposed to being scaled from no intrinsic motivation to very high intrinsic motivation. Identifying this scale supported interpreting the relationship between intrinsic motivation and self-efficacy. The last two questions on the left half inquired about who they were with (friends, classmates, teacher, etc.) and provided an area to describe any strong emotions they experienced. These two sections were not analyzed for this dissertation. The sections that I did not analyze were included in the ESF to mirror the ESF's used in prior research, which will facilitate the comparison between the data collected in this research and data collected in other ESM studies. The right half of the ESF consisted of 20 Likert-scale questions on which participants indicated the type and level of affect they were experiencing at the moment they were signaled. These Likert-scale questions along with the main activity students were engaged in and their gender were the focus of my analysis of the state data.

2.2.2 State data collection procedures

Signals to fill out the ESF were sent to students' cell phones as text messages. I scheduled signals semi-randomly across each day such that there was a signal once during each two-hour block between 8 am and 10 pm and all signals were greater than 30 minutes apart. A constraint on the schedule was that a signal was scheduled for every physics course meeting (two lectures, two recitations and one laboratory), resulting in a higher rate of sampling for physics than for other experiences. I collected data from throughout students' daily lives so that physics and STEM courses would not appear to the participants to be treated differently than other activities.

I provided participants with a one-hour briefing on the data collection procedures. This briefing covered filling out the ESF, and participants completed one ESF for practice. I instructed participants to fill out the ESF as soon as possible


Date_______ Time signalled_______ Time responded_______

What is the main thing you were doing?

What else were you doing at the same time?

What was on your mind?

Where were you?

Thinking back on how you got into this activity...
Were you doing the main activity because you...(check all that apply)
Wanted to / Had to / Had nothing else to do

Were you with:
Classmates / Friends / T.A. / L.A. / Professor / Alone
Other:

How many people were you directly working with? ______

If you have had any strong emotions in this activity please describe them here:

Not at all / Slightly / Moderately / Very / Extremely
Extremely / Moderately / Slightly / Neither / Slightly / Moderately / Extremely

How did you feel in the main activity? (fill in one circle)

Determined	**Attentive**	Free	Constrained
Frustrated	Worried	**Excited**	**Bored**
Stressed	**Inspired**	Detached	Involved
Active	Alert	**Confused**	**Clear**
		Hard Concentration Easy	

Please indicate how you felt about the main activity. (fill in one circle for each question)

Not at all / Slightly / Moderately / Very / Extremely

How much were you concentrating in the activity?
Did you enjoy what you were doing?
How skilled were you in the activity?
How challenging was the activity?
Did you feel in control of the situation?
How important was this activity to your future goals?
Were you succeeding at what you were doing?

Figure 2.1: The experience sampling form used in this study.

after the signal and to fill it out no matter how much time had passed. Participants dropped off the completed booklets periodically in a marked and locked box in the physics building.

2.2.3 State data processing

I entered the ESM data for both the third and tenth weeks of the semester into a spreadsheet database. Only participants who completed at least fifteen surveys were included in the data analysis because that is considered to be a minimum threshold for establishing student's average experience when using the ESM (Hektner et al., 2007). I entered all surveys into the spreadsheet, but I only analyzed surveys that were completed within fifteen minutes of the signal. This time limit is standard practice in ESM studies to ensure that the samples are random and to be consistent with the in-the-moment nature of the ESM (Hektner et al., 2007). I coded the responses to the open-ended question "What were you doing?" using the two-digit coding scheme presented in Table 2.3. The first digit of the coding scheme was used to indicate the primary activity type and the second digit was used to distinguish between course components for school activities. School activities included being in a course or doing work for a course outside of the normal course meeting times. Because the coding was low inference, a single researcher, myself, completed all coding. I reduced the school activities to three approximately equal sized categories in order to maximize statistical power while providing a similar and a different activity to compare physics experiences to. The three school activities were: (1) non-STEM courses, which included a diverse range of courses such as English, anthropology, art, etc., (2) STEM courses, which excluded IE physics and primarily consisted of the introductory chemistry course for STEM majors, first semester calculus, and introductory engineering courses, and (3) the focal IE physics course. The process and results of this categorization are presented in Chapter 4.

Table 2.3: Coding scheme for the activities students engaged in. The school activities were coded with a secondary code for the course component.

Primary Code	Activity	Secondary Code	Course Component
00	Blank		
10	Non-School		
2*	STEM - Chemistry	*1	Lecture
3*	STEM - Calculus	*2	Laboratory
4*	STEM - Other	*3	Recitation
5*	Non-STEM Courses	*4	Homework
9*	IE Physics		

I used SPSS 19 to run statistical analyses on the data. In order to use this program I built a spreadsheet of the key variables of interest. Initially the spreadsheet was structured with each case being a single ESF and included the participant's identification number, gender and responses to each of the questions on the ESF. After I confirmed that the 20 Likert-items could be reduced to four affective constructs using principle components analysis, which is detailed in Chapter 3, I calculated a raw score and a Z-score for each state construct (i.e., self-efficacy and the three complementary affective states). I calculated the raw scores by averaging the component questions on a 5-point, 0-4, ratio scale. I discuss my reasons for using Z-scores in the effect size section of this chapter. To create Z-scores I first converted the twenty Likert-scale affect questions to Z-scores based on a given participant's mean and standard deviation for each question for a given week. Similar to the raw scores, I averaged the Z-scores for the component questions to get the Z-score for the affective state constructs.

2.2.4 Trait Instrumentation and data processing

I collected the bulk of the trait data using three different pre and post measures relying on standard survey instruments that were used by Kost et al. (2009a) and Kost-Smith (2011): physics self-efficacy traits (Kost-Smith, 2011), attitudes about learning physics (Adams et al., 2006), and conceptual knowledge (Thornton and Sokoloff, 1998). The three trait measures also measured sub-categories. However, I only used the overall scores for each of these instruments. Minimizing the number of variables used in the subsequent analysis made the results easier to understand and minimized the likelihood of spurious results.

I measured students' *self-efficacy traits* in physics by adapting the twenty-one Likert-scale self-efficacy questions from the Physics Self-Efficacy and Identity Survey developed by Kost-Smith (2011). To adapt this survey I removed the sources of self-efficacy questions and the identity questions because they were not the focus of my investigation, and I was concerned about participant attrition if the instrument was too long. Therefore I truncated the name to Physics Self-Efficacy Survey (PSES). The PSES measures self-efficacy across four constructs, but only the overall self-efficacy score, which ranged from zero to four, was used in this study. This self-efficacy instrument has only been used in one prior study (Kost-Smith, 2011) and was the only instrument available that had been used in an IE physics course.

I measured student *attitudes* about learning physics with the Colorado Learning Attitudes about Science Survey (CLASS) (Adams et al., 2006), which has been used extensively. The CLASS measures eight separate categories of student beliefs compiled from student responses to 42 questions. Responses are coded as favorable, neutral, or unfavorable based on agreement with expert responses. Like the PSES, the CLASS is multidimensional, having eight sub-constructs of expert-like response, but it also allows for an aggregate score. I used only the overall favorable score in the present study.

I measured student *conceptual knowledge* in the focal IE physics course with the Force and Motion Conceptual Evaluation (FMCE) (Thornton and Sokoloff, 1998), which is a 47 question multiple-choice exam commonly used for this purpose. The

FMCE was scored out of 37 points following the methods of Thornton et al. (2009) using a spreadsheet developed for that purpose (Wittmann, 2009).

A fourth comparative measure was *course grades.* I obtained course grades for the focal IE physics course from the instructor and analyzed them on a 4.0 scale, such that an A was 4.0 , an A- was 3.7, a B+ was 3.3, etc. This was the scale used at the institution and was the same scale used by Kost et al. (2009a).

2.2.5 Trait data collection procedures

Students completed the surveys for trait measurements during the first and last week of the course. The conceptual knowledge measurement (FMCE) was done during class. The FMCE was not graded, but it was a mandatory class activity for students in attendance. Students took the attitude and self-efficacy surveys (CLASS and PSES) outside of class via an online platform as a part of weekly homework assignments. Students received course credit equal to one homework problem for completing each survey. I obtained course grades from the instructor after the course had ended.

2.3 Approaches to sampling

Due to the intensive nature of the ESM it is common to collect state data from a representative sample of participants in a given context, such as a course or a school, rather than taking data from all students. Using this approach, from a physics course of 242 students (222 students completed the course and are represented by the largest circle in Figure 2.2) I conducted the ESM with 33 *ESM participants* who participated in the first, second or both data collections. This number of participants was dictated by both the number of students who volunteered and the maximum number of participants allowed by resources. By contrast, trait data was much easier to collect, and I obtained complete sets of all seven trait measures from a

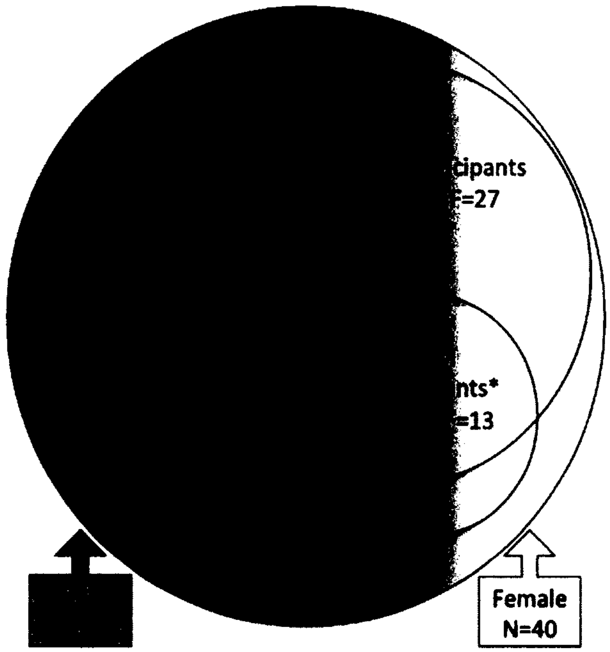


Figure 2.2: Diagram of the ESM participants and trait participants. * Eight male and five female ESM participants were not trait participants, two of the female ESM participants did not receive course grades. The shading distinguishes between male and female students in each population.

larger sample of 117 *trait participants.* Unfortunately, not all 33 ESM participants were part of the 117 trait participants (represented by overlapping populations in Figure 2.2).

I recruited ESM participants from their IE physics course through a brief announcement during class describing the research as investigating students' experiences in college. Students interested in participating submitted a slip of paper with their name and email. I contacted all students who submitted slips using a brief email message describing the research and directing them to an online form to sign up for a training session. All students over eighteen who desired to participate in the research were included, one student under eighteen years old was not included due to issues of informed consent. Participants were given a small amount of extra credit for the first week they participated in the ESM and a stipend of fifty USD for the second week they participated. No actions were taken to enlarge the proportion of women in the sample of ESM participants. However, I expected women to be over-represented given their over-representation in other studies in physics (Kost et al., 2009a; Kost-Smith, 2011). This turned out to be the case, and it was useful in providing sufficient samples to investigate gender differences in affective experience. As I will further discuss in Chapter 4 and Chapter 5, ESM participants over-represented high achieving students in the focal IE physics course. This is common in ESM studies (Hektner et al., 2007) and was also the case for the participants in the research by Kost et al. (2009a) and Kost-Smith (2011), which provided important points of comparison for the present study.

2.3.1 Participants

Considering the ESM participants and trait participants as a sample of the course populations raised some complex issues. Figure 2.2 illustrates this complexity. Out of 242 students who started the course, 222 completed the course and received grades (the largest circle in Figure 2.2). Of the students who received grades 40 (18%) were female. Of the 20 students who dropped or withdrew from the course, 5 (25%)were female. Of the 117 trait participants, 90 were male and 27 (23%) female (middle sized circle in Figure 2.2). The ESM participants initially included 44 students who signed up for and completed the one-hour training. Of these 44 students, 2 students did not follow the procedures correctly, and 9 students did not provide the minimum fifteen responses needed to be included in the analysis. This left 33 ESM participants (the smallest circle in Figure 2.2). Of these, 20 were male and 13 (39%) were female, as shown by the dark and light shading in the figure. 20 ESM participants were also trait participants, 12 male and 8 female. Among the 33 ESM participants 12 provided data during both data collections. Two of the female ESM participants withdrew from the course and did not receive final course grades.

2.4 Context

The study took place at the University of Maine, a four-year public university located in the northeastern part of the United States. The university was the leading research university for the state and was a PhD-granting institution in many STEM fields.

The focal IE physics course met five times in total each week: twice for 50 minutes of lecture with approximately 150 students, twice for 50 minutes of recitation with 24 students and once for 110 minutes of laboratory with 24 students. The instructor of the course was male and had thirty-five years of teaching experience. The data was collected during the fifth year the instructor taught this course. The course was modeled on IE physics courses described in Kost et al. (2009a). Almost all lectures used several conceptual multiple-choice questions embedded throughout the lecture (Mazur, 1997). Students answered these questions using a personal electronic response system and discussed these questions with their neighbors (i.e., they answered clicker questions). The course instructor called on volunteers to explain their reasoning for their answers. Students earned a small portion of their final course grade, 3%, by participating in the clicker questions. Three mid-term exams and one final exam were given in the lecture portion of the course. There was a weekly homework assignment with a written and an online component. Homework and tests included both conceptual and calculation problems. In the two recitation sections students spent most of their time solving conceptual problems in small groups. One recitation per week made use of a standard set of tutorial lessons (McDermott and Shaffer, 2002). The other recitation used a mix of locally-generated conceptual and calculation physics problems. Each recitation and laboratory class was facilitated by a graduate teaching assistant (TA). An undergraduate learning assistant (LA) assisted the TA during each recitation. The LA's had previously completed the course and were enrolled in a weekly seminar on pedagogy (Otero et al., 2010). The TA's and LA's met with the instructor of the course to go over the content and pedagogy used during recitation. These meetings emphasized the use of Socratic dialogue to support students in generating their own conceptual understanding in the activities in the recitation.

2.5 Methods of data analysis

2.5.1 Identifying meaningful and reliable differences: The three-step statistical process

When comparing means on a given measure (e.g., difference in mean self-efficacy state between physics and other STEM courses, or difference between men's and women's mean self-efficacy states) the first step I took was to directly compare the means of interest to identify if numerical differences existed. If mean differences existed I then used three standard steps to check for the statistical significance of the differences, which I refer to as the three-step statistical process. Not all of these steps were necessary for every comparison. In the first step I used an omnibus multivariate analysis of variance (MANOVA) to see if any statistically significant differences in means existed overall. This step was necessary whenever multiple dependent variables were being tested, such as all four affective constructs or all seven trait measures. If the MANOVA was statistically significant I used factorial univariate analysis of variance (ANOVA) to identify which dependent variables in the multivariate analysis were statistically significant for a main effect of each independent variable or an interaction effect between two or more independent variables. Lastly, I used post hoc tests to identify statistically significant differences between groups for each dependent variable. The last step I took when comparing means

Table 2.4: Assumptions for the statistical tests.

MANOVA	ANOVA	T-Test
Sufficient and equal cell size	Sufficient and equal group size	N/A
Multivariate normality	Normal	Normal
Absence of outliers	Absence of outliers	Absence of Outliers
Homogeneity of Variance-Covariance Matrices	Homogeneity of Variance	N/A
Linearity	N/A	N/A
Lack of multicollinearity and singularity	N/A	N/A
DV are continuous, IV are categorical	DV are continuous, IV is categorical	DV are continuous, IV is categorical
Independence of variables	Independence of variables	Independence of variables

was to use Cohen's *d*, histograms of the raw score responses, Z-scores of the affective constructs or some combination thereof to interpret the size of the differences.

2.5.1.1 Assumptions for the statistical tests

The practical issues and assumptions for each of the three steps in the statistical analyses are laid out in Table 2.4. These assumptions were tested for all of the statistical tests used in this dissertation and were all met with the exception of those that I discuss below.

In the case of the comparisons for the state data not all of the assumptions were met. The MANOVA's met six of the eight assumptions but did not have equal cell size or homogeneity of variance-covariance matrices. The MANOVA is robust against these two violations separately but not concurrently (Tabachnick and

Fidell, 2001). I used Box's M test to determine homogeneity of variance covariance matrices was not met. However, this test is understood to be quite sensitive and potentially overly sensitive (Tabachnick and Fidell, 2001). Therefore, I inspected the variance-covariance matrices directly and determined that the cells with the most samples, which in this case were non-school activities, had larger variances and covariances than the smaller cells. This indicated that the lack of homogeneity was conservative for my purposes because dishomogeneity would tend to produce false negative results (Tabachnick and Fidell, 2001). Based on this evidence I concluded that the MANOVA could be used for statistical testing despite not meeting the assumption of homogeneity of variance-covariance matrices.

Similar to the MANOVA the ANOVA did not have equal group sizes or homogeneity of variance. Like the MANOVA the ANOVA is robust to violating either of these assumptions independently but not concurrently (Tabachnick and Fidell, 2001). I used Levene's test to check the homogeneity of variances and found that the variance was not homogenous for all four dependent variables ($p < 0.05$). Subsequently I inspected the variances for each of the cells and found that the ratio of the variances between the cells ranged from 1.5 to 2.6. Thus, some of the data violated some of the rules of thumb for acceptable variance ratios; acceptable values vary from two (De Muth, 2014) to four (Durrheim and Tredoux, 2014). However, the variance was largest in the larger cells for self-efficacy, activation and motivation, but not for stress. This indicated that the lack of homogeneity was conservative for my purposes in these three tests because dishomogeneity would tend to produce false negative results. Since stress was a secondary focus and I could use consistency between the differences for all of the affective states to inform the results of the ANOVA's, the possibility of spurious differences in stress did not preclude me from using the ANOVA. Based on this evidence I concluded that the ANOVA could be used for statistical testing despite not meeting the assumption of homogeneity of variance-covariance matrices or having equal group sizes.

Step three in the statistical analysis used Tamhane's T2 test when comparing between activities or two-sided T-tests when comparisons were limited to two groups; examples include comparing men's and women's self-efficacy states in physics or comparing women's self-efficacy states in physics to their self-efficacy states in other STEM courses. Tamhane's T2 test was used because it is robust against violations of homogeneity of variance. In the case of two-sided T-tests when the variance between the two groups was not equal, as indicated by Levene's test for inequality of variances, I used the results for the test not assuming equal variances.

2.5.2 Effect sizes

I used Cohen's d, histograms of the raw score responses and Z-scores of the affective constructs to interpret the size of the differences in affective states. I used Cohen's d because it is a common measure in comparing the distance between two groups, and it facilitates comparisons between studies. Cohen's d measures the size of the difference between two groups in a normalized unitless measure scaled in terms of the pooled standard deviation of the two groups. Z-scores are a frequently used method in ESM studies (Hektner et al., 2007) to address differences in how individual participants used the scales. I used Z-scores of the affective states to situate students' mean experiences in activities within their overall experience. I used histograms to identify differences in the actual distributions of students' affective states.

To produce histograms I reduced the raw scores into four bins for each of the affective constructs. This reduction was non-standard in that both the very low and very high bins included both respective endpoints. Subsequently the bins were: very low [0,1], moderately low (1,2], moderately high (2,3) and very high [3,4]. I used this non-standard binning to have the very low and very high bins span equal ranges since the most important differences would occur in those bins.

2.5.3 Confirming the structure of the affective state constructs

I used Principle Components Analysis (PCA) to provide validity for the self-efficacy state construct by checking that the groupings of the Likert-scale questions for each of the affective state constructs matched the expected groupings for that construct. Constructs with groupings inconsistent with those that I expected could invalidate the affective state constructs. I used PCA because it is a common method of data reduction for multi-dimensional instruments and is often used in ESM studies (Hektner et al., 2007). PCA is a method for identifying relationships within a set of individually correlated variables in order to identify coherent subsets of those variables that are relatively independent (Tabachnick and Fidell, 2001). In PCA the factors are calculated from the factor loadings that are output by the analysis as this maximizes the variance explained by the extracted factors. In this case I elected to average the variables within each factor rather than to calculate the factors based on the factor loadings from the PCA. Averaging the variables simplified the data analysis and facilitates the use of this ESF in future studies. Therefore, I referred to the extracted factors as affective constructs.

The most important and difficult choice in PCA is determining how many factors to extract. Several methods exist to select the number of components to extract and I opted for extracting factors with an eigenvalue greater than one as that indicated that the factor explained more variance than the average single item (Girden, 2014). PCA using eigenvalues greater than one extracted four factors. I also performed PCA with a forced extraction of five factors, the number expected based on the original design of the ESF, and determined that the four factor model found by including factors with eigenvalues greater than one more closely aligned with my expected model. As is common practice I used Varimax rotation to concentrate the factor loadings and to clearly identify which items loaded on each construct. Rotated matrix factor loadings greater than 0.25 were noted and loadings greater than 0.40 were used to assign questions to individual factors. This accords with accepted PCA procedures in which factor loadings of 0.32 are considered a minimum for assigning items to factors (Tabachnick and Fidell, 2001).

The goal of the analysis was to investigate if the structure within the Likert-scale data was consistent with the expected theoretical groupings of the data. Therefore, I sought to show that the data set from the ESM met the assumptions of the PCA in order to support the strength of the interpretation. As is common practice I used the Kaiser-Meyer-Olkin statistic to determine the overall suitability of the data for PCA; results of my analysis exceeded the standard of 0.60 (Tabachnick and Fidell, 2001). I investigated the multivariate normality of the data, multivariate outliers and the linear relationships between the affective Likert-scale questions. Multivariate normality is the extension of a normal distribution into n-dimensional space. Testing of the multivariate normality was supported by testing the normality of the individual Likert-scale items on which the PCA was run and by testing the normality of the subsequent principle components that were generated based on the results of the PCA (Jackson, 2003). Normality of the individual questions was established using the rule of thumb that the skew and kurtosis both fell within the range -2 to 2. I used correlations between the items included in the PCA to verify linear relationships existed within the data. I used Mahalnobis distances to test for multivariate outliers. I did not remove multivariate outliers from the data set. Instead I confirmed that the results of the PCA were the same whether they were included or removed.

Cronbach's Alpha is a measure of the reliability of a combination of variables. I have included it as a measure of the internal validity of the affective constructs because it is a simple measure that is commonly used and is easily interpreted. A useful rule of thumb is that scores greater than 0.70 are evidence of satisfactory internal consistency.

2.5.4 Measuring relationships between affective state constructs

There was a need to check the strength of the relationships between the different state and trait measures. I used bivariate correlations, partial correlations or both to measure the strength of these relationships. Bivariate correlation is a method of assessing the degree of relationship between two continuous variables (Tabachnick and Fidell, 2001). Partial correlation is a method of determining the degree of relationship between two variables while taking into account the role of other variables (Tabachnick and Fidell, 2001). I used partial correlations to investigate the relationships between self-efficacy, activation and stress because I expected the positive relationship between stress and activation to conceal some of the strength of the positive relationship between self-efficacy and activation since self-efficacy and stress were negatively related.

2.5.5 Investigating the relationships between states and traits

In order to investigate the relationships between state and trait data I needed to check these relationships while controlling for other variables. Multiple linear regression (MLR) allows investigating the relationships between one dependent variable and several independent variables. I used MLR to inform the strength of the relationships between students' mean self-efficacy states in physics and their traits. MLR informed how the strengths of these relationships changed as other independent variables (traits) are added to the analysis. In using MLR I built multiple models of the dependent variable, which was always students' mean self-efficacy states experienced in the focal IE physics course. Each subsequent model included additional dependent variables with the first model only including gender. The relationships between the dependent variable and each independent variable are represented by the β_k in Equation 2.1. Thus, in each model β_0 represented the intercept value and β_1 represented the relationship between gender and mean self-efficacy states. Gender was a dummy variable with one for females and zero for males so that β_1 indicated how different females' mean self-efficacy state experiences were than males' mean self-efficacy states. The change in β_1 between models provided a measure of how much of the gender difference in self-efficacy states was explained by the trait variables. I also reported on the total variance explained by each model. Total variance explained is a measure of the information that the model provides and comparing total variance explained across the models informs the additional information that each independent variable adds to the model.

$$Mean\ SE\ state = \beta_0 + \beta_1 \times Gender + \sum_{k=2}^{N} \beta_k \times VAR_k \qquad (2.1)$$

A practical issue that arose in using MLR in this investigation was the structure of the trait and state data. Each participant in the study had, at most, one measure for each of the trait instruments at the beginning and again at the end of the course, but the number of experiences in physics they reported varied. Therefore in order to structure the data as one case for each individual I averaged each participant's self-efficacy states in the focal IE physics course and these means were used in the MLR.

In using MLR I built several models to investigate the relationships of interest and determine how these relationships changed with the addition of other independent variables. The first step in building these models was to identify variables that correlated with the dependent variable but did not have excessively high correlations with other independent variables. My first model included only the independent variable of greatest interest, gender, to determine the variance explained by gender alone and to determine the strength of the relationship between gender and mean self-efficacy states as indicated by the standardized and unstandardized betas. This model acted as a baseline against which the subsequent models were compared. Subsequently, I added the other independent variables of interest to the model, noting how the variance explained of the overall model shifted and how the betas for each independent variable shifted. I reported and interpreted both the corrected and uncorrected variance explained. The corrected variance explained takes into account the number of variables included in the analysis and provides an overall measure of how much information the model is providing. This correction comes at the cost of obscuring how much unique information the new variable is contributing in comparison to the variables in the previous models, which is more clearly illustrated by the uncorrected variance.

Several practical issues and assumptions for MLR impact interpreting the results of the analysis. These guidelines include:

- sufficient sample size,
- absence of outliers,
- absence of multicollinearity,
- normality, linearity and homoscedasticity of residuals,
- independence of errors and
- outliers in the solution.

These guidelines were tested using the procedures recommended by Tabachnick and Fidell (2001) and were met with the exception of sufficient sample size. Sufficient

Table 2.5: Statistical methods used in the dissertation.

Chapter	Question	Method
3	How strong and unique were the relationships between the core components of the self-efficacy state construct?	PCA
3	How closely did the relationships between the self-efficacy states and the complementary states match the expected relationships?	Cor
4	How different were self-efficacy states in physics compared to in other STEM courses?	3-Step
4	How representative were the ESM participants?	3-Step
4	How different were the self-efficacy states experienced in the components of the focal IE physics course?	3-Step
5	How big were the gender differences for self-states in physics as compared to those in other STEM courses?	3-Step
5	How representative were the ESM participants when taking gender into account?	3-Step
5	How representative were the trait participants?	3-Step
5	How large were the gender differences on the trait measures?	3-Step
5	How similar were the means and gender differences in means for the trait measures in the focal IE physics course to other IE physics courses?	3-Step
6	What were the relationships between pre-course traits, gender and self-efficacy states experienced in the focal IE physics course?	MLR

sample size is not an assumption or requirement of MLR, but it is a guideline for minimizing the chance that the results are unreliable. Meeting this guideline, however, is not strictly necessary to employ MLR for investigating the relationships between variables (Berk, 2003). Even when sufficient samples are not available MLR is still useful for informing the relationships between the variables. Therefore, I have included the statistical significance of all betas, but I am more concerned with the size of the relationships, the shifts in the models and the overall statistical significance of the models.

2.6 Questions and analyses used throughout the dissertation

In Table 2.5 I lay out the structure of the questions that were investigated in each of the results chapters in this dissertation and the methods that were used to investigate them. I used some of these methods infrequently such as principle components analysis, Cronbach's alpha, multiple linear regression analysis and correlations. In contrast, I used the three-step statistical process, or components of it, to investigate several questions in Chapter 4 and Chapter 5.

Chapter 3
INTERNAL VALIDITY OF THE SELF-EFFICACY STATE MEASURE

3.1 Introduction

My goal in this chapter is to establish a foundation of the validity of the self-efficacy state construct by investigating two forms of internal validity. These measures of internal validity were available due to the multiple measurements the ESM captured during each experience. First, I used principle components analysis (PCA) to show that the component questions of the self-efficacy state measure grouped as expected and with sufficient strength and independence. Second, I used correlations to investigate if the relationships between self-efficacy states and the complementary states matched those expected in the theoretical framework. The following chapters of the dissertation also provide additional markers of internal and external validity of the self-efficacy state construct. I will discuss these markers in the conclusion of this chapter.

Secondary to this primary goal I also present analysis of the motivation state construct to demonstrate that it varied from extremely extrinsic to extremely intrinsic, as it was expected to, and not from none to extreme intrinsic motivation. Determining how motivation varied informed the relationship between the self-efficacy state and the intrinsic motivation state measures and supported characterizing student experiences. For instance, it is different to find that low self-efficacy and low motivation tended to occur concurrently as opposed to low self-efficacy and very extrinsic motivation tended to occur concurrently. The first case implies that in very low self-efficacy states students are not motivated. The second case, in contrast, implies that in very low self-efficacy states students draw on external sources of motivation.

This second case implies that students have the capacity to overcome difficulties they encounter in the learning environment, while the first case portrays students as unable to find motivation in challenging situations.

3.1.1 There is a need for a dynamic measure of self-efficacy

"Efficacy beliefs should be measured in terms of particularized judgments of capability that may vary across realms of activity, under different levels of task demands within a given domain and under different situational circumstances." (Bandura, 1997, p. 42). In this statement Bandura describes self-efficacy as a highly dynamic construct that can vary across domains, challenges and circumstances. These characteristics demand a highly dynamic method of measuring self-efficacy that can assess how self-efficacy is affected by each of these variables. Bandura further proposed that self-efficacy is a causal force and key factor in human agency (Bandura, 1997, p. 3) and findings support this causal relationship in terms of cognitive outcomes (Bouffard-Bouchard, 1990) and interest and choice in science majors (Luzzo et al., 1999). Self-efficacy, however, is not the only factor in human agency and acts in concert with many other latent variables.

An example of work attempting to link the dynamics of self-efficacy to other affective constructs is that of Bledow (2013). Bledow (2013) proposed that self-efficacy is the "tip of the iceberg" of a broader set of self-regulatory processes that includes conscious and subconscious processes he calls self-motivation. Bledow (2013, p. 15) claims, "...a positive rate of change in task specific self-efficacy indicates self-motivation." and is "...inseparably linked to an opponent process, the oscillating perception of demands." Investigating the role that changes in individual's self-efficacy beliefs have on their actions and feelings requires a measure of self-efficacy state dynamics.

To test Bledow's claims the dynamic measure must be able to capture self-efficacy in every day activities across a much longer time scale than those possible in laboratory studies. The self-efficacy state measure introduced in this dissertation has these capabilities. It can measure the dynamics of self-efficacy throughout daily life and across time in a specific domain. Of equal importance the ESM can concurrently capture the experience of other affective states such as motivation. These characteristics of the ESM and the self-efficacy state construct raise the possibility of future research investigating Bledow's ideas.

3.1.2 Designing the self-efficacy state measure

The survey utilized in ESM studies is called the Experience Sampling Form (ESF). Hektner, Schmidt and Csikszentmihaly (2007) provide an overview of constructs and their component questions used in prior ESM studies in both education and non-education settings. This overview acted as the starting point for the design of the ESF. In addition to self-efficacy, four complementary constructs were designed to be on the ESF. I discussed four of these five constructs, self-efficacy, activation, intrinsic motivation and stress, in Chapter 2. The fifth construct was cognitive efficiency, the depth and difficulty of concentration. I did not discuss it earlier because, as I will discuss, results indicated that cognitive efficiency did not form a unique construct. The expected components of these constructs are laid out in Table 3.1.

3.2 Research questions

The central question I sought to answer in this chapter was the extent to which the ESM could be used to measure self-efficacy states. In order to answer this question I asked two research questions to investigate if the self-efficacy state measure

Table 3.1: Expected and actual components of the affective constructs. DifCon is a measure of difficulty concentrating from hard to easy. Clarity is a measure from clear to confused

Construct	Expected Components	Actual Components
Self-Efficacy	Skill, Control, Success	Skill, Control, Success, *DifCon* , *Clarity*
Activation	Determined, Active, Attentive, Alert, Inspired, *Challenge*	Determined, Active, Attentive, Alert, Inspired, *Detached/Involved*, *Concentrating*
Intrinsic Motivation	Free/Constrained, Enjoy, Excited/Bored, Importance *Detached/Involved*	Free/Constrained, Enjoy, Excited/Bored, Importance
Stress	Stress, Worry, Frustration	Stress, Worry, Frustration
Cognitive Efficiency	*Concentrating*, *DifCon*, *Clarity*	

was consistent with the literature on self-efficacy both by the component questions that formed it and by its relationship to the other affective states.

1. To what extent did the three primary component questions (skill, control and success) form a strong and independent self-efficacy state construct?

2. To what extent did the relationships between this self-efficacy state and the three complementary affective states match those described in the theoretical framework?

Together these two questions can show that the self-efficacy state measure was consistent with the current understanding of self-efficacy in terms of both what it was composed of and how it related to the other affective constructs. This provides a robust first step in demonstrating the validity of the self-efficacy state measure.

3.3 Design of the analysis

I designed the analysis presented in this chapter to reduce the data to the self-efficacy state construct and the three complementary affective state constructs and to provide internal validity for these constructs. I reduced the data to these four constructs using principle components analysis (PCA) as described in Chapter 2. Results of the PCA tested the internal validity of the self-efficacy state construct.

Once the four affective constructs were confirmed and calculated I investigated the correlations between the four affective constructs to further investigate the validity of the self-efficacy state construct as described in Chapter 2.

I used an additional question on the ESF regarding student autonomy to determine if intrinsic motivation varied from intrinsic to extrinsic. The autonomy question identified if experiences were intrinsically motivated in that students were doing them because they "wanted to" or if they were extrinsically motivated in that students were doing them because they "had to". Thus, I expected that extrinsic motivation would occur more frequently in the situations students felt they "had to" do and intrinsic motivation would occur more frequently when students "wanted to" do the activity.

3.3.1 Procedures

The procedures for investigating the internal validity of the self-efficacy state measure are described in Chapter 2. In this section I present the procedures and results of the investigation to confirm that the motivation state varied from extrinsic to intrinsic. I include these results here as this analysis was a secondary analysis to support the analyses of the self-efficacy state constructs internal validity.

I investigated the differences in motivation states between autonomy conditions by comparing the means and histogram distributions for the four autonomy con-

Table 3.2: Motivation versus autonomy.

Autonomy Condition	Motivation: Very Extrinsic	Moderate Extrinsic	Moderate Intrinsic	Very Intrinsic	Total Count	Mean	S.D.
Neither	2%	22%	46%	30%	87	2.56	0.76
Wanted To	1%	10%	38%	51%	643	2.91	0.72
Had To	24%	56%	17%	3%	577	1.49	0.67
Both	4%	32%	53%	11%	133	2.14	0.65
Total	11%	31%	32%	27%	1440	2.25	0.96

ditions: wanted to, had to, neither or both. I tested the statistical significance of differences in the means using the ANOVA and Tamhane's T2 test as described in Chapter 2.

The numerical value for motivation was greatest in the "wanted to" condition and lowest in the "had to" condition with the difference being very large in terms of both the total scale, 0 to 4, and the standard deviations, Table 3.2. The distributions of motivation states show that the majority of "had to" experiences were extrinsic while the majority of "wanted to" experiences were intrinsic. The "neither" condition and "both" condition were primarily experienced with intrinsic motivation, but they were less skewed than the "wanted to" condition.

I investigated the reliability of the differences in motivation states between the four autonomy conditions using an ANOVA. Results of the ANOVA indicated that there was a statistically significant difference $F(3,1436)=431.5$, $p<0.001$. Post hoc analysis using Tamhane's T2 revealed that all two-way comparisons were statistically significant ($p<0.001$).

The large and reliable difference in motivation states between the "had to" and the "wanted to" conditions was consistent with interpreting the motivation state as varying from extrinsic to intrinsic. The distributions of the conditions indicated that the midpoint of the scale was an acceptable point for differentiating between intrinsic and extrinsic motivation. The vast majority of experiences having either been "had to" or "wanted to" also supported using a bi-polar scale rather than two distinct unipolar scales as this indicated that intrinsic and extrinsic motivation were seldom experienced concurrently.

3.4 Results

3.4.1 Investigating if skill, control and success formed a unique and strong self-efficacy state construct

Principle components analysis (PCA) with Varimax rotation was used to clearly identify which variables loaded on each construct. PCA found four factors with an eigenvalue greater than 1. I compared these results to a forced extraction of five factors, the number expected based on the design of the ESF, and determined that the four factor model more closely aligned with the theorized model. Rotated matrix factor loadings greater than 0.25 were noted and loadings greater than 0.40 were used to assign questions to individual factors.

In the first step all data for the 20 Likert-scale questions were included in the analysis. This revealed that the question "How challenging was the activity?" loaded on all four factors greater than 0.250 and I subsequently removed it from the analysis because it could not be readily identified with any single factor. I reran the PCA on the remaining 19 questions with no noteworthy differences and results for those 19 questions are shown in Table 3.3. All 19 questions had acceptable levels of variance explained by the model and were included in calculating the four affective constructs.

These four constructs explained 58.5% of the variance in the data as shown in Table 3.3. Based on the design of the research I named these factors activation, self-efficacy, stress and intrinsic motivation. Throughout the dissertation I refer to these factors as constructs because I used the average of the component questions to calculate the value for each construct as opposed to using the factor loadings. I discussed this decision in the PCA section in Chapter 2. Analysis did not identify the expected factor for cognitive efficiency; its expected components (Table 3.1) were split between activation and self-efficacy as shown in Table 3.3. I judged it as reasonable for students' experiences of self-efficacy to have been tied to their cognitive efficiency as they were largely engaged in and judged on their thinking.

Factor loadings for skill, control and success were all very good or excellent on the self-efficacy state construct and none of these component questions had factor loadings greater than 0.250 on any other factor. Factor loadings for other questions were straightforward to judge except for three questions: clarity, bored and enjoy, all of which loaded greater than 0.4 on two constructs. I expected some cross loading to occur as I designed the survey to include factors that were interrelated. I included clarity in the self-efficacy construct because I designed it to go with concentration difficulty, though it makes sense that clarity, which is a measure of confusion, would be related to the level of stress students experienced. I included bored in the intrinsic motivation construct because it loaded more heavily on that factor than on the activation factor. I included enjoy in the intrinsic motivation construct because it loaded more heavily on that construct and because it is more directly applicable to my definition of intrinsic motivation than to self-efficacy. One other individual question, detached/involved, loaded differently than expected based on the design of the ESF and results from the pilot study data. I expected detached/involved to align with the intrinsic motivation construct. Instead it aligned with the activation construct. I am unsure of why this occurred, but it is reasonable for this question to align with either of these categories. I also ran PCA on subsets of the data to compare the constructs for male and female participants and to compare constructs in school versus non-school activities. The PCA for these subsets aligned with the analysis presented here.

Table 3.3: Principle components analysis results.

Likert-Scale Question	Normalized Variance Explained	Factor Loadings			
		Activation	Self-Efficacy	Stress	Motivation
Determined	0.55	0.67			0.27
Active	0.46	0.59			
Attentive	0.64	0.80			
Inspired	0.36	0.56			
Alert	0.60	0.77			
Involved	0.57	0.73			
Concentration	0.63	0.75			
Clarity	0.50		0.52	-0.41	
ConDif	0.45		0.51		-0.30
Skill	0.64		0.79		
Control	0.50		0.68		
Success	0.70		0.82		
Frustrated	0.65		0.31	0.71	
Stressed	0.76			0.83	
Worried	0.66			0.80	
Constrained	0.57		-0.28	0.27	0.65
Bored	0.65	-0.44			0.64
Enjoy	0.65	0.26	0.49		0.57
Importance	0.58	0.26			0.71
	Total	Variance Explained (%)			
	58.5	25.6	20.2	6.6	6.1
		Cronbach's Alpha			
		0.87	0.76	0.79	0.70

Cronbach's alpha was satisfactory for each construct and indicated that each construct was made up of consistent questions, Table 3.3.

3.4.2 Relationships between the self-efficacy state construct and the complementary affective states

I used correlations to measure the relationships between the four affective constructs. These correlations, presented in Table 3.4, confirmed the relationships that were expected and were all statistically significant with $p<0.01$. The strongest

Table 3.4: Correlations and partial correlations between the four affective state constructs. Right of the slash are partial correlations. The control for SE and Mot was Stress, the control for stress was SE. All correlations and partial correlations were statistically significant ($p<0.01$).

	Self-Efficacy	Stress	Motivation
Activation	0.16/0.28	0.15/0.28	0.14/0.23
Self-Efficacy		-0.52	0.54
Stress			-0.41

correlations were the positive relationship between self-efficacy and intrinsic motivation and the negative relationship between self-efficacy and stress. Activation was very weakly related to the other constructs. I expected that this was a result of the complex relationships between stress, self-efficacy and activation. Therefore, I used partial correlation analysis to determine the relationship between each of the other affective constructs and activation while controlling for the effects of the other constructs. I analyzed the partial correlation between stress and activation while controlling for self-efficacy. This partial correlation was much higher than the bivariate correlate, 0.28 versus 0.15, and is presented in Table 3.4. All partial correlates were statistically significant at the $p<0.01$ level. I analyzed the partial correlations between activation and self-efficacy and between activation and intrinsic motivation while controlling for stress and both of the partial correlations were much higher than the bivariate correlations were, Table 3.4. This indicates that there were two distinct times in which activation was heightened during instruction and life in general. Activation was higher when students' sense of efficacy or intrinsic motivation or both were heightened or in the opposing scenario when students experienced heightened levels of stress.

3.5 Discussion

Skill, control and success, which are central attributes of self-efficacy, all loaded on the self-efficacy state construct with very good or excellent factor loadings. The other questions that loaded on the self-efficacy construct all did so in a manner that was consistent in that higher self-efficacy was associated with greater enjoyment, less frustration, easier concentration and greater clarity. The consistency and strength of the self-efficacy state construct provides strong evidence for the ability of the ESM to measure self-efficacy states directly in the activity at hand.

Similarly, the relationships between self-efficacy and the complementary affective states matched those expected. Stress was negatively related to self-efficacy, which was consistent with their opposing relationships to ability. Stress arises when ability is lacking and self-efficacy arises when ability is perceived to be sufficient for the task at hand. Intrinsic motivation was positively related to self-efficacy, which follows from activities that were more efficacious being more intrinsically motivated and less efficacious activities being more extrinsically motivated. Once the opposing effect of stress was taken into account the relationship between self-efficacy and activation was positive, but small, and indicated a greater level of activation in those activities that participants felt skillful in. These relationships all matched those expected and indicate that the self-efficacy state construct aligned with self-efficacy in its relationship to other affective states.

That these measures of internal validity were consistent is a strong indication that the self-efficacy construct is a reliable and distinct measure of the efficacy that students experienced in activities throughout their daily lives. I did not expect two of the three cognitive efficiency questions to load on to the self-efficacy construct. However, both difficulty in thinking and clarity of thought logically align with the three primary questions. Both can be interpreted as extensions of skill and success, particularly for students who are frequently engaged in activities explicitly designed to stimulate thinking and who are judged on their ability to think. Based on this logic and the results of the PCA I decided to include these two cognitive efficiency questions in the self-efficacy state construct. It is possible, however, that other populations may not have similarly strong relationships between their cognition and their self-efficacy.

3.6 Conclusion

The strong and unique loading of skill, control and success on the self-efficacy state construct along with the relationships between the self-efficacy state construct and the complementary state constructs matching the expected relationships provide a strong foundation of validity that I will continue to build upon in the following chapters. Additional evidence will focus on further indicators of internal validity between self-efficacy and the complementary states and external validity between self-efficacy states and traits. The relationships between the self-efficacy state measure and the complementary state measures should also show up in how different activities were experienced. Activities experienced with greater self-efficacy should have also been experienced with greater activation, more intrinsic motivation and less stress. This argument also follows for populations of students. If male students experience an activity with greater self-efficacy than female students they should also experience greater activation, more intrinsic motivation and less stress. I used a similar argument to check the relationship between self-efficacy states and traits. If women had more negative shifts in their self-efficacy traits from pre to post instruction in the focal IE physics course then they should also have experienced much lower self-efficacy states during instruction.

Further support for the validity and generalizability of the self-efficacy state construct can be achieved by confirming the indicators of internal validity that I presented in this chapter for other populations such as younger students and especially non-students. Here I found that cognition was a part of students' self-efficacy states, but this may not generalize to other populations given that the participants in the study were high achieving students in college STEM courses. This broad base of validity could be achieved without any additional data collection, as some extensive ESM data sets are available for researchers to analyze. Further, analyzing these data collections from the perspective of self-efficacy could elaborate on the research here. In particular the Sloan Survey of Youth and Social Development, described in Chapter 1, could be used to investigate if there was a gender difference in self-efficacy states experienced during physics learning for secondary students.

Chapter 4
SELF-EFFICACY STATES IN PHYSICS

Increasing the number and diversity of students who enter and remain in the STEM education pipeline is an important goal in the United States and in many developed countries (National Commission on Mathematics and Science Teaching, 2000; Osborne, J. and Dillon, J., 2008). Attainment of this goal depends on positive learning outcomes for diverse students in K-16 STEM education. One important and often overlooked class of outcome is student's development of coherent affective traits, such as their attitudes about learning science or their self-efficacy traits. The development of these affective traits is a core goal of science education (Redish et al., 1998; Brewe et al., 2009) because it supports students in pursuing their STEM education and their future STEM careers. Survey-based studies in STEM have measured the mediating effect on student achievement for interest (Koller et al., 2001), motivation (Singh et al., 2002; Mujtaba and Reiss, 2013), more general attitudes and beliefs (Perkins et al., 2004; Madsen et al., 2015) and self-efficacy traits (Lee, 2009; Marra and Bogue, 2009; Chemers et al., 2001).

The relationships between affect and student achievement are also well established for self-efficacy in physics learning, which is the focus of this study. Kost-Smith (2011) found that self-efficacy traits correlated with students' test grades in an introductory physics course. Sawtelle et al. (2012b) found that self-efficacy traits predicted students passing introductory physics courses.

In addition to studies showing that self-efficacy traits predict student attainments in educational settings, experimental design studies have shown causal relationships between self-efficacy and both immediate cognitive performance and long term career interest. Using an experimental design providing students with bogus feedback to either increase, decrease or neutrally impact students' self-efficacy, Bouffard-Bouchard (1990) found that increasing students' self-efficacy caused students to set higher standards for themselves, use more efficient strategies and achieve greater intellectual performance. Using an experimental design to increase students' self-efficacy traits, Luzzo et al. (1999) found that increasing students' self-efficacy traits increased students' interest in pursuing science and mathematics careers.

Despite the importance of self-efficacy for student performance and persistence in STEM education, in my review of the literature I found no studies reporting positive shifts in students' physics self-efficacy. Instead, I found that most of the studies in physics described negative shifts in physics self-efficacy traits, but no studies described negative shifts in self-efficacy traits for other STEM courses. Self-efficacy traits consistently decreased from pre to post instruction in both research-based physics courses and traditional physics courses (Kost-Smith, 2011; Sawtelle et al., 2010; Lindstrom and Sharma, 2011). The best outcomes I found were no changes in self-efficacy traits in a research-based course for majors (Sawtelle et al., 2010) and an inquiry-based course for non-majors (Cavallo et al., 2004). By comparison, in introductory chemistry courses (Dalgety and Coll, 2006; Villafane et al., 2014; Ferrell and Barbera, 2015), an introductory algebra course (Brewer, 2009) and introductory biology courses for non-majors (Lawson et al., 2007; Roster, 2006) students' self-efficacy traits increased from pre to post instruction. This comparison is both limited and strengthened by the diversity of these studies. It is limited because in the studies of physics and in the studies of other STEM courses both the type of instruction and the instruments to measure self-efficacy were highly varied. Some of the studies were for courses with traditional instruction while others studied research-based courses; some of the studies were in courses for majors and others for non-majors. In some of the studies the self-efficacy measure was a part of the instrument; in other studies the entire instrument solely measured self-efficacy. These differences make it difficult to reliably compare the effects of the courses subject matter on self-efficacy. The differences between these studies leaves the possibility open that there is some broad negative impact on students' self-efficacy that is common in introductory college STEM courses for STEM majors. However, the diversity of these studies strengthens the comparison of the effect of physics instruction on self-efficacy to the effect of other STEM courses in that there was a consistently negative or neutral shift in self-efficacy traits in physics and a consistently positive shift in other STEM courses for a wide range of course types.

The more negative shifts in self-efficacy in physics motivates the central focus of this investigation, which is to investigate the extent to which the focal IE physics course caused negative impacts on students' self-efficacy traits. Tying the shift in self-efficacy traits in physics to the experience of learning physics follows from the central assumption I make in this dissertation that self-efficacy traits in a domain are, to some extent, a collection of the experiences that students have in that domain, which I described in Chapter 1. Prior research on self-efficacy in physics has not been able to address this central question because it (Kost-Smith, 2011; Sawtelle et al., 2010; Lindstrom and Sharma, 2011; Cavallo et al., 2004), like most research on self-efficacy (Sawtelle et al., 2012a), has not directly measured self-efficacy in the process of learning. Therefore the prior research has not addressed the possibility that the negative shift is caused by experiences outside of the learning process. The one investigation that did measure self-efficacy in the process of learning (Sawtelle et al., 2012a) was not designed to show causal relationships. Furthermore, these studies on self-efficacy in physics have only investigated self-efficacy in physics courses, leaving open the possibility that the negative shift in physics is an epiphenomenon of a larger trend in introductory college STEM courses.

Further, looking at the effects of courses on students' affective traits through the perspective of experiences is a paradigm shift that can support educators to improve student outcomes. The retrospective trait measures that are commonly used to investigate self-efficacy and other affective traits situate the effect within the individual since traits, by their definition, belong to and reside within the individual. Because traits are so personal, educators may feel that they have little power to influence something that resides so seemingly wholly within the individual. This discomfort of educators confronted with affect may in part explain why Pintrich et al. (1993) found it necessary to explain that affect and cognition are inherently interlinked and it may also, in part, explain why affect has received so little attention in college settings (Pekrun and Stephens, 2010). States, in contrast, situates the effect as arising between the environment and the individual. This shifts the focus of affect from seemingly wholly within the individual to, at least in part, arising in response to the learning environment. Creating and maintaining the learning environment is well within the scope of educators' responsibilities. Thus, focusing on the role of the environment in students' affect may empower educators to make changes to the learning environment for the explicit purpose of improving students' affective outcomes.

4.1 Purpose of the research

Given the central focus just outlined, the purpose of this chapter was to investigate if the self-efficacy states students experienced in the focal IE physics course were much lower than those they experienced in other STEM courses. This difference would be consistent with physics instruction harming students' self-efficacy traits and other STEM courses supporting students' self-efficacy traits. Furthermore, this difference would locate that harm within the physics-learning environment. The self-efficacy states in the focal IE physics course being different than those in other STEM courses would also be evidence against the negative effect of physics instruction on self-efficacy traits being an epiphenomenon of a larger trend throughout introductory STEM instruction.

4.2 Research Questions

Negative shifts in self-efficacy traits are consistently measured from pre to post instruction in introductory physics courses, but positive shifts are consistently measured in other introductory STEM courses. And, as I will further describe in Chapter 5, there were negative shifts in students' physics self-efficacy traits in the focal IE physics course. Assuming self-efficacy traits in a domain develop in response to students' experiences in that domain, these negative shifts in physics, and only physics, indicate that self-efficacy states in physics were likely much lower than those experienced in other STEM courses. This possibility forms the basis the central question that I sought to answer in this chapter, which was the extent to which the negative shift in students' self-efficacy traits in physics was caused by their experiences learning physics. To answer this central question I asked two research questions. The first compared self-efficacy states between IE physics and other STEM courses. The second compared self-efficacy states between the different components of the IE physics course. I report on each in turn.

1) To what extent was the focal IE physics course experienced with lower self-efficacy than other STEM courses?

If students experienced much lower self-efficacy states in the focal IE physics course than in their other STEM courses this would locate the negative effect on students' self-efficacy within the process of learning physics. Furthermore, lower self-efficacy states in physics than in other STEM courses would be evidence against the negative effect on students' self-efficacy traits in physics being an epiphenomenon of a larger trend.

In order to further understand the experience in the focal IE physics course I also investigated the extent to which the experience varied across the different components of the focal IE physics course: lecture, lab, recitation and homework. I reasoned that if the components were all experienced similarly then this would indicate that some characteristic that they all shared in common, such as the physics content, the IE pedagogy or both, was potentially the cause of any differences in how the focal IE physics course was experienced. If, in contrast, some components were experienced with lower self-efficacy than others, the list of potential causes would be narrowed to specific aspects of course components where the effect occurred, such as lab work, tutorials or answering clicker questions. Therefore, to further investigate the potential causes of any difference in how the focal IE physics course was experienced I asked:

2) How different were students' self-efficacy states in the four components of the focal IE physics course?

4.3 Designing the research to make comparisons, check reliability and test representativeness

Chapter 2 describes how I collected 1440 measures of self-efficacy state and the three complementary states from 33 students throughout their daily lives for two different weeks. This allowed me to investigate the first research question by comparing self-efficacy experiences in the focal IE physics course to the self-efficacy experiences in other STEM courses. Following from the design laid out in Chapter 2, I also designed the research to check the reliability of any differences in self-efficacy states that were measured and to check the representativeness of the self-efficacy states that were measured. This overall design structure is laid out in Figure 4.1.

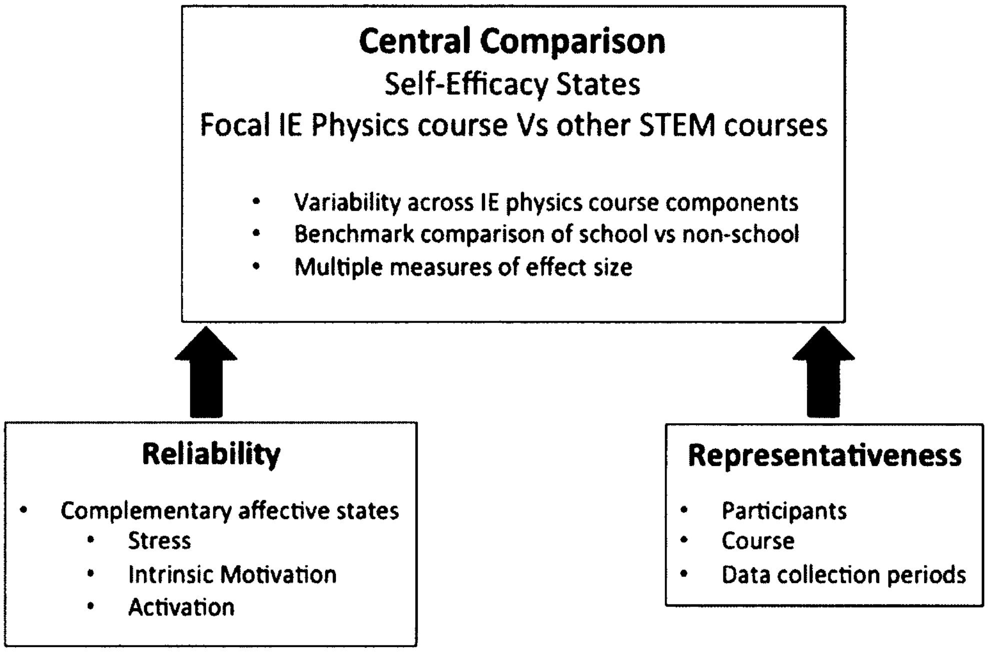


Figure 4.1: The structure of the design of the research. The primary design focus was making and informing the central comparison. I also designed the research to provide checks for the reliability of any differences found in the central comparison and the representativeness of any differences measured as part of a broader trend in IE physics instruction.

The central comparison of the research, comparing self-efficacy states in the focal IE physics course to those in other STEM courses, allowed me to investigate if the cause of the negative shift in self-efficacy traits that is commonly found in physics courses lies within the experience of learning physics. I designed the research to support interpreting the size of any differences found in the central comparison by using differences in self-efficacy states between school and non-school activities as a benchmark for very large differences. I reasoned that school and non-school were more different than any school activities would be and that this difference would act as a benchmark for very large differences. In addition to comparing the differences in the means I also compared the frequency of extreme self-efficacy states, i.e. very high and very low states. I made this comparison because I reasoned that these extremes of experience were much more likely to impact students' self-efficacy traits than more moderate experiences were.

To identify potential sources of low self-efficacy states in the focal IE physics course I compared self-efficacy states across the four components of the focal IE physics course (lab, lecture, recitation and homework). I checked the reliability of any differences in self-efficacy states that I found in either of these comparisons using the three complementary affective states as detailed in Chapter 2. Large differences in self-efficacy states should have had corresponding differences in stress and intrinsic motivation.

I investigated the representativeness of the samples of self-efficacy state experiences in three distinct ways in order to evaluate the extent to which any findings in the focal IE physics course could be generalized to other IE physics courses, as detailed in Chapter 2. First, I compared the ESM participants to the non-participants using their course grades. Second, I compared the affective states in the two different weeks of data collection to each other. Third, I compared students pre and post course traits in the focal IE physics course to those in other IE physics courses investigated by Kost et al. (2009b) and Kost-Smith (2011). The results of this last comparison are presented and discussed in Chapter 5.

An overall design goal of this research was to investigate the validity of the self-efficacy state construct. I did this in Chapter 3 through principle components analysis and the relationships between self-efficacy states and the three complementary states. In this chapter I report on the three complementary states when comparing means between activities in order to investigate the extent to which differences in the means for the complementary states were consistent with the differences in the means for the self-efficacy states.

4.4 Procedures of Analysis

The statistical analyses and other methods used in this chapter are described in detail in Chapter 2. All assumptions for these analyses were tested and met with the exception of those described in Chapter 2.

4.4.1 Central comparison, supporting comparisons and evidence of validity for these comparisons

To make the central comparison of self-efficacy states in the focal IE physics course to those in other STEM courses I first compared the means for self-efficacy states in these two activities to identify if differences existed. I then used the three-step statistical process (MANOVA, ANOVA, Tamhane's T2) to determine if these differences were reliable. The three-step statistical process is described in Chapter 2. In using this process I tested the differences for all four affective constructs across all four activities. This process allowed me to test the reliability of: the central comparison for the self-efficacy states, the validity comparisons for the complementary states and the benchmark comparisons for school versus non-school, with the minimal amount of statistical analysis. I used means, Cohen's d and histograms to characterize the size of the differences in these comparisons, as described in Chapter 2. I compared the difference in self-efficacy states between physics and other STEM courses to the difference between school and non-school activities because the latter comparison acted as a benchmark for large differences.

I compared the means for each of the four affective states in each component of the focal IE physics course to identify how large the differences in experience between the course components were. I used the three-step statistical process to identify if any reliable differences existed in how the physics course components

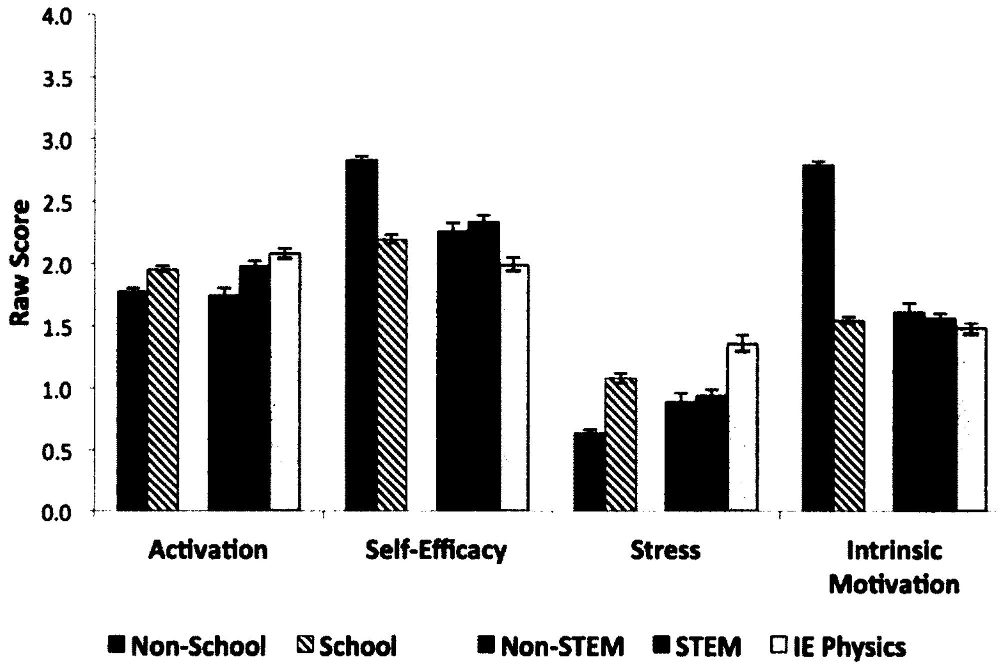


Figure 4.2: ESM data for all participants in all four activities for all four affective states.

were experienced. I tested all four affective state constructs to test the validity of any differences in self-efficacy states that I found.

4.4.2 Self-efficacy in school and non-school activities

I used means, Cohen's d and histograms to investigate the size of the differences for self-efficacy states between school and non-school activities. I used means and Cohen's d to describe the size of the differences for the three complementary states in the results section as a check on the the validity of the self-efficacy state construct.

The size of the difference in self-efficacy states between school and non-school activities acted as a benchmark for very large differences in self-efficacy states. Students' self-efficacy states were lower in school (2.19) than in non-school activities (2.86), Figure 4.2. This numerical difference in self-efficacy states between school and non-school was the second largest of the four affective state constructs (0.67). The lower self-efficacy in school was consistent with school being experienced pri-

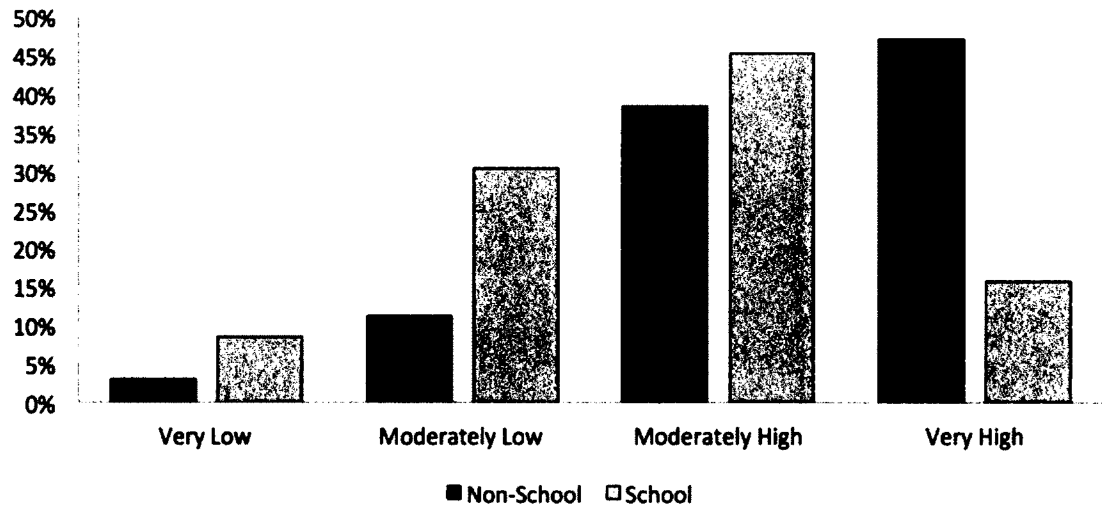


Figure 4.3: Distributions of self-efficacy states in school and non-school activities.

marily with extrinsic motivation and with greater stress than non-school activities. The 0.67 difference in the means between the self-efficacy states in school and non-school was a large effect size, d=0.75. This large difference supports my use of the difference between school and non-school as a benchmark for large differences in self-efficacy states. The distributions of self-efficacy states in school and non-school, Figure 4.3, provided additional benchmarks. Very low self-efficacy states occurred 2.8 times as frequently and very high self-efficacy states occurred 0.33 times as frequently in school activities compared to non-school activities.

4.4.3 Representativeness of the sample

I analyzed how representative the ESM participants were of the course population by comparing the final course grades of 31 of the 33 ESM participants to the other 191 students who received final course grades. I used a two-tailed T-test to verify the reliability of any differences between the ESM participants and non-participants. Limiting this analysis to just course grades maximized the number of students included in the analysis. This increased the likelihood of identifying reliable differences between the students in the course who did and did not participate by increasing the statistical power of the analysis.

I investigated how representative the self-efficacy states I sampled were of all of students' experiences in the course by comparing the two weeks of data collection. I compared the means for all four affective states for each activity between the two data collections. I then used the three-step statistical analysis to see if there was a reliable overall difference between the two data collections, a main effect, or a reliable difference specific to one of the activities, an interaction.

I investigated how representative the course was of IE physics in general by comparing the means for male and female students and the gender differences in means for all seven trait measures in the focal IE physics course to those in the IE physics courses studied by Kost et al. (2009b) and Kost-Smith (2011). I used Cohen's d to determine the size of the differences. For differences with $d > 0.2$ I discussed the size of these differences in terms of that specific measures to further characterize the size of the difference. I do not present these results in this chapter; instead I cover them in detail in Chapter 5.

4.5 Results

4.5.1 Self-efficacy states in the focal IE physics course compared to other STEM courses

The focal IE physics course was experienced with the lowest self-efficacy states (1.99) of all four activities, Figure 4.2 and Table 4.1. This was numerically lower (0.34) than the self-efficacy states experienced in the other STEM course (2.33), and this difference was approximately half of the size of the benchmark (0.67) for very large differences. In comparison, the difference for mean self-efficacy states between STEM and non-STEM courses was numerically much smaller (0.08). The

Table 4.1: Means and standard deviations for the four affective constructs in each activity.

	Non-School		School		Non-STEM		STEM		IE Physics	
N	816		624		161		233		230	
	M.	S.D.	M.	S.D.	M.	S.D.	M.	S.D.	M.	S.D.
Self-Efficacy	2.83	0.79	2.19	0.80	2.26	0.81	2.34	0.72	1.99	0.84
Stress	0.64	0.86	1.08	0.96	0.89	0.87	0.94	0.89	1.36	1.03
Intrinsic Mot.	2.79	0.77	1.55	0.69	1.62	0.79	1.56	0.63	1.48	0.67
Activation	1.77	0.82	1.96	0.69	1.74	0.71	1.98	0.66	2.08	0.66

numerically higher stress experienced in the focal IE physics (1.36) than in other STEM courses (0.94) was consistent with the numerically lower self-efficacy states. The difference in motivation states between physics and STEM was very small. This small difference may have resulted from a ceiling effect of extrinsic motivation in school given that it was very different than non-school activities and it was very consistent within school. Even though the difference for motivation was small it was still consistent with the lower self-efficacy states in physics in that motivation states were more extrinsic in physics than in other STEM courses.

I tested the reliability of the differences in affective states between the four activities using a 1X4 MANOVA. The four affective states were the dependent variables, and activity was the independent variable. The MANOVA showed a statistically significant effect of activity overall $F(4,1435) = 328$, $p < 0.001$. Subsequent one-way ANOVAs showed that all four affect variables had statistically significant difference in means between the four different activities: self-efficacy $F(3,1436) = 85.1$, $p < 0.001$, activation $F(3,1436) = 13.3$, $p < 0.001$, stress $F(3,1436) = 41.1$, $p < 0.001$ and intrinsic motivation $F(3,1436) = 338.3$, $p < 0.001$.

Post-hoc tests comparing means for the various activities using Tamhane's T2, Table 4.2, indicated that the self-efficacy states experienced in physics were statisti-

Table 4.2: Comparisons of the four affective states for non-school versus school and for STEM versus physics.

	Non-School - School		STEM -Physics		STEM V Physics
	Mean	*d*	Mean	*d*	p
Self-Efficacy	0.64	0.75	0.35	0.44	<0.001
Stress	-0.44	-0.48	-0.43	-0.44	<0.001
Intrinsic Motivation	1.24	1.30	0.08	0.13	0.69
Activation	-0.19	-0.24	-0.10	-0.15	0.69

cally significantly lower than the self-efficacy states experienced in the other STEM courses. This was the central result of this chapter, showing that the depression in self-efficacy was not part of an epiphenomenon and locating the depression of self-efficacy in the physics-learning environment.

The effect size of the difference in self-efficacy states between the focal IE physics course and the other STEM courses (d=0.44) was moderate in size and smaller than the benchmark measure (d=0.75). In contrast, the relative likelihoods for the very high and very low self-efficacy states in the focal IE physics course compared to in the other STEM courses, Figure 4.4, were much closer to the benchmark measures. Very low self-efficacy states occurred 13.9% of the time in the focal IE physics course and occurred 2.7 times more frequently than in the other STEM courses. Very high self-efficacy states occurred 9.6% of the time in physics and occurred 0.50 times as frequently as in the other STEM courses. Thus, the ratio of low self-efficacy states in physics to other courses was very close to the benchmark measure of 2.8, and the ratio for very high self-efficacy states was also close to the benchmark measure of 0.33.

Taken together the measures of the size of the difference in self-efficacy states between physics and other STEM courses characterize the differences as being moderate to large in size. The differences in the means and effect size indicated a moderate difference, while the frequencies and relative frequencies of the extreme

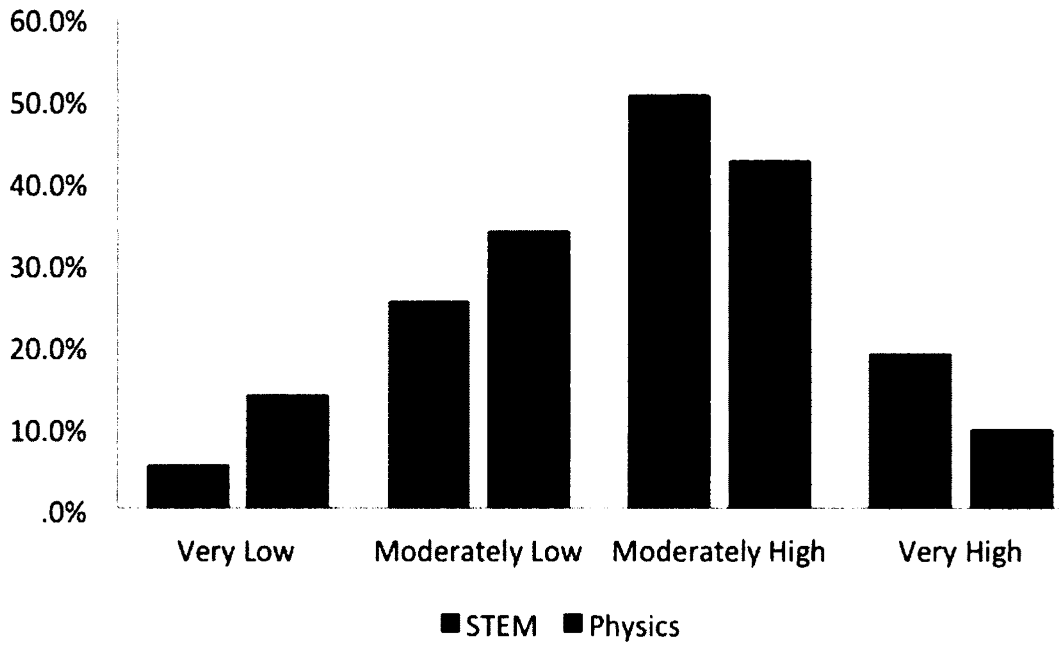


Figure 4.4: Distributions of self-efficacy states in physics and in other STEM courses.

self-efficacy states in the focal IE physics course indicated a large difference. The similarity of these activities and the prevalence and much greater frequency of very low self-efficacy states, which likely harm student self-efficacy traits, indicated that the negative shift in physics self-efficacy traits was caused by students' experiences in the physics-learning environment.

4.5.2 Differences within the focal IE physics course

The lowest average self-efficacy states were experienced while in lecture (1.96) followed by homework (1.98). These states were only slightly numerically lower than the average self-efficacy states experienced in laboratory (2.02) and in recitation (2.05). I investigated the reliability of these differences in experience across the components for the physics course with a 1x4 MANOVA with dependent variables of the four affective constructs and one independent variable for the four physics course components: lab, lecture, recitation and homework. The two weeks of data collection were combined for this analysis. Results of the omnibus analysis showed a statistically significant difference in how the course components were experienced, F(4, 223) = 2.421, p = 0.049. Subsequent ANOVA's showed no significant differences between any of the course components for each of the affective constructs: activation F(3, 224)=1.907, p=0.129, self-efficacy F(3, 224)=0.122, p=0.947, stress F(3, 224)=1.092, p=0.353 and intrinsic motivation F(3, 224)=1.470, p=0.224. Activation, which had the lowest p value and largest numerical differences, had a lower mean value in the lecture component, 0.2-0.3 points, than in the other course components, but all three two tailed T-tests indicated that these differences were not reliable, p>0.20. This lack of any reliable differences in affective states between the course components indicated that the source of the low self-efficacy states was common to all of the components of the focal IE physics course.

4.5.3 Representativeness of the weeks in which data was collected

The average self-efficacy states experienced in the physics course during the first week (2.00) was numerically similar to that experienced during the second week (1.99). I investigated the reliability of differences between the two weeks of data collection using a 2X4 MANOVA with the four affective states as dependent variables and independent variables for the data collection period and the activity. The MANOVA showed that there were no statistically significant overall difference in the affective states between the two data collection periods, F(4,1429)=0.793, p=0.530, and that there was no interaction between the activity and the data collection period, F(4,1431)=1.525, p=0.192, which indicated there wasn't a difference specific to some of the activities. This indicated that the experience in the two weeks of data collection was similar and therefore that the experience throughout the physics course, at least in those times distant from exams, was represented by the experiences measured here. Given that exams were unlikely to increase students' self-efficacy, or decrease their stress, these experiences were likely a conservative measure of students' experiences in physics overall.

4.5.4 Representativeness of the participants

Grades were numerically higher for ESM participants, M=3.05, SD=0.98, which represented an average grade of B, than for non-participants, M=2.31, SD= 1.31, which represented an average grade of C+. Results from the two-tailed T-test indicated that these differences were statistically significant, T(220)= 2.995, p=0.003. This result indicated that there was a sample bias toward higher achieving students. Given that self-efficacy and achievement are positively related this bias likely resulted in a conservative measurement of self-efficacy states, as higher achieving students would tend to experience higher self-efficacy states. Furthermore, the within-subjects design of the study meant that sample bias would not have posed a very large problem for comparing physics SE states to those of other STEM courses.

4.6 Discussion

Students' much lower self-efficacy states in the focal IE physics course than in other STEM courses indicated that the negative shift in self-efficacy traits in physics was caused by the physics-learning environment. The size of the difference in self-efficacy states between physics and other STEM courses was moderate to large, and the measurement located the negative effect within the experience of instruction. Very low self-efficacy states that likely undermined students' self-efficacy traits, were relatively common in physics, occurring 13.9% of the time, which was 2.7 times more often than in other STEM courses. Very high self-efficacy states that likely maintained or improved self-efficacy traits occurred only 9.6% of the time, which was half as often as in other STEM courses. Furthermore, the consistency of the affective states across the two distinct weeks of the semester suggests that these very low self-efficacy states were a consistent part of the experience in the focal IE physics course. Self-efficacy states being so much lower in physics than in other STEM courses was a strong indication that the negative effect on students' self-efficacy was unique to physics rather than an epiphenomenon of general trend throughout introductory college STEM courses. Furthermore, the practices and outcomes in the focal IE physics course closely aligned with those in other IE physics courses suggesting that this course was representative of IE physics instruction in general. Thus, assuming the representativeness of the focal course and given the consistent finding of negative shifts in self-efficacy traits in physics courses, it appears that the experience of learning physics in IE courses harms students' beliefs in their ability to learn and do physics. This seems to be the case even in courses using research-based materials and practices that support much greater learning than traditional lecture-based courses.

The physics-learning environment includes a wide range of variables. These variables can be separated by the degree to which the instructor or the institution has control over them. For example, instructors have a choice in what content they teach and how they teach it; choosing to use either traditional physics instruction or research-based physics instruction. Instructors have less control over who enrolls in their courses, and subsequently they have little control over the traits of their students. The focal IE physics course was experienced with much lower self-efficacy than other STEM courses, indicating that it was something specific to physics instruction that was detrimental to students' self-efficacy. Therefore sources of these low self-efficacy states were likely not the physical classroom, the inclusion of laboratories or homework since these are common characteristics of other STEM courses. Instead, important aspects of the physics learning-environment were those specific to physics. These important aspects may include how the physics content was taught and the expectations established in this physics course, which the instructor has some control over. Alternatively the important aspects may be largely outside of the instructor's control, including inherent aspects of the physics content or larger social trends such as limited interest in and exposure to physics or stereotypes about who can and who cannot learn physics. In the next section I discuss some of these possible causes of students poor self-efficacy states in physics.

4.7 Possible causes of the low self-efficacy states in the focal IE physics course

I separate the possible causes of the low self-efficacy states that students experienced in the focal IE physics course into three different but interacting categories. In the following I will tentatively explore two of these categories: the physics content and the IE pedagogy. The third category, gender, I explore in Chapter 5 because of the principal role that I found gender played in the self-efficacy states that students experienced in physics. Here I will explore the possible sources behind students' poor self-efficacy states that are related to the physics content or the IE pedagogy in order to inform future research that could investigate these relationships.

One possibility is that the conceptual and mathematical demands of the physics content may distinguish it from other STEM courses, making students feel less efficacious. However, mathematical demands can largely be ruled out as a categorical cause for the lower self-efficacy in physics because the mean self-efficacy states experienced in mathematics courses (2.36) was similar to that in all STEM courses other than physics (2.34) and was much higher than the average self-efficacy states experienced in the focal IE physics course (1.99). However, it is possible that the mathematical demands of the physics course were different than in math courses. The way that mathematics is presented and used in physics is tied to conceptual models of physical systems, whereas in mathematics courses, procedural and schema-based knowledge involving symbol manipulations may take greater precedence. Therefore, in a physics course additional demands may interact with the mathematical demands to cause students to experience low self-efficacy states.

Conceptual demands of learning physics may be a source of students' poor self-efficacy states in the focal IE physics course. Students do not start physics instruction as blank slates. They live in a physical world and have developed a stable set of informal conceptions about how that physical world works based on their experiences. Student's informal conceptions form a relatively coherent, logical and robust framework for understanding the physical world (Vosniadou and Skopeliti, 2014). This framework is deeply ingrained due to it being based on and confirmed by experiences throughout daily life (Vosniadou and Skopeliti, 2014). Subsequently, physics instruction creates a conflict between students' informal concepts and formal physics concepts. This conflict is not unique to physics as it is the case for all STEM learning (Hammer, 1996; Vosniadou and Skopeliti, 2014) that students start with informal conceptions. Physics, however, is the subject area in which students have the greatest number of informal conceptions (Duit and Treagust, 2003; Stewart et al., 2007). The breadth of students informal conceptions and their basis in daily life may have been one source of the much lower self-efficacy states in the focal IE physics course.

Because I investigated only one physics course, my results cannot inform the categorical role of the courses explicit focus on conceptual learning in the poor self-efficacy states students experience. It is an open question as to how students' experience differs between research-based and traditional physics instruction. The high levels of learning in this and other IE physics courses should provide greater opportunities for students to experience high self-efficacy states. However, the focus on conceptual learning may have undermined students' self-efficacy by consistently confronting students with the discrepancy between their informal concepts and formal physics concepts. Future investigations should seek to identify the relationships between conceptual knowledge development and self-efficacy. These studies are warranted given that the poor self-efficacy states in the focal IE physics course are an indication that research-based materials and strategies do not provide students with the affective resources and experiences to develop self-efficacy traits. Undermining students' self-efficacy is problematic since self-efficacy traits are important for students' long-term and short-term outcomes, including their conceptual learning (Sinatra, 2005)

4.8 Conclusion

Introductory physics courses act as a gateway to physics degrees and most other STEM degrees. In the most common forms of introductory physics courses students' self-efficacy decreases. The low self-efficacy states in the focal IE physics course and the the similarity of that course to other IE physics courses indicates that this negative shift is caused by the physics-learning environment in these courses. Subsequently, physics instruction is closing the path to STEM careers for many students. Indeed, the low self-efficacy states experienced in physics may be an important reason why physics is one of the smallest STEM disciplines. Undermining students' self-efficacy undermines their ability to learn physics content and pass physics courses, which are both necessary for students to persist in physics majors. Decreased self-efficacy causes students to set lower standards for themselves, use less efficient strategies and achieve lower levels of intellectual performance (Bouffard-Bouchard, 1990). This effect has been shown to occur in physics courses where self-efficacy predicts student performance (Sawtelle et al., 2012b; Kost-Smith, 2011). A second reason why low self-efficacy would tend to reduce physics enrollment is that self-efficacy traits are an important predictor of the college major students choose

(Betz and Hackett, 1983; Marra and Bogue, 2009). It is reasonable that students will not choose to pursue physics majors if they do not believe they can succeed in learning and doing physics. If physics is going to recruit and retain students then it must support them in both learning the physics content and developing strong self-efficacy traits; two goals that are mutually supportive of one another.

Chapter 5
GENDER, EXPERIENCE AND SELF-EFFICACY IN INTRODUCTORY PHYSICS

This chapter has been accepted for publication in *Physical Review Special Topics - Physics Education Research*. The authorship of this article is: Jayson Nissen and Jonathan Shemwell. This chapter is a verbatim copy of the submitted article and, subsequently, has several redundancies with other chapters in this dissertation.

5.1 Abstract

There is growing evidence of persistent gender achievement gaps in university physics instruction, not only for learning physics content, but also for developing productive attitudes and beliefs about learning physics. These gaps occur in both traditional and interactive-engagement (IE) styles of physics instruction. We investigated one gender gap in the area of attitudes and beliefs. This was men and women's physics self-efficacy, which comprises student's thoughts and feelings about their capabilities to succeed as learners in physics. According to extant research using pre and post course surveys, the self-efficacy of both men and women tends to be reduced after taking traditional and IE physics courses. Moreover, self-efficacy is reduced further for women than for men. However, it remains unclear from these studies whether this gender difference is caused by physics instruction. It may be, for instance, that the greater reduction of women's self-efficacy in physics merely reflects a broader trend in university education that has little to do with physics per se. We investigated this and other alternative causes, using an in-the-moment measurement technique called the Experience Sampling Method (ESM). We used ESM to collect multiple samples of university students' feelings of self-efficacy dur-

ing four types of activity for two one-week periods: (1) an introductory IE physics course, (2) students' other introductory STEM courses, (3) their non-STEM courses and (4) their activities outside of school. We found that women experienced the IE physics course with lower self-efficacy than men, but for the other three activity types, women's self-efficacy was not reliably different from men's. We therefore concluded that the experience of physics instruction in the IE physics course depressed women's self-efficacy. Using complementary measures showing the IE physics course to be similar to others in which gendered self-efficacy effects have been consistently observed, we further concluded that IE physics instruction in general is likely to be detrimental to women's self-efficacy. Consequently, there is a clear need to redress this inequity in IE physics, and probably also in traditional instruction.

5.2 Introduction

Over the last 60 years, physics has lagged behind other Science, Technology, Engineering and Mathematics (STEM) disciplines in the proportion of women who pursue undergraduate degrees. For many STEM disciplines, the number of women relative to men is now at or near parity. As examples, between 2000 and 2010 women made up 50% of degree recipients in chemistry and 41% in mathematics. However, during this same period only 21% of bachelors degrees in physics were received by women (National Science Foundation, 2012).

One reason why so few women may be pursuing physics degrees is that the physics-learning environment preferentially favors male students over female students. This possibility is backed by research showing persistent differences in how women and men experience physics in which women are disadvantaged. In introductory courses, women tend to both start out and end at lower levels of conceptual knowledge than men (Madsen et al., 2013; Kost et al., 2009a). Furthermore, women tend to have less productive attitudes about learning physics, including interest, sense making effort and problem solving confidence (Kost et al., 2009a). For both conceptual knowledge and attitudes, these gender differences increase from pre to post course measurement (Kost et al., 2009a; Madsen et al., 2013).

The gender gap in attitudes and beliefs about physics learning also extends to self-efficacy, our subject here. Self-efficacy is the belief in one's ability to succeed in a given domain (Bandura, 1997). It is an important predictor of academic performance and persistence, both in general (Multon et al., 1991), and in introductory physics courses (Sawtelle et al., 2012b; Kost-Smith, 2011). Kost-Smith (2011) found that women entered introductory physics courses with lower self-efficacy than men, and this difference increased from pre to post course. Sawtelle et al. (2010) obtained the same result in lecture-based physics courses, as did Cavallo et al. (2004) and Lindstrom and Sharma (2011).

While it seems fairly clear that there is a gender gap in self-efficacy in physics, it remains an open question whether physics instruction somehow causes this inequity. This is a very important question and the central one of the present study. It may be, for instance, that the negative shift in women's self-efficacy that is consistently observed in physics is not unique to physics courses. Rather, this shift may be an epiphenomenon, or secondary effect, of a broader trend that would tend to occur in most courses, or perhaps most STEM courses. So long as this and other broad-based causes of the inequity (these are discussed later in this article) cannot be ruled out, then there is no particular urgency to redress it in physics courses. However, if gender differences in self-efficacy could be shown to be caused by physics instruction, then there would be an obvious need for concerted action within the physics community to bring about more equitable classroom experiences. The purpose of the present study is to see whether this and other explanations could be ruled out, thus resting the source of the gender inequity more squarely on physics instruction.

We engaged with the question of causality by measuring men's and women's feelings of self-efficacy as they were learning in physics and in other STEM and non-STEM courses over two weeks of instruction. The measurement used an established quantitative technique called the experience sampling method (ESM) in which students responded to a signal to briefly record their thoughts and feelings of self-efficacy in the midst of their activities. We reasoned that if women could be observed to experience lower self-efficacy than men in physics, but not in other courses, physics instruction would have to be seen as a primary cause of the gender difference.

Bandura (1997, p. 3) defined self-efficacy as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments." The term "beliefs" in this definition is potentially confusing because it suggests that self-efficacy is a fairly stable characteristic. However, Bandura considered self-efficacy to be "a dynamic fluctuating property, not a static trait" (Bandura, 1997, p. 406). Furthermore, he recognized that it was highly responsive to a person's behavior and their environment. Therefore, self-efficacy "beliefs" are sometimes better thought of as being dynamic states. On the other hand, Bandura acknowledged that self-efficacy was often associated with habitual patterns of behavior (Bandura, 1997, 2006). Accordingly, self-efficacy is sometimes measured using surveys that ask people to rate their confidence in their ability to accomplish tasks, with the results interpreted as being trait-like characteristics (Bandura, 2006; Kost et al., 2009a). Thus, a person's self-efficacy in physics can be said to go up or down after a semester of instruction.

To encompass both dynamic and stable aspects of self-efficacy, hereafter we refer to (and measure) them as two distinct components. One is the dynamic response that may shift from moment to moment, which we call self-efficacy state. The second is a more stable attitude (or belief) about one?s ability to succeed in a domain, which we refer to as self-efficacy trait.

So far as we can tell, our approach to measuring self-efficacy states separately from traits is unique. Often, researchers skirt the issue by making the sources of self-efficacy the object of measurement, rather than self-efficacy itself. Typically, this is done by asking people to rate their agreement with statements about experiences they had in the domain of interest (Sawtelle et al., 2010; Fencl and Scheel, 2005). Since the sources of self-efficacy are assumed to underlie both dynamic states and longer term patterns, there is no need to distinguish between these aspects. In physics, the Sources of Self-Efficacy in Science Courses- Physics (SOSESC-P) (Fencl and Scheel, 2005) takes this approach, asking students' to reflect on sources of self-efficacy in their experiences of physics instruction. Another instrument used in physics asks students about their self-efficacy via their confidence in their ability to succeed at physics tasks, obtaining a measurement of self-efficacy "beliefs", which we refer to as self-efficacy traits. This is the Physics Self-Efficacy and Identity Survey (PSEIS) (Kost-Smith, 2011). This instrument also includes sources of self-efficacy items from the SOSESC-P.

5.2.1 The gender gap in physics self-efficacy

Leaders in the field have pointed out that the development of coherent attitudes and beliefs about learning and doing science should be a core goal of physics education (Redish et al., 1998; Brewe et al., 2009). Unfortunately, these attitudes and beliefs generally erode over time in physics courses, even when using research-based pedagogies that manifestly benefit learning (Kost et al., 2009a; Kost-Smith et al., 2010; Brewe et al., 2009). Furthermore, there are consistent gender differences in which negative shifts in attitudes and beliefs are larger for women than for men (Kost et al., 2009a; Kost-Smith et al., 2010). These differences extend to self-efficacy. Using the PSEIS, Kost-Smith (2011) demonstrated that women had larger negative shifts than men for both self-efficacy traits and sources of self-efficacy. This result was reliable across four instructors and three different offerings of a research-based introductory course known as interactive engagement physics. The author also found that there was a gender difference in conceptual knowledge at posttest in these courses (typical for physics instruction), and that 12% of the effect was predicted by gender differences in self-efficacy beliefs. Sawtelle et al. (2010) used the SOSESC-P to show that students' physics self-efficacy became less positive across three different semesters of lecture-based physics courses, with the negative shift consistently larger for female students. Corroborating evidence for a reliable gender gap in self-efficacy beliefs in physics, at least in introductory courses, comes from studies of general attitudes and beliefs in physics. Most notably, Kost et al. (2009a) and Kost-Smith et al. (2010) found that women started interactive engagement physics courses with less expert attitudes about learning and doing physics than men, and these differences tended to increase from pre to post instruction.

Sawtelle et al. (2010) pointed out a notable exception to the trend in negative and gendered self-efficacy outcomes in physics. Studying a course that used modeling instruction the authors measured self-efficacy traits at the beginning and end of the course for three different semesters using the SOSESC-P. They found neither positive nor negative shifts in either men's or women's self-efficacy traits. Sawtelle et al. (2012a) investigated the source of this salutary outcome using video and interviews of three students engaged in modeling activities. They showed that creating and working on models in small group settings (the primary instructional mode of the course), provided many opportunities for self-efficacy development, such as when students received positive feedback from their classmates and vicariously from seeing their classmates succeed. They proposed that that these opportunities might be what differentiated modeling instruction from other physics courses with regard to self-efficacy outcomes.

Assuming that modeling instruction in general does not negatively affect women's physics self-efficacy, then the gendered self-efficacy outcomes found in traditional and IE physics would be more likely to be caused by the experience of instruction in those formats and less likely to result from a persistent, broad-based trend in university education. However, Sawtelle and colleagues' (Sawtelle et al., 2010, 2012a) research was not intended to be conclusive about the causes of self-efficacy outcomes in modeling instruction. The authors' second, more diagnostic study in particular was not intended to explain variance, but rather to reveal processes by which self-efficacy could be supported. Thus, direct evidence of the impact of more mainstream (i.e., non-modeling) physics instruction on men's and women's self-efficacy is needed if its gender effects are to be squarely established.

5.2.2 Classroom environments, experiences and gender

Much of the more general education research on differences in how male and female students experience STEM instruction has focused on the tenor of the classroom set by the professor. Using interviews Hall and Sandler (1982) found that women experienced "chilly" classrooms in which male instructors maintained classroom inequalities such as spending disproportionate amounts of time talking to male students and ignoring female students' questions. Seymour and Hewitt (1997) used interviews to show low levels of faculty support and highly competitive environments were typically the starting point of students' paths out of STEM majors. They concluded that many highly capable students, including women, were leaving STEM disciplines because of their poor experiences and not because of an inability to perform well in their coursework. In physics, Mujtaba and Reiss (2013) analyzed high school students' end-of-course surveys to show a gender difference in the level of encouragement to continue in the discipline they felt from their teachers. This measure was correlated with students' intentions to take additional physics courses in the future. Similarly Kost et al. (2009b) used a survey to show that women reported experiencing less support in physics courses, for instance compared to men, women more frequently agreed with the item "I felt like I didn't belong in this course."

The gender inequities just described are relevant to the present study because they are attributed to the experience of learning rather than within a broader gender-based trend. However, these studies used retrospective measures, wherein the distance from the experience of instruction leaves open the possibility of alternative causal factors. In particular, gender differences could arise because men and women focus on different aspects of their experiences in retrospection; not because they actually experienced instruction differently. For example, Hyde et al. (1990) found that women retrospectively reported greater levels of anxiety about mathematics than men. They inferred from this result that women experienced higher levels of anxiety during their mathematics courses. Goetz et al. (2013) called this interpretation into question by combining retrospective reports with an in-the-moment measure of anxiety, which was the experience sampling method (ESM) used in the present study. Retrospective surveys found that women reported higher levels of mathematics anxiety than men, but the in-the-moment measure showed that women and men experienced very similar levels of anxiety. Bieg et al. (2014) referred to this mismatch as a state-trait discrepancy. They found that much of it was explained by students' mathematics self-concept, which they described as a measure of students' feelings of control over their performance in the course. They proposed that the state-trait discrepancy arose when students with lower math self-concepts focused more on their anxiety in retrospective reporting than did students with higher math self-concept.

5.3 Theoretical Framework

As we discussed earlier there are inconsistencies in how self-efficacy is described as both dynamic and static, which we have addressed by separating self-efficacy into states and traits. Albert Bandura (Bandura, 1986, 1997) proposed that internal states are one of the three major classes of determinants in human agency, along with behavior and environment. States arise within the individual, have a complex latent structure consisting of affect, cognition and biological events and are dynamically responsive to both the perceived environment and the individual's behavior (see Figure 5.1) (Bandura, 1997). In contrast traits are the relatively stable patterns of behaviors and internal states, including thoughts and feelings, that habitually occur in different circumstances and contexts (Jackson et al., 2012). We propose thinking of traits as representing the patterns that arise between the three major classes of determinants: internal states, environment and behavior. This framework is consistent with both the definitions of traits and of self-efficacy in that self-efficacy traits are context and situation dependent, tend to be very stable and result in habitual patterns of behavior (Bandura, 1997, 2006).

The development of self-efficacy traits is rooted in experience (Luzzo et al., 1999). High levels of performance support the development of stronger self-efficacy traits which subsequently support future performance (Williams and Williams, 2010). Because self-efficacy states are a measure of experience and, to some degree, a measure of personal performance, we expected a similar causal reciprocal relationship to exist between self-efficacy states and traits measured in the present study. Therefore self-efficacy states experienced in physics integrated over time should produce physics self-efficacy traits. Therefore, because self-efficacy traits predict student performance in physics (Sawtelle et al., 2012b; Kost-Smith, 2011), we viewed very

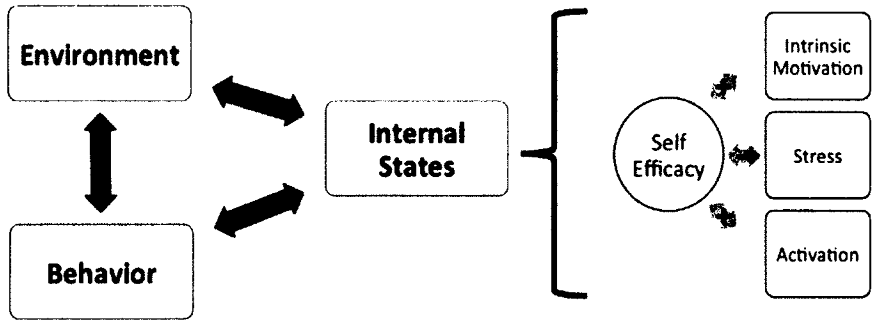


Figure 5.1: The three majors classes of determinants. The three majors classes of determinants according to Social Cognitive Theory are shown on the left. The arrows represent the reciprocal causal relationships that exist between each of the classes. On the right the internal state class is broken down into the four affective states that were measured in this study. The relationships between the self-efficacy state and each of the complementary states are shown in the arrows.

low self-efficacy states experienced in physics as harmful to students' persistence and success in physics.

5.4 Research Questions

The negative shift in women's physics self-efficacy traits measured across introductory physics instruction (Kost-Smith, 2011; Sawtelle et al., 2010; Cavallo et al., 2004; Lindstrom and Sharma, 2011) suggests that there is something about physics instruction that is particularly harmful to women's self-efficacy compared to men's. However, as we discussed earlier rival explanations that this is caused by factors outside of the experience of instruction must be dealt with before locating the cause within physics instruction and not elsewhere. The main body of research to date has primarily focused on post course measures and/or only on physics courses, so it has not effectively addressed these rival possibilities. In order to address the overarching question about the cause of the larger negative effect on women's self-efficacy being situated in the experience of physics instruction we asked two principle research questions:

1. To what extent did women experience IE physics instruction with lower self-efficacy states than men?

2. How did the differences between men's and women's self-efficacy states in IE physics compare to the differences in other STEM and non-STEM courses?

5.5 Methods

5.5.1 Context

The study took place at a four-year public university located in the northeastern part of the United States. The university was the leading research university for the state it served and was a PhD-granting institution in many STEM fields.

We collected data in one interactive engagement (IE) physics course, the focal IE physics course. Interactive engagement promotes (Hake, 1998, p. 65) "conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors." IE has been used to describe courses that use research based teaching practices (Crouch and Mazur, 2001; Turpen and Finkelstein, 2009; Kost et al., 2009a) such as Peer Instruction (Mazur, 1997) and Tutorials in Introductory Physics (McDermott and Shaffer, 2002).

We collected data in an IE physics course, as opposed to a traditional physics course, because we expected IE instruction to be a more conservative measure of gender differences in self-efficacy experiences. We based this decision on IE instruction better supporting student conceptual learning and because gender differences in conceptual knowledge tend to be smaller after IE instruction than after traditional physics instruction (Madsen et al., 2013).

The focal IE physics course met five times in total each week: twice for 50 minutes of lecture with approximately 150 students, twice for 50 minutes of recitation with

24 students and once for 110 minutes of laboratory with 24 students. The instructor of the course was male and had thirty-five years of teaching experience. The data was collected during the fifth year the instructor taught this course. The course was modeled on IE physics courses described in Kost et al. (2009a). Almost all lectures used several conceptual multiple choice questions embedded throughout the lecture, i.e. ConcepTests (Mazur, 1997). Students discussed these questions with their neighbors and the course instructor called on students to explain their reasoning for their answers. Students earned a small portion of their final course grade, 3%, by participating in the ConcepTests. Three mid-term exams and one final exam were given in the lecture portion of the course. There was a weekly homework assignment with a written and an online component. Homework and tests included both conceptual and calculation problems. In the two recitation sections students spent most of their time solving conceptual problems in small groups. One recitation per week made use of a standard set of tutorial lessons (McDermott and Shaffer, 2002). The other recitation used a mix of locally-generated conceptual and calculation physics problems. A graduate teaching assistant (TA) facilitated the recitation periods and the lab. An undergraduate learning assistant (LA) assisted the TA during the recitation. The LA had previously completed the course and was enrolled in a weekly seminar on pedagogy (Otero et al., 2010). The TA and LA were provided weekly training on the content and pedagogy used during recitation. This training emphasized the use of Socratic dialogue to support students in generating their own conceptual understanding in the activities during recitation.

5.5.2 Design

The study used a within-subject design comparing students' self-efficacy states in the focal IE physics course to their self-efficacy states in other introductory STEM courses the students were taking in the same semester (see Figure 5.2). This design enabled us to address five research goals to: provide evidence for answering the research questions (goals 1 and 2), provide validity for that evidence (goals 3 and 4) and generalize the findings (goal 5). The first three goals were addressed with the state data the last two goals were addressed with the trait data.

For goal one we identified any gender differences in the self-efficacy states students experienced during instruction in the focal IE physics course and the size of those differences. For goal two we determined the extent to which any gender differences in self-efficacy states were unique to the focal IE physics course or whether they also occurred in the other courses, potentially as part of a broader trend, by comparing self-efficacy states in the focal IE physics course to those in other STEM courses (Figure 5.2 - left). For goal three we determined the extent to which gender differences in the complementary states were consistent with gender differences in the self-efficacy states experienced in the focal IE physics course.

An important feature of the design was the use of complementary measures to provide validity for any identified gender difference in self-efficacy states that were measured in the focal IE physics course, goals three and four. In addressing goal four self-efficacy trait data was used to validate the self-efficacy state data. The self-efficacy trait measure complemented the self-efficacy state measure, as shown by the dark arrow in Figure 5.2, in that gender differences in self-efficacy states experienced in the IE physics course should show up as gender differences in the means of, and shifts in, self-efficacy traits across the semester. A secondary objective of goal four was the use of complementary trait measures, bottom right of Figure 5.2, to support the validity of the self-efficacy trait measure by identifying the extent to which gender differences were consistent across all traits.

Because we studied only one semester of a single IE physics course we designed the research to collect evidence of how well this focal course represented IE physics courses in general, goal five. We compared the scores and gender differences in scores

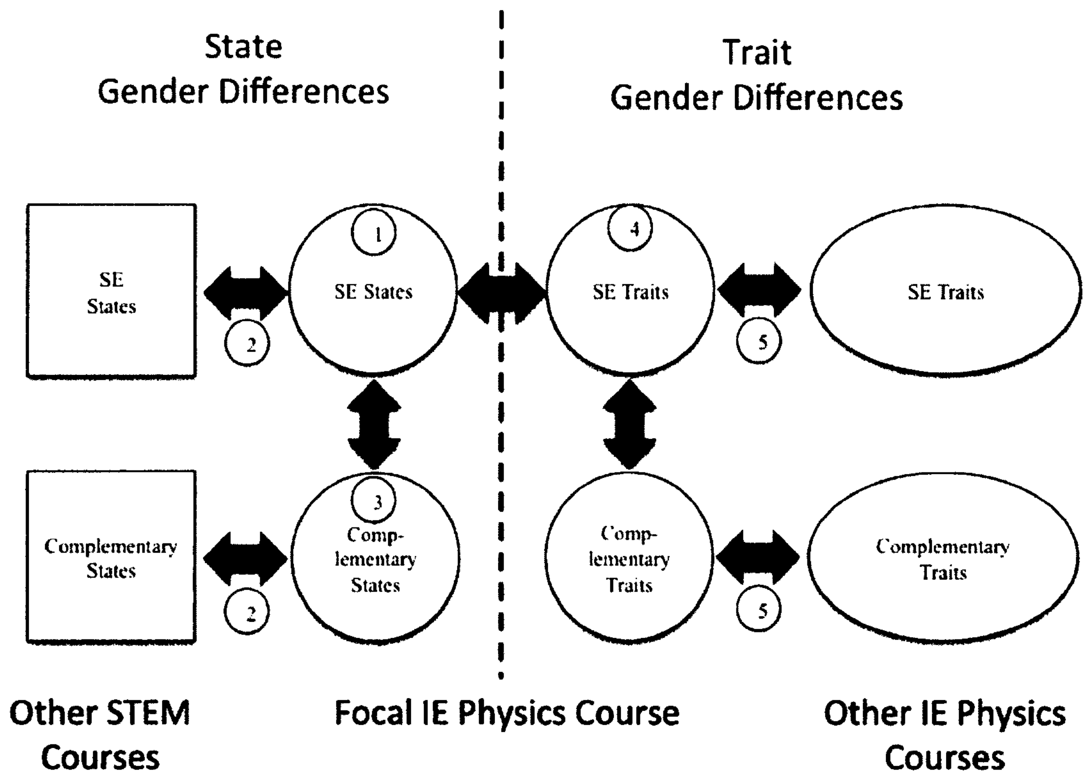


Figure 5.2: Design structure of the research illustrating the five goals of the design. 1) Identify gender differences in self-efficacy states experienced in the IE physics course, 2) identify if the gender differences in state experiences were unique to IE physics, 3) consistency of gender differences in the focal IE physics course for self-efficacy states and the complementary states, 4) consistency between gender differences for self-efficacy states and traits and 5) identify how similar trait outcomes and gender differences were in the focal IE physics course to courses studied by Kost et al. (2009a).

to those scores for similarly designed courses at another institution as reported by Kost et al. (2009a). To do this, we used three different pre-post measures relying on standard survey instruments: self-efficacy traits (Kost-Smith, 2011), attitudes (Adams et al., 2006), and conceptual knowledge (Thornton and Sokoloff, 1998). A fourth comparative measure was course grades. In Figure 5.2, the latter three measures are grouped at the lower right under the collective heading of complementary traits.

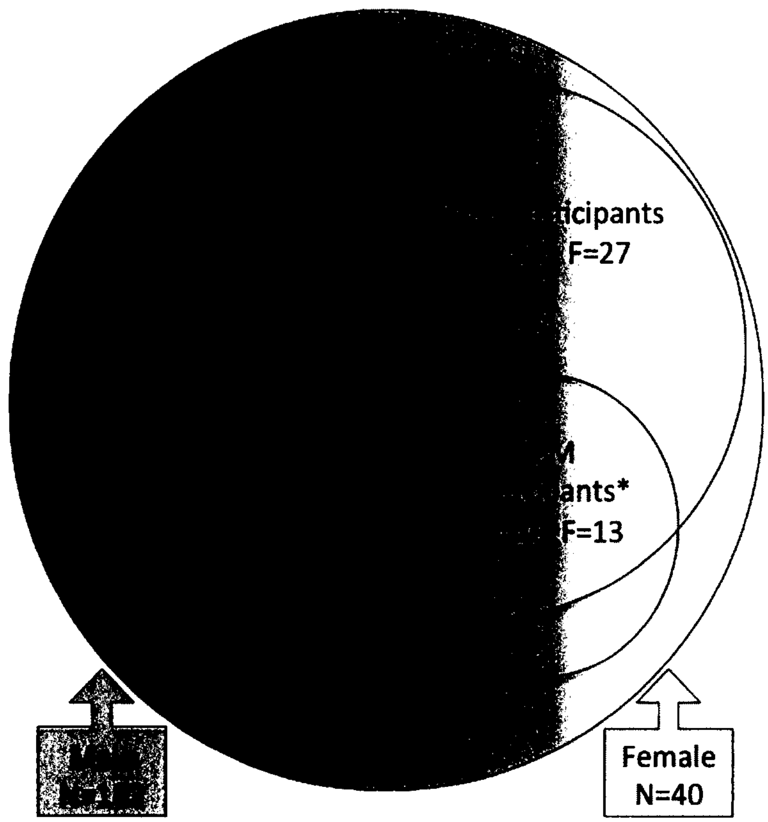


Figure 5.3: Diagram of the overlapping ESM participants and trait participants. *Eight male and five female ESM participants were not trait participants.

Due to the intensive nature of the ESM, it is typical to collect state data from a representative sample of participants in a given context, such as a course or a school, rather than taking data from all students. Using this approach, we conducted ESM with 33 *ESM participants* from a physics course of 242 students. By contrast, trait data was much easier to collect and we obtained this from a larger sample of 117 *trait participants.* Unfortunately, not all 33 ESM participants were part of the 117 trait participants (see Figure 5.3). Therefore, the two overlapping groups were used as independent samples for different purposes. Trait participants were used to represent the effect that the course had on students' traits. ESM participants were used to characterize how students experienced instruction.

5.5.3 Participants

Out of 242 students who started the course, 222 completed the course and received grades. Of these, 40 (18%) were female (see Figure 5.3). Of the 20 students who dropped or withdrew from the course, 5 were female. Of the 117 trait participants 90 were male and 27 (23%) female. Of the 33 ESM participants 20 were male and 13 (39%) were female. Overall, there were 20 ESM participants who were also trait participants, 12 male and 8 female. Two of the female ESM participants withdrew from the course and did not receive final course grades.

ESM participants were recruited from their IE physics course through a brief announcement by the first author describing the research. They were informed that the research was investigating their experiences as college students. All students who wished to participate in the study were allowed to do so. Participants who completed the ESM were given a small amount of extra credit and a stipend of fifty USD.

We defined gender as the self-identification as either male or female.

5.5.4 Instrumentation for trait data collection

We measured students' *self-efficacy traits* in physics by using the twenty 5-point Likert-scale self-efficacy questions from the Physics Self-Efficacy and Identity Survey developed by Kost-Smith (2011). We truncated the name to Physics Self-Efficacy Survey (PSES) because we did not include the identity questions or the sources of self-efficacy questions. The PSES measures self-efficacy across four constructs, but only the overall self-efficacy score was used in this study. We measured students' *attitudes* about learning physics with the Colorado Learning Attitudes about Science Survey (Adams et al., 2006). The CLASS measures eight separate categories of student beliefs compiled from student responses to 42 questions. Responses are coded as favorable, neutral, or unfavorable based on agreement with expert responses. Like the PSES, the CLASS is multidimensional, having eight subconstructs of expert-like response, but it also allows for an aggregate score. We used only the overall favorable score in the present study. We measured students' *conceptual knowledge* in the focal IE physics course with the Force and Motion Conceptual Evaluation (FMCE) (Thornton and Sokoloff, 1998), a 47 question multiple-choice exam. The FMCE was scored out of 37 points following the methods of Thornton et al. (2009) using a spreadsheet developed for that purpose (Wittmann, 2009). We obtained *course grades* for the focal IE physics course from the instructor and analyzed them on a 4.0 scale, such that an A was 4.0 , an A- was 3.7, a B+ was 3.3, etc. This was the scale used at this institution and was the same scale used by Kost et al. (2009a).

5.5.5 Experience Sampling Form

The data collection instrument for ESM studies is a short survey that participants fill out when randomly signaled, or shortly thereafter, about the activity they were engaged in at the moment of the signal. ESM studies typically refer to this instrument as the Experience Sampling Form (ESF). Our ESF was modeled on those used in studies overviewed by Hektner et al. (2007). It was the single side of one standard-sized page split into two sections. The first section asked four free-response questions: (1) the main and (2) the secondary activities students were doing, (3) where they were and (4) what they were thinking about. For the present study, only the first of these free-response items was analyzed. The second portion of the survey, on the right half of the page, consisted of 20 Likert-scale questions. Students indicated the type and level of affect at the moment they were signaled by responding to the question, "How did you feel in the main activity?" which was followed by the twenty emotions. Principle components analysis of all surveys confirmed that 19 of the 20 Likert-scale items reliably loaded onto the four affect constructs as shown in Table 5.1.

Three of the Likert-scale questions, skill, control and success, formed the basis of our self-efficacy state measure. We designed the study to include these questions because control and capability are central attributes of self-efficacy (Bandura, 1997). These feelings have also been statistically grouped in prior ESM studies (Schmidt and Shumow, 2012; Hektner et al., 2007) and principle components analysis confirmed their structure in the present study. The other Likert-scale questions formed the three complementary affective states, activation, intrinsic motivation and stress, which are defined in Table 5.1. Self-efficacy, activation and stress were measured on a unipolar scale from none to extreme. Intrinsic motivation was measured on a bipolar scale from extremely extrinsic to extremely intrinsic (Nissen and Shemwell, 2014).

We used the relationships between self-efficacy and each of the complementary state measures to provide additional validity for the self-efficacy state measure. The relationship between self-efficacy and stress was expected to be negative. When self-efficacy is higher, stress should be lower because self-efficacy is a measure of personal skill and stress arises when skill does not meet the demands of the situation. The relationship between self-efficacy and both activation and intrinsic motivation was expected to be positive. People are more likely to become activated when they feel efficacious (Bandura, 1997) and people are also more likely to internalize motivation for activities that they feel efficacious in (Ryan and Deci, 2000).

5.5.6 Procedures

ESM data was collected for two different seven-day periods during the third week and then again in the tenth week of the semester. These weeks were chosen so as not to fall during an exam or other significant assessment. Signals to fill out the ESF were sent to students' cell phones. These were semi-randomly scheduled across each day such that there was a signal once during each two-hour block between 8

Table 5.1: Affective state constructs, definitions and component questions with construct reliability measures and factor loadings. Italicized questions were asked in a 7-point bipolar format. All other questions were asked in a 5-point unipolar format. Parentheses in the left hand column are (Cronbach's alpha/ percent variance explained). Parentheses in the right hand column are the rotated factor loading for that question.

Construct	Definition	Components
Self-Efficacy (0.76/20.2%)	Dynamically responsive judgments of one's ability to organize and execute the courses of action required to produce given attainments in the activity at hand.	skill (0.79), control (0.68), success (0.82), *difficulty concentrating easy/hard* (-0.51), *confused/clear* (0.52)
Activation (0.87/25.6%)	An elevated level of excitement and involvement in the task, consistent with Thayer (1996) and in contrast to a relaxing state (Thayer, 1967).	determined (0.67), active (0.59), attentive (0.80), alert (0.77), inspired (0.56), *detached/involved* (0.57), concentrating (0.63)
Intrinsic Motivation (0.70/6.1%)	A drive to engage in the activity at hand, derived from within, either because it is personally enjoyable or valuable, as opposed to extrinsic motivation, which is driven by external pressures or rewards (Deci et al., 1999; Csikszentmihalyi, 1975).	*free/constrained* (-0.65), *excited/bored* (-0.64), enjoy (0.57), importance (-0.71)
Stress (0.79/6.6%)	Negative feelings resulting from an individual's perception that they do not have the resources to cope with a perceived situation (Lazarus and Folkman, 1984).	stress (0.83), worry (0.80), frustration (0.71)

am and 10 pm and all signals were greater than 30 minutes apart. A constraint on the schedule was that a signal was scheduled for every physics course meeting, resulting in a higher rate of sampling for physics than for other experiences. We did this to ensure enough samples in physics for a reliable measurement, since students spent less time there than in the other, broader categories of experience. To prepare participants for the first of the two data collection periods, we gave them a one-hour briefing on the data collection procedures.

Surveys for trait measurements were given during the first and last week of the course. The knowledge measurement (FMCE) was done during class. This was not part of student grades, but it was a mandatory class activity for students in attendance. Students took the attitude and self-efficacy surveys (CLASS and PSES) outside of class via an online platform as a part of weekly homework assignments. Students received credit equal to one homework problem for completing each survey. We obtained course grades from the instructor after the course had ended.

5.5.7 Methods of Analysis

ESM data was analyzed to compare means between genders across all four activities for each of the four affective constructs. To check for statistical significance, we used a three-step process beginning with an omnibus multivariate analysis of variance to see if a statistically significant difference in means existed for the gender X activity interaction. Then factorial univariate analysis identified if statistically significant differences existed for the gender X activity interaction on each of the affective constructs. Last, post hoc tests were run to identify statistically significant differences in means between males and females for each of the four affective constructs in each of the four activities.

The ESM data for both the third and tenth weeks of the semester was entered into a spreadsheet database. Principle components analysis was conducted on the raw responses and verified the individual questions aligned with the four expected affective constructs, summarized in Table 5.1. Averaging the component questions of each construct on a 5-point, 0-4, scale created the raw score for each construct. The data for what students were doing was reduced to four activities: non-school, non-STEM, STEM, and IE physics and the two weeks of data collection were combined. Analysis of variance confirmed that no statistically significant differences existed for either of these reductions. Results of these analyses are reported in Chapter 6.

We used Cohen's *d*, histograms of the raw score responses and Z-scores of the affective constructs to interpret the size of the gender differences measured in the focal IE physics course. The histograms allowed comparing the distribution of students' responses across the scale for each affective construct. This supported interpreting the meaningfulness of the differences, for instance in the case where one population never experienced a very high level of a state but the other population frequently experienced that high level.

Z-scores allowed identifying how the experience in physics was situated in students' overall experiences in two ways. First, they allowed showing how males' and females' average experiences in physics compared to their overall experiences, for example bottom 20% or top 10%. Second, they allowed seeing how often physics experiences were above average. To create Z-scores the twenty Likert-scale affect questions were converted to Z-scores for each response based on that participant's mean and standard deviation for that question for that week. This conversion minimized the effects of participants using the scales differently by describing responses as above or below average for that person and scaling the distance from average in units of standard deviation for that person's response to that question (Hektner et al., 2007). Averaging the component question Z-scores created the Z-score for each affective construct.

Each of the four trait measures yielded a single overall score. We compared means for these scores between male and female students for all of the trait participants (i.e., all of the students for whom we had a complete set of trait data, see Figure 5.3). We assessed the effect size of any differences between men's and women's mean scores on each trait measure using Cohen's *d*. To check for statistical significance, we used a two-step process beginning with an omnibus multivariate analysis of variance to see if a statistically significant differences in means existed for gender. Then factorial univariate analysis identified if statistically significant gender differences existed on each measure.

We used results from the trait analysis to assess the similarity of the focal IE physics course to those investigated by Kost et al. (2009a) by comparing means for male and female students on each measure between the two course contexts. In particular, we compared the effect sizes for gender differences to see if the focal course maintained, increased or decreased gender differences in similar ways to other IE physics courses (Kost et al., 2009a).

Representativeness of trait participants was investigated by comparing mean grades of trait participants to mean grades for all other students while controlling for gender using analysis of variance. Assessing the representativeness of ESM participants was more challenging because we sought to balance the number of students included in the analysis with the number of trait measures over which we analyzed the representativeness. First, analysis of variance was used to compare means on all trait measures between ESM participants and non-participants. However, this limited the ESM participants included in the analysis in a biased way and the small N resulted in low statistical power. Therefore, analysis for representativeness of the ESM participants was accomplished by comparing means for ESM participants and non-participants on each trait measure for all students who completed that measure using two tailed T-tests.

Cohen's d was utilized as a measure of the effect size between male and female students for both traits and state experiences as recommended by Rodriguez et al. (2012). Cohen (1977) provided guidelines of small (0.2), medium (0.5) and large (0.8) for interpreting effect sizes for interventions, but he cautions that these are not hard and fast rules. Thus, we used these guidelines for interpreting effect sizes loosely and described the differences in experience as descriptively as possible in order to substantiate the size of those differences.

5.6 Results

In presenting the results we first present the state data and then the trait data. We begin the state results by describing how well ESM participants represented the course population. Next, we present the results for self-efficacy states experienced in the IE physics course compared to other types of courses and day-to-day experiences. This addresses the two research questions and the first three design goals. We begin the trait results by describing the representativeness of the trait participants. Then we present the trait results to address the fourth and fifth design goals: checking the extent to which self-efficacy states in the focal IE physics course were consistent with physics self-efficacy traits and assessing the degree to which the focal IE physics course should be taken as representative of IE physics courses in general.

5.6.1 The representativeness of ESM participants

Differences in traits between students who participated in the ESM and those who did not were tested by comparing the means between ESM participants and non-participants for male and female students for all students who completed each trait measure as shown in Table 5.2. These comparisons showed that, first of all, both the male and female ESM participants in the study were high achieving students in the sense that they learned more conceptual knowledge and had higher grades than other students in the course. While there were no differences in selectivity between men and women with respect to achievement, the other trait measures suggested that men who participated in the ESM might have had especially robust attitudes and self-efficacy traits compared to the other men in the course. Whereas female ESM participants had more novice-like and more malleable attitudes, but similar self-efficacy traits, compared to other women in the course. These differences, or

Table 5.2: Representativeness of male and female ESM participants. Includes traits for ESM participants and non-participants by gender for all students who completed each trait instrument.

		Male Participants			Male Non-Participants			
Measure	Scale	Mean	N	SD	Mean	N	SD	*d*(95% CI)
Pre-FMCE	%	32.9	17	17.5	29.7	163	23.3	0.14 (-0.36, 0.64)
post FMCE	%	69.1	16	26.0	63.2	145	28.7	0.20 (-0.31, 0.72)
Pre-CLASS	%	72.0	19	13.4	62.0	164	16.1	0.64 (0.16, 1.11)
post CLASS	%	64.8	16	13.8	58.3	101	17.5	0.38 (-0.15, 0.91)
Pre-PSES	1-5	3.61	19	0.41	3.43	156	0.59	0.31 (-0.17, 0.79)
post PSES	1-5	3.62	17	0.49	3.41	97	0.73	0.30 (-0.22, 0.81)
Course Grade	0-4	3.05	20	1.11	2.31	162	1.31	0.57 (0.10, 1.04)

		Female Participants			Female Non-Participants			
Measure	Scale	Mean	N	SD	Mean	N	SD	*d*(95% CI)
Pre-FMCE	%	22.7	13	10.3	23.4	31	18.2	-0.04 (-0.69, 0.60)
post FMCE	%	55.6	12	27.8	48.1	28	27.5	0.27 (-0.41, 0.95)
Pre-CLASS	%	55.8	13	19.4	61.8	28	16.6	-0.34 (-0.99, 0.33)
post CLASS	%	46.5	11	14.4	58.5	22	19.0	-0.68 (-1.41, 0.08)
Pre-PSES	1-5	3.23	11	0.62	3.34	29	0.58	-0.19 (-0.88, 0.51)
post PSES	1-5	3.14	10	0.40	3.10	23	0.55	0.08 (-0.67, 0.82)
Course Grade	0-4	3.05	11	0.74	2.34	29	1.35	0.58 (-0.13, 1.28)

biases, provide a caveat for generalizing gender differences in the sample to the course population.

5.6.2 Gender differences in self-efficacy states

The largest gender difference for self-efficacy states occurred in the focal IE physics course where women experienced much lower average self-efficacy (1.57) than men (2.25). There was a much smaller gender difference for mean self-efficacy states in other STEM courses with women having slightly lower means (2.25) than men (2.45). Thus, women experienced the focal IE physics course with much lower self-efficacy than the other STEM courses, whereas the difference was relatively small for men. The second largest gender difference in experience was for intrinsic motivation in the focal IE physics course. Women experienced more extrinsic motivation (1.25) than men (1.61). There was a much smaller gender difference in other STEM courses

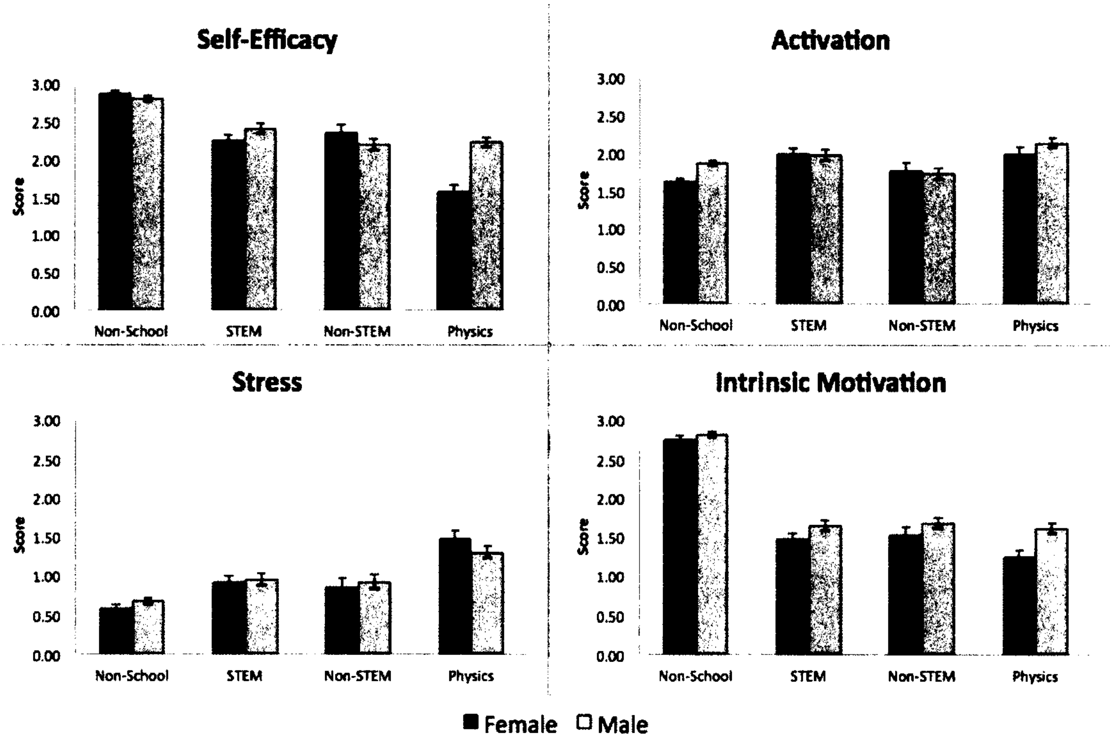


Figure 5.4: Students' affective state experiences by gender and activity. States were measured on a 5 point Likert-scale and ranged from 0, not-at-all, to 4, extremely, for self-efficacy, stress and activation. Intrinsic motivation ranged from extremely extrinsic, 0, to extremely intrinsic, 4. Compared to men, women in IE physics experienced lower self-efficacy, more extrinsic motivation, lower activation, and higher stress. Analysis indicated that the large gender differences for self-efficacy states were unique to the focal IE physics course. Error bars are one standard error.

with women having more extrinsic motivation (1.47) than men (1.64). Similar to self-efficacy, the difference between men's motivation in the focal IE physics course and in other STEM courses was small, whereas women's motivation was much more extrinsic in the focal IE physics course. Consistent with women's lower mean self-efficacy states and more extrinsic-motivation states in the focal IE physics course women also experienced greater stress in physics (1.48) than men (1.30) and lower activation (1.99) than men (2.13).

Analysis of variance was used to determine if any statistically significant differences existed between male and female students' experiences. A 2X4 MANOVA with independent variables for activity and gender and dependent variables for the

Table 5.3: MANOVA and ANOVA results the state data.

	Gender		Activity		Gender X Activity	
	F(df, error df)	p	F(df, error df)	p	F(df, error df)	p
MANOVA	5.37(4, 1429)	<0.001	335(4, 1431)	<0.001	12.8(4, 1431)	<0.001
ANOVA Results						
Self-efficacy	8.02 (1, 1432)	0.005	96.8 (3, 1432)	<0.001	13.4 (3, 1432)	<0.001
Activation	3.02 (1, 1432)	0.082	13.4 (3, 1432)	<0.001	3.02 (3, 1432)	0.029
Stress	0.002 (1, 1432)	0.964	41.9 (3, 1432)	<0.001	1.31 (3, 1432)	0.270
Motivation	15.0 (1, 1432)	<0.001	337 (3, 1432)	<0.001	2.43 (3, 1432)	0.064

four affective constructs identified statistically significant effects for gender, activity and for the gender X activity interaction, Table 5.3. The statistical significance of the gender X activity interaction indicated that there might have been statistically significant differences in experience for male and female students for some of the affective constructs specific to certain activities. This was tested with univariate analysis of variance and was statistically significant on the gender X activity interaction for self-efficacy and activation. The analysis for the activity condition is discussed in Chapter 4 and elsewhere (Nissen and Shemwell, 2014).

Post hoc analysis further investigated the statistical significance of gender differences for each affective construct in each activity using two tailed T-tests. Only two gender differences were statistically significant outside of the focal IE physics course: activation in non-school and motivation in STEM courses, Table 5.4. In addition to the large gender differences for self-efficacy states in the focal IE physics course being statistically significant so was the moderately large difference for intrinsic motivation. The small difference for activation was marginally statistically significant and the small difference for stress was not statistically significant. These results portray a consistent picture of the focal IE physics course having been experienced more negatively by women, with the largest gender difference measured for self-efficacy states. In no other activities were there large or consistent gender differences in experience.

Table 5.4: Gender differences in raw experience across activities and affective constructs. Abbreviations in the first column are: affective construct (Aff), self-efficacy (SE), activation (Act), intrinsic motivation (Mot) and stress (Str).

	Female			Male				
Aff	Mean	SD	N	Mean	SD	N	d	p
				Physics				
SE	2.57	0.82	82	3.23	0.76	148	0.77	< 0.001
Act	2.99	0.58	82	3.13	0.70	148	0.22	0.09
Mot	2.25	0.56	82	2.61	0.69	148	0.53	< 0.001
Str	2.48	1.12	82	2.30	0.97	148	-0.18	0.22
				STEM				
SE	3.25	0.73	107	3.41	0.71	126	0.22	0.09
Act	2.99	0.63	107	2.97	0.69	126	-0.03	0.84
Mot	2.47	0.60	107	2.64	0.65	126	0.27	0.04
Str	1.92	0.92	107	1.95	0.87	126	0.04	0.78
				Non-STEM				
SE	3.36	0.89	62	3.20	0.75	99	-0.20	0.24
Act	2.77	0.67	62	2.73	0.74	99	-0.06	0.71
Mot	2.53	0.91	62	2.68	0.70	99	0.19	0.27
Str	1.85	0.99	62	1.92	0.78	99	0.08	0.64
				Non-School				
SE	3.88	0.92	326	3.80	0.69	490	-0.10	0.21
Act	2.62	0.82	326	2.87	0.81	490	0.31	< 0.001
Mot	3.76	0.77	326	3.81	0.76	490	0.07	0.30
Str	1.59	0.90	326	1.68	0.82	490	0.10	0.15

Z-score transformed data illustrated the size of the difference in men's and women's self-efficacy states experienced in the focal IE physics course. We accomplished this in two steps first by ranking their physics experiences within their overall experiences and second by seeing how often their experiences in physics were above their average overall experience, Z-score = 0. Women's experiences in physics were amongst their worst self-efficacy experiences overall, with a rank of 21%. Men's mean self-efficacy experiences in physics ranked 14 points higher at 35%. Furthermore, with the exception of women's mean self-efficacy states in physics all other mean self-efficacy states in school activities ranked between 35% and 42%; a range half the size of the difference between men's and women's ranks in physics. Men

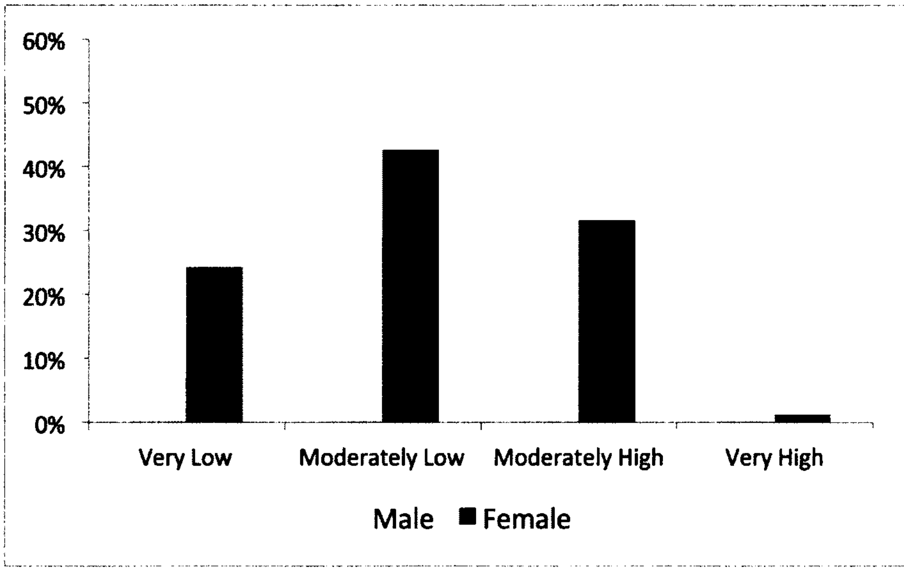


Figure 5.5: Distribution of self-efficacy state experiences in IE physics by gender.

were also two and a half times as likely to have above average self-efficacy states in the focal IE physics course, 28% for men versus 11% for women. And the 17 percent difference in above average self-efficacy states between men and women in physics was much larger than the 10 percent range, 28-38%, of self-efficacy states experienced above average in all school activities excluding women's experiences in the focal IE physics course.

Female students primarily experienced the IE physics course with low self-efficacy whereas male students experiences tended toward high self-efficacy, Figure 5.5. Approximately 1 in 4 of women's experiences were very low self-efficacy, whereby women experienced little to no control, success, or skill. Less than 10% of male students' experiences fell into this very low self-efficacy category. Female students had almost no experiences, 1%, of very high self-efficacy states whereas male students had 14% of their experiences be very high self-efficacy. These differences in the distribution of experience provide further evidence that women experienced much lower levels of self-efficacy states in the focal IE physics course than their male peers did.

5.6.3 Representativeness of trait participants

Male trait participants had higher mean grades (M=2.69, SD=1.28) than male non-participants (M=2.10, SD=1.28). Female trait participants had higher mean grades (M=2.78, SD=1.26) than female non-participants (M=2.05, SD=1.16). There were only small differences between males and females within the trait participant and non-participant groups. Univariate analysis of variance with a dependent variable of course grade and independent variables for trait participation and gender was statistically significant for participation F(1, 218)=8.05, p=0.005 but not for gender, p=0.94, or the gender X participation interaction, p=0.77. Similar to Kost et al. (2009a) those students who completed all trait measures and who make up the data set used to analyze gender differences over-represent high achieving students and this trend was similar for male and female students.

5.6.4 Gender differences in the focal IE physics course for trait measures

Male students started the course with slightly higher self-efficacy traits (3.47) than female students (3.29), Table 5. Self-efficacy traits decreased for both male and female students with a very small shift for male students to a mean of 3.43, and a small shift for female students to a final mean of 3.13. These shifts were small, but the larger negative shift for women was consistent with the much worse self-efficacy states experienced by women in the IE physics course. The larger negative shift for self-efficacy traits for women resulted in the gender gap increasing a small amount from d=0.34 to d=0.47.

Consistent with the gender differences for self-efficacy traits, male trait participants' mean scores were higher than females' on the pre and post measures for all other measures except for course grades. The female trait participants had slightly higher grades than the male participants, 2.78 versus 2.68. Gender differences on all

Table 5.5: Gender differences in trait measures for the focal IE physics course and the courses studied by Kost et al. (2009a) and Kost-Smith (2011). * indicates p < 0.05 and ** indicates p < 0.01.

		Gender Differences in the Focal IE Physics Course						
			Male			Female		
	Scale	Mean	N	SD	Mean	N	SD	*d*(95% CI)
Pre-FMCE	%	30.0	90	23.0	25.5	27	18.0	0.20 (-0.23, 0.63)
post FMCE	%	64.3	90	27.9	52.8	27	27.7	0.41* (-0.02, 0.84)
Pre-CLASS	%	64.1	90	15.2	57.9	27	18.5	0.39** (-0.05, 0.82)
post CLASS	%	60.8	90	17.5	52.8	27	19.4	0.44** (0.01, 0.87)
Pre-PSES	1-5	3.47	90	0.55	3.29	27	0.54	0.34 (-0.10, 0.77)
post PSES	1-5	3.43	90	0.69	3.13	27	0.46	0.47* (0.03, 0.90)
Course Grade	0-4	2.69	90	1.28	2.78	27	1.23	-0.07 (-0.50, 0.36)

		Gender Differences in Other IE Physics Courses						
			Male			Female		
	Scale	Mean	N	SD	Mean	N	SD	*d*(95% CI)
Pre-FMCE	%	32.2	1566	21	22	533	16	0.51 (0.41, 0.61)
post FMCE	%	67.3	1566	27	56.8	533	29	0.38 (0.28, 0.48)
Pre-CLASS	%	65.7	1380	16	63.6	522	15	0.13 (0.03, 0.23)
post CLASS	%	56	-	-	52.3	-	-	-
Pre-PSES	1-5	-	-	-	-	-	-	0.29
post PSES	1-5	-	-	-	-	-	-	0.36
Course Grade	0-4	2.53	2715	0.99	2.41	848	0.92	0.11 (0.03, 0.19)

measures except course grades, were small to medium in size, 0.20 to 0.47, favored male students and increased from pre to post measurement.

We used a 2X7 omnibus MANOVA to identify if there was a main effect of gender on trait measures for the trait participants. A secondary purpose was to identify if there were also effects for participating in the ESM and for ESM X gender interaction that were discussed earlier. Independent variables were student gender and participation in the ESM data. Dependent variables were course grade and the pre and post measures for the FMCE, CLASS and PSES. The MANOVA showed a statistically significant difference for Gender F (7, 107) = 2.85, p=0.009. Statistically significant gender differences identified by the subsequent factorial analysis of variance are indicated in Table 5.

The results for trait measures for both overall scores and gender differences in scores were consistent with the results reported for other IE physics courses (Kost et al., 2009a; Kost-Smith, 2011). Means for the measures were mostly very similar, $d < 0.2$, and there was no consistent pattern of one course having higher means than the other course. For women the differences greater than 0.2 standard deviations were small for the post FMCE, d=0.22, and moderate for both the pre-CLASS, d=0.38, and course grades, d=0.40. With the women in the focal course having higher grades and post FMCE scores but lower CLASS scores. For male students the only noteworthy difference was the small effect on the post CLASS, d=0.27. All other differences were very small. These results indicate mostly small and inconsistent differences between the two courses and indicate that the students in the courses started and ended instruction similarly.

The gender differences and shifts in gender differences were also similar in the two courses. All of the differences favored male students and increased from pre to post instruction. While the gender differences on the PSES were very similar between the focal and other IE courses those on the CLASS and FMCE had some variability. However, this variability can be explained by relatively small differences, on the order of one question, between the means for men and women in the two courses. Subsequently, we concluded that these courses had similar populations of students and that shifts in students' traits from pre to post instruction were similar.

5.7 Discussion

While learning physics, women did not experience high self-efficacy states, as men sometimes did. Instead women frequently experienced low or very low self-efficacy states, and correspondingly, their self-efficacy traits were significantly reduced from pre to post course. Men, by contrast, had very small negative shifts in their self-efficacy traits, consistent with the higher levels of self-efficacy states that they experienced. Furthermore, there was no other activity in which either men or women had such low self-efficacy states as women experienced in the focal IE physics course. Supporting the validity of our self-efficacy measures is the finding that the gender differences we observed were consistent with gender differences on the complementary measures, both state and trait. This is especially true for state measures, which showed that women experienced the focal IE physics course with less activation, more extrinsic motivation and greater stress than men. For traits, similar to self-efficacy, gender differences increased for conceptual knowledge and attitudes about learning physics. Thus, our overall conclusion is that the larger negative shift in women's self-efficacy traits in the physics course was caused by the experience of instruction in which women's self-efficacy states were much worse than men's.

Earlier we brought up three alternative explanations for the larger negative shift in women's self-efficacy traits: that they were the result of a broad trend across many college courses, that they were a result of differences in experience in marginal activities in physics learning or that there was no difference in experience only a difference in retrospection. Our findings demonstrate all three alternative explanations had little to do with the disparate effects on women's self-efficacy. Most importantly, there was no indication that the larger negative shift in women's self-efficacy traits was a part of a larger trend. The large gender differences in self-efficacy states only occurred in the focal IE physics course and did not occur in other STEM courses. Secondly, while it is possible that women experience marginal activities in physics learning differently than men, the large differences that we measured for much more common activities make it unlikely that marginal experiences play a more important role than the experience of learning physics that we measured. Lastly, the large gender differences in the experience of learning physics ruled out the possibility that the larger negative shifts were due only to differences in retrospection.

Based on the similarity of the focal IE physics course to other IE courses and the consistency between the large gender differences in self-efficacy states and concurrent larger negative shift in women's self-efficacy traits in the focal course we think it is probable that similar gender differences in self-efficacy states exist in other physics courses using either IE or traditional instruction. This is supported by most investigations revealing that IE and traditional lecture physics courses had larger negative impacts on female students' self-efficacy traits (Kost-Smith, 2011; Sawtelle et al., 2010; Cavallo et al., 2004; Lindstrom and Sharma, 2011). To be sure, there may have been something idiosyncratic to the focal course that was depressing women's self efficacy in our study, something that would not be a regular feature of other IE physics courses. An important example of course idiosyncrasies are differences between instructors, which can cause large differences in how students experience otherwise similar courses (Sanders and Horn, 1994). Given that the present study focused on a single IE physics course we cannot firmly rule out the possibility that idiosyncrasies in that course uniquely affected women's experiences. However, if idiosyncrasies uniquely affected women's self-efficacy states then the self-efficacy trait outcomes for the focal IE physics course should have been more severe than in other IE physics courses. In fact, the self-efficacy trait outcomes for the focal IE physics course featured in this study were similar to other IE physics courses, suggesting that this course was representative of IE physics instruction in general. Furthermore, the focal course had similar conceptual learning outcomes to other courses implementing IE pedagogies (Madsen et al., 2013; Kost et al., 2009a; Hake, 1998) and had similar gender differences to other IE physics courses on all four trait measures including: conceptual knowledge, self-efficacy traits, attitudes about learning physics and course grades (Kost et al., 2009a; Kost-Smith, 2011).

Although our favored explanation is that the IE physics instruction negatively impacted women's self-efficacy, two alternative explanations bear discussing. Both attribute the cause of the effects to the female students rather than to the learning environment. First, it may be that the gender differences in state experiences were not a result of gender per se, but rather a result of the trait factors that varied with gender, namely female students' lower conceptual knowledge, less expert-like attitudes and lower self-efficacy traits at the beginning of the course. In Chapter 6 I checked on this possibility using regression analysis to determine whether gender explained significant variance in mean self-efficacy states when controlling for trait effects. The results were that gender was the largest predictor of an individual's average self-efficacy state experiences while controlling for traits. Thus, there was something about being a woman in physics, over and above the measured physics traits, that made the experience of IE physics harmful to women's self-efficacy. A second alternative explanation, particularly for the large gender difference in self-efficacy states, can be attributed to the recruitment process. It is possible that recruiting students in their IE physics class biased the sample by attracting female volunteers who were particularly interested in having their negative experiences in physics be understood. The recruitment process did not indicate that the study was about students' emotions or feelings to limit this possibility. Nevertheless, students could easily have inferred that our study of "experiences" would include affect, so the possibility of sample bias cannot be ruled out. The results reported for self-efficacy traits, however, provide some evidence that, if there was sample bias, it was in a conservative direction. Namely, female ESM participants had smaller negative shifts pre to post course than female non-ESM participants and women's low self-efficacy states were evenly distributed across a sample of one in four women in the course who over-represented higher-achieving women. Thus, even if the sample were biased these results apply to a significant subpopulation of women taking physics.

What is it about IE physics instruction that is harmful to women's self-efficacy? STEM courses naturally have many features in common with physics in terms of physical environment, course structure, assignment of grades and so on. It is the nature of instruction and subject matter that varies most from course to course. One possibility is that there is something about the domain content in physics that somehow makes women feel less efficacious than does the content of other introductory courses such as mathematics, chemistry, and engineering principles. Along these lines Taasoobshirazi and Carr (2008) suggested that women may be disadvantaged in physics because of its emphasis on spatial thinking, which interacts with gender differences in spatial ability. The results of the present study are only partially consistent with this interpretation in that men did exhibit higher scores on conceptual knowledge than women. However, contrary to this interpretation, women ended the course similar to men with respect to grades. Thus, it is not obvious that men were leveraging their presumably higher average spatial ability very well. Meanwhile, as reported earlier in this article modeling instruction physics courses tentatively do not negatively impact students' self-efficacy traits or provide for differential shifts between men and women. Therefore, we think that pedagogy likely plays a larger role than subject matter in the observed effects within IE Physics.

5.8 Conclusion

Here we have used the ESM to situate the experiences of interest, self-efficacy in physics, within the breadth of experience while minimizing the effects of retrospection. This demonstrated how the ESM can be scaled to capture a large collection of experiences across a broad range of activities for a large number of participants. These features complement and bridge the fine-grained detail that can be achieved with case studies using interviews or video analysis and the large-scale data that surveys provide. The present study did not allow us to conjecture about which aspects of physics pedagogy substantially impacted student's experiences, however, different designs leveraging the strengths of the ESM in combination with other methods can achieve this goal. For instance a useful form of research to identify the possible causal relationships and mechanisms between specific aspects of instruction and student experience would be case studies detailing the experiences of ESM participants whose self-efficacy states fell in either of the extremes. A second approach would be to use a similar design to the present study, but with a larger sample of experience, to increase the resolution of the ESM. This would allow investigating self-efficacy states within specific aspects of instruction, such as answering clicker questions in lecture, and linking these experiences to students' shifts in physics self-efficacy traits. Both of these approaches would benefit from collecting data in multiple courses and across different pedagogies.

The poor experiences, poor outcomes and underrepresentation of women in physics warrants future research to inform addressing and resolving these issues. Many students leave STEM and physics because of their poor experiences and despite being fully capable of succeeding in the material (Seymour and Hewitt, 1997). By physics instruction undermining women's self-efficacy traits, physics instruction is also likely undermining women's performance, persistence and selection of physics as a major. Here we have shown that self-efficacy is important to understanding the underrepresentation of women in physics. Developing physics instruction that supports positive self-efficacy states is a starting point for instruction that supports all students in meeting both the affective and the cognitive demands of learning physics, especially the development of self-efficacy traits. Such instruction is necessary for physics to inclusively support diverse populations of students (Redish et al., 1998; Hazari et al., 2007). Otherwise, physics is likely to continue to lag behind other

STEM disciplines in diversity, threatening its survival as a major subject of study as it becomes an anachronism in an ever more diverse world.

Chapter 6
EQUITY OF GENDER DIFFERENCES IN SELF-EFFICACY

The large gender difference in self-efficacy states experienced in the focal IE physics course that I report on in Chapter 5 indicated that the physics-learning environment does not equitably impact male and female students. Interpreting this difference in experience as an inequity is supported by the resultant increase in the gender differences in self-efficacy traits from pre to post instruction that is in part caused by the difference in experience. The increase in the gender difference in self-efficacy traits is an example of inequity in how the physics-learning environment impacts men and women. It is possible, however, that the large difference in self-efficacy states experienced in the focal IE physics course was not explicitly the result of gender, but was instead the result of pre-existing gender differences. Determining the extent to which each of these two possibilities, the learning environment and the pre-existing differences, were the sources of the gender difference in experience is important because it can inform how best to approach resolving these differences.

Before investigating the issue of equity it is necessary to lay out the possible interpretations of what equity is (Rodriguez et al., 2012). A starting point is to clearly state that equity is not the increase in the size of differences favoring the advantaged group. For example, IE physics does not equitably impact male and female students because, as I have shown in Chapter 5 and has been reported by Kost et al. (2009a) and Kost-Smith (2011), gender differences tend to increase from pre to post instruction in IE physics courses for self-efficacy traits, conceptual knowledge and attitudes about learning physics. Rodriguez et al. (2012) discusses two models of equitable outcomes that are applicable to interpreting the gender difference in self-efficacy states. Equity of parity is achieved when pre-existing differences are eliminated. In the case of self-efficacy states equity of parity would occur if male and female students had similar experiences while learning physics in spite of pre-existing differences. Equity of fairness is achieved if the actual gender differences in experience were similar to the gender differences in experience that were predicted by the pre-existing trait differences. If predicted and actual experiences were similar then gender would not be a categorical factor in student experiences, and the focal IE physics course could be said to be fairly experienced by male and female students. Thus, these two types of equity can be located on a spectrum with inequity. Inequity and equity of parity are on opposite ends of the spectrum and equity of fairness lies somewhere between them. This spectrum follows from inequity increasing differences, fairness maintaining differences and parity eliminating differences. My goal in this chapter is to locate the learning environment of the focal IE physics course on this continuum of equity in terms of the self-efficacy states that were experienced in that physics-learning environment.

In pursuit of this goal I asked the following research question:

- To what extent is the gender difference in mean self-efficacy states predicted by students' pre-course traits?

If the gender differences in experience were mostly predicted by students' pre-course traits then this would locate the physics-learning environment as fair on the equity spectrum. A fair learning environment would indicate that both efforts to improve the learning environment and efforts to minimize the gender differences in traits prior to introductory college physics instruction were warranted to address these gender inequalities. In contrast, if the gender difference in experiences is a result of inequity this would indicate that there is a strong need to improve the physics-learning environment if the underrepresentation of women in physics is to be addressed.

6.1 Design of the analysis

The first step I took to investigate the equity in the focal IE physics course was to select the pre-course traits that would act as independent variables in predicting students' self-efficacy states in the focal IE physics course. Because not all ESM participants completed all of the pre-course trait measures, this step also involved selecting the population of ESM participants to include in the analysis. I sought to maximize the number of participants included in the analysis and to not skew the results of the analysis in a biased way. Therefore, I used correlations between the variables and several iterations of multiple linear regressions (MLR) to select the variables that balanced the information they provided with the largest population of participants.

All MLR models investigated mean self-efficacy states as the dependent variable, Equation 6.1, which is accepted practice in ESM studies (Hektner et al., 2007). Each model also included gender as the first independent variable entered as a dummy variable with 0 for males and 1 for females. As subsequent models included additional trait variables the shift in beta for the gender variable demonstrated how much of the gender differences were accounted for by those additional traits. If the beta for gender were to be near zero this would indicate that the course had equity of fairness. If the beta for gender were to become positive this would indicate that the gender differences were mitigated and that the course fell between fairness and parity in terms of equity. I limited the independent variables to pre-course trait data in order to investigate the extent to which the gender difference in experience was related to the traits students brought to the course rather than the traits that they developed in the course.

$$Mean\ SE\ state = \beta_0 + \beta_1 \times Gender + \sum_{k=2}^{N} \beta_k \times VAR_k \quad (6.1)$$

I calculated means and standard deviations for all state and trait data for male participants, female participants and in total. The means facilitated interpreting the betas that were generated by the MLR because it gave a size of the gender differences on these measures. The standard deviations facilitated interpreting the standardized betas because they are normalized in terms of the total standard deviation for that variable. I calculated the correlations between the variables in order to inform the strength of the relationships between the variables. I added independent variables to the MLR in order of highest correlation with mean self-efficacy states.

The data met all of the assumptions of MLR, which are detailed in Chapter 2.

6.2 Selecting variables for the MLR analysis

Twenty-eight of the thirty-three ESM participants provided data for all three of the pre-course trait measures: FMCE, CLASS and PSES. Of the five students who did not provide all pre-course trait data one male did not provide either CLASS or PSES scores, three males did not provide FMCE scores and one female did not provide PSES data.

After completing several iterations of MLR with all possible combinations of the CLASS, FMCE and PSES I decided to not include the PSES in the final analysis. The MLR analysis revealed that the PSES did not provide any additional information above and beyond that provided by CLASS and FMCE. I have included a fourth model with the PSES and limited to 28 participants in Table 6.3 to illustrate the limited additional information the PSES provided. By excluding the PSES from the MLR I was able to include 29 of the 33 ESM participants.

Correlations between students' mean self-efficacy states and the pre-course traits were strongest for the CLASS followed by the FMCE, Table 6.1, and this is the order I entered these independent variables into the MLR analysis.

Table 6.1: Correlations between mean self-efficacy states and pre-course traits. Correlations are for all 33 ESM participants with the number of participants included in each correlation indicated in parenthesis. * indicates p<0.05 and ** indicates p<0.01.

	1.	2.	3.	4.
1. SE State				
2. Gender	-0.47** (33)			
3. CLASS	0.47** (32)	-0.46** (32)		
4. FMCE	0.40* (30)	-0.33 (30)	0.20 (29)	
5. PSES	0.31 (31)	-0.26 (31)	0.57** (31)	0.08 (28)

Table 6.2: ESM participants' grades and mean self-efficacy for those included in the multiple linear regression by gender.

Measure	Male non-MLR	Male (MLR)	Female
N	4	16	13
	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)
Grade	2.48 (1.65)	3.19 (0.95)	3.05 (0.95)
Self-Efficacy	1.79 (0.27)	2.38 (0.72)	1.58 (0.61)

6.3 Participants and Attrition

The four male ESM participants that I did not include in the analysis because they had not provided the necessary pre-course traits had lower mean self-efficacy states and lower course grades than the male ESM participants that I did include in the analysis, Table 6.2. The attrition of these male students with lower self-efficacy states brings into question the representativeness of the relationships that the subsequent MLR revealed because it increased the gender difference in self-efficacy states. However, because the purpose of the MLR was to investigate how much the beta for gender changed, and not the size of that beta, it is likely that this shift will have little effect on the resultant analysis. In order to test the representativeness of the findings I replicated the first two MLR models with this larger group of 32 students to investigate the extent to which the shift in the beta for gender that resulted from including the CLASS in the model was similar between the two populations.

Table 6.3: Means and standard deviations for pre-course traits used in the MLR.

Measure	Male (16) Mean (S.D.)	Female (13) Mean (S.D.)	Total (29) Mean (S.D.)
CLASS (%)	72.3 (14.0)	55.8 (19.4)	64.9 (18.4)
FMCE (%)	33.1 (18.1)	22.7 (10.3)	28.4 (15.7)

The male students included in the MLR started the course with moderately higher levels of conceptual knowledge and much more expert-like attitudes about learning and doing physics than the female students, Table 6.2. The standard deviation for the CLASS was slightly higher than for the FMCE which will make the standardized beta differ from the non-standardized beta more for the FMCE than for the CLASS.

6.4 Results: The role of trait gender differences in self-efficacy state gender differences

I built three models, shown in Table 6.4, to demonstrate the relationships that were found between mean self-efficacy states, gender, pre-CLASS and pre-FMCE. Each model follows Equation 6.1 and included gender in order to determine how much of the gender differences in mean self-efficacy states were explained by pre-course traits. I included a fourth model that added PSES to the prior three models to illustrate that doing so provided no additional information to the prior models.

The upper portion of the MLR table describes the model overall. The variance explained, row 2, describes the overall explanatory power of that model. The adjusted variance, row 3, describes the overall explanatory power of the model taking into account the fact that adding additional variables can exaggerate the explanatory power of the model. I included both because the variance, row 2, is a better measure for comparing the additional information that is provided by the additional variables in each model. The p value, row 4, for the model describes the statistical reliability of the model overall. The lower half of the model provides information for each of the betas in the model. The beta for the constant is the intercept for that model with all other independent variables at zero. The beta for each independent variable describes how much a one-unit change in that variable will change the dependent variable. The standardized beta describes how much a one standard deviation change in that independent variable will change the dependent variable. Together, both of these betas are useful for interpreting the relative impact of each variable because they show how much and, to a degree, how easily each independent variable can effect the dependent variable. The p value for each beta describes the reliability of that beta within the model.

I have included the statistical significance for each beta to provide insight into the strength and reliability of the relationships. However, as Berk (2003) points out MLR is useful for investigating the relationships between variables even when statistical significance of the results cannot be assured. Therefore, I focused on the change in the overall variance explained by the model and the shift in the beta for gender that resulted from the inclusion of each variable to determine the additional information provided by adding that variable.

Model 1 only included gender. This model shows the size of the gender differences for self-efficacy states and the variance explained by this model provides a baseline to compare subsequent models to. Female students experienced 0.81 units lower mean self-efficacy states in the IE physics course than male students on a 0 to 4 scale and this model explained 27.8% of the variance in the mean self-efficacy states. In Model 2 the addition of pre-course CLASS scores added substantively to the model, and raised the variance explained by 7.5% to 35.3%. Pre-course CLASS traits reduced the role of gender to 73.5% of the original effect. In Model 3 I added pre-course FMCE scores to those predictors included in Model 2. The inclusion of pre-course FMCE scores reduced the gender difference to 59.4% of that in Model 1 and raised

Table 6.4: Multiple linear regression analysis results. Note that model 4 was used to illustrate the limited additional information the PSES provided and included one less participant than the other models.

Model	1			2			3		
Variance	0.278			0.353			0.404		
Adj. Var.	0.251			0.303			0.333		
p	0.003			0.004			0.004		
	β	β_{std}	p	β	β_{std}	p	β	β_{std}	p
Constant	2.389	-	0.001	1.447	-	0.001	1.099	-	0.001
Gender	-0.811	-0.527	0.003	-0.596	-0.387	0.038	0.482	-0.313	0.095
CLASS	-	-	-	0.013	0.307	0.095	0.012	0.292	0.105
FMCE	-	-	-	-	-	-	0.012	0.242	0.153
PSES	-	-	-	-	-	-	-	-	-

Model	4*	N=28	
Variance	0.403		
Adj. Var.	0.299		
p	0.015		
	β	β_{std}	p
Constant	1.814	-	0.079
Gender	-0.465	-0.297	0.135
CLASS	0.011	0.258	0.250
FMCE	0.012	0.238	0.177
PSES	0.104	0.070	0.727

the explained variance by 5.1% to 40.4%. These findings demonstrate that gender was the most important predictor of mean self-efficacy states experienced in the focal IE physics course. Taking into account the attitudes and conceptual knowledge that students had at the beginning of the course only accounted for 40.6% of the gender difference in how they experienced the course. This strongly suggests that the focal IE physics course was inequitable and biased in how it was experienced by women.

6.5 Results: Representativeness of the MLR analysis

In the following analyses I investigate the representativeness of the above MLR analysis. To do this I included all 32 students who provided ESM data and CLASS data by including the three male ESM participants that did not provide FMCE data. Model Rep. 1, Table 6.5, had a smaller gender difference than in Model 1, Table 6.4. Women experienced 0.72 units lower self-efficacy states in the focal IE

Table 6.5: MLR models to investigate the representativeness of the 29 participants included in the analysis. These models included the 32 students who provided ESM data and pre-CLASS scores.

Model	Rep. 1			Rep. 2		
Variance	0.232			0.312		
Adj. Var.	0.207			0.265		
p	0.005			0.004		
	β	β_{std}	p	β	β_{std}	p
Constant	3.300	-	0.001	2.337	-	0.001
Gender	-0.722	-0.482	0.003	-0.505	-0.337	0.061
CLASS	-	-	-	0.013	0.318	0.076

physics course than male students. This was 89% the size of the gender gap found in the earlier analysis, which was 0.81 units. In Model Rep. 2 the beta for gender was reduced to 70.0% of that in Model Rep. 1. This was similar in size but slightly larger than the reduction of 73.5% from Model 1 to Model 2 in the earlier analysis. If FMCE scores had been available it is likely that they would have had a similar or smaller effect on the findings in the earlier models since the FMCE was not as strongly related to mean self-efficacy states as the CLASS was. Therefore it is likely that the findings in the earlier analysis that pre-course traits explained only 40.6% of the gender difference in self-efficacy states was representative and reliably identified the focal IE physics course as having an inequitable and biased effect on how women experienced learning physics.

6.6 Discussion

Female students started the physics course with less expert-like attitudes about learning physics and lower conceptual knowledge than male students. These two pre-course traits predicted some of the difference in the self-efficacy states that men and women experienced in the focal IE physics course. In total, however, they explained less than half of the difference in experience. An alternative interpretation of the size of the gender difference in the mean self-efficacy states students experienced in the focal IE physics course was that they were more than double the difference predicted by students' pre-course traits. That the pre-course traits predicted less than half of the gender difference in self-efficacy states is a strong indication that the physics-learning environment was inequitable in that it was detrimentally biased against women. This locates the source of this bias, at least in part, within students common experiences in the physics-learning environment rather than in more marginal aspects of experience like receiving course grades.

Given that this was a relatively small study in terms of the number of participants, the number of background variables included and the focus on one IE physics course, I took steps to inform the representativeness of these findings. All of these steps indicated that these findings were representative of the students' experiences in the focal IE physics course and in IE physics courses in general. One concern remaining is that the small sample of students left the MLR analysis with limited statistical power, leaving open the question of the reliability of many of the relationships in the models; this can be seen by the p value for gender increasing from Model 1 to Model 3. However, it is unlikely that the result that the physics-learning environment was inequitable was so unstable as to grossly misrepresent the status of equity in IE physics instruction given that the gender difference in self-efficacy traits increased for a much larger sample of students both in the focal IE physics course and in other IE physics courses (Kost-Smith, 2011). A larger sample of students in IE physics courses could identify how stable and reliable the relationships between self-efficacy states, gender and students' pre-course traits are. Further investigations including physics courses using different teaching practices could also inform the role of the IE instruction in these relationships

Further limiting these findings was the pre-FMCE scores having had a relatively small range and limited variance for female students, which may have limited the variance in the mean self-efficacy states that the pre-FMCE scores could explain. This limited variance in pre-FMCE scores raises the possibility that other measures of prior knowledge could better explain variance in the mean self-efficacy states students experienced since self-efficacy and performance are consistently related in physics (Sawtelle et al., 2012a; Kost-Smith, 2011; Cavallo et al., 2004). Other measures of performance could include prior physics course grades, mathematical ability or college GPA. It is also possible that other traits that were not measured were particularly germane to women's self-efficacy state experiences such as women's endorsement of gender stereotypes in science, contentment with their gender, mathematical ability or spatial ability. I was unable to include these other measures in this study because of the extensive nature of the data collection. Notwithstanding these limitations, the relationships that I found indicated that the learning environment in the IE physics course strongly favored male students beyond what would be expected based on the advantages those male students started the course with.

I expected the pre-course physics self-efficacy traits to be significant predictors of the self-efficacy states students experienced in the focal IE physics course. Instead the pre-course self-efficacy traits were not strongly related to student's mean self-efficacy state experiences and did not add substantively to the models of self-efficacy state experiences. Nonetheless, there was a positive moderate relationship between the pre-course self-efficacy traits and the students' mean self-efficacy states consistent with the expected relationship. It is possible that this relationship was not as strong as the relationships with the other pre-course traits because students did not have stable physics self-efficacy traits prior to instruction in this physics course. This could have been the case because students likely had very few experiences in physics courses similar to the focal IE physics course. This possibility was supported by students' low conceptual knowledge at the beginning of the course, which was consistent with most students not having experienced physics instruction similar to the instruction in the focal IE physics course. The potential instability of students' pre-course self-efficacy traits aligns with mixed findings about the relationship between achievement and physics self-efficacy in introductory college physics courses. Kost-Smith (2011) found a relationship between students' course test grades and their self-efficacy beliefs. Lindstrom and Sharma (2011), in contrast, found a relationship between self-efficacy and performance only after students' self-efficacy traits had developed over several months of physics instruction. Both my findings and the mixed results in these prior studies suggest that a more robust and reliable measure of students' pre-instruction self-efficacy traits that are applicable to physics learning should be investigated in order to further explore the role of self-efficacy traits in the experiences students have while learning physics.

6.7 Conclusion

Female students tend to start college physics courses at a disadvantage in terms of their cognitive and affective traits. College physics instruction fails to mitigate or even just maintain these differences and instead the experience of instruction perpetuates and grows these differences. There is evidence that gender inequities common in college physics courses may be mitigated by IE instruction (Kost et al., 2009b). However, the large gender difference in the self-efficacy states experienced in the focal IE physics course and the similarity of the focal course to other IE physics courses indicate that physics instruction has a long way to go before an equity of fairness, in which gender differences do not grow, is achieved. One option for informing how to improve equity in physics is to investigate research-based instruction that demonstrates at least equity of fairness. These courses could be used to identify practices that support equity. If equity in college physics instruction is not addressed physics will continue to dramatically underrepresent women and will continue to stand out as one of the least diverse science disciplines (National Science Foundation, 2012).

In order for physics to recruit and retain a larger and more diverse population of students, physics instruction must support students with diverse backgrounds, skills and abilities. Students should be supported in attaining the content knowledge and skills they need to succeed and in developing the expert attitudes and self-efficacy traits that will sustain their perseverance in physics learning. Efforts to support student development of these important characteristics should focus on the experiences students have while they learn physics as these experiences are the source of many of these characteristics.

Chapter 7
SUMMARY

Findings from across all of the chapters indicated that female students experienced much lower self-efficacy states than men in physics and only in physics, and these gender difference were more than twice as large as those predicted by students' pre-course traits. Subsequently, women had much more negative shifts in their self-efficacy traits than men. Taken altogether these findings indicated that the physics-learning environment in the focal IE physics course was a significant cause of the negative impact on female students' self-efficacy. The shift in self-efficacy traits located this effect as a result of physics instruction, and the low self-efficacy states experienced in physics located the effect directly in the physics-learning environment. Furthermore, the much lower self-efficacy states that women experienced in physics than in other STEM courses indicated that this negative effect was neither an epiphenomenon of a larger trend nor a consequence of differences in retrospection.

That the gender difference in self-efficacy states was much larger than that predicted by students' pre-course traits indicated that the negative impacts on women's self-efficacy were consequences of inequities in the physics-learning environment. The difference between men's and women's self-efficacy states were more than twice those predicted by pre-course traits. The size of this difference indicated that not only did the physics instruction in the focal IE physics course perpetuate pre-existing gender differences, but the physics instruction increased those differences. Furthermore, the gender differences for conceptual knowledge, attitudes and self-efficacy traits all increased from pre to post instruction. Taken altogether these increases in gender differences for traits and states indicated that the physics-learning environment in the focal IE physics course perpetuated and amplified gender differences, and that a source of this inequitable effect was students' experiences in the physics-learning environment.

Investigating the effect of physics instruction on students' self-efficacy directly in the physics-learning environment required developing a measure of self-efficacy states. A review of the prior research during the development of my own research revealed no previous study that used the ESM to measure self-efficacy states. Therefore, it was necessary for me to develop and validate the self-efficacy state measure. I developed the self-efficacy state construct by using Bandura's (1997) description of self-efficacy to identify three Likert-scale questions used in prior ESM studies that were consistent with self-efficacy: skill, control and success. Validity was supported by these three principle questions uniquely and strongly loading on the same construct. Further evidence of the validity of the self-efficacy state construct was provided by the relationships between self-efficacy states and the complementary states matching those that I expected. Additional evidence was also provided by the consistency between the gender differences for both the self-efficacy state and trait measures. Women had more negative shifts in their physics self-efficacy traits than men and women experienced much lower self-efficacy states in physics than men. Taken together these results indicated that the ESM provided a reliable and valid measure of students' self-efficacy states.

In developing and validating a measure of self-efficacy states I sought to address a discrepancy in how self-efficacy has been historically described, as a state-like dynamic property, and how it has been measured, often as a static trait-like property. Measuring self-efficacy states within the activity at hand in peoples' daily lives provides a method for investigating how self-efficacy develops and the role of self-efficacy in daily life. In this research I have focused on physics learning, but self-efficacy pervades daily life and the self-efficacy state measure can substantially contribute to understanding the role of self-efficacy in human agency.

For conceptual knowledge, attitudes and self-efficacy traits the shifts in students' scores and the shifts in gender differences in the focal IE physics course were similar to those in the IE physics courses studied by Kost et al. (2009b) and Kost-Smith (2011). The focal IE physics course used the same research-based materials and strategies as those used in other IE physics courses. These similarities indicated that the negative impact on female students' self-efficacy in the focal IE physics courses was likely representative of IE physics in general. Furthermore, research-based teaching practices, including IE physics, tend to better support students' cognitive and affective outcomes (Kost et al., 2009a; Brewe et al., 2009; Sawtelle et al., 2010; Madsen et al., 2013; Hake, 1998) and are less biased against women (Madsen et al., 2013; Sawtelle et al., 2010; Brahmia, 2008) than traditional lecture-based instruction. Therefore, the biased and negative impact on women that I found in the focal IE physics course likely represents a best-case scenario for the effects of traditional lecture-based introductory college physics courses on women's self-efficacy.

7.1 Using stereotype threat to inform the scope of the physics-learning environment

One possible source of the large gender difference in self-efficacy states in the focal IE physics course that bears discussing is stereotype threat. Stereotype threat is an extra-pressure that individuals in negatively stereotyped groups experience in activities where they could confirm that stereotype. For example, women or African American students in mathematics courses. This pressure takes two forms: negative emotions while engaged in the activity that consume cognitive resources that would otherwise be available and an incentive to disengage from the stereotyped domain (Aronson et al., 1999). Through these pressures, stereotype threat undermines performance (Aronson et al., 1999) and learning (Rydell et al., 2010; Mangels et al., 2012). In physics, stereotype threat has been used to investigate and mitigate the underperformance of women in introductory physics courses (Miyake et al., 2010) and the underrepresentation of women in physics (Deemer et al., 2014). Stereotype threat is salient to the gender differences in self-efficacy in physics that I found in this study because, as I discussed in Chapter 2 and confirmed in Chapter 3, stress and self-efficacy states are related and stress is one of the negative emotions that blocks learning under stereotype threat (Mangels et al., 2012). Furthermore, self-efficacy traits may mitigate the effects of stereotype threat (Deemer et al., 2014). The measure of self-efficacy states that I developed and validated in this research may be a useful tool for investigating the mechanisms by which stereotype threat impacts students in the learning environment. A necessary first step in this research would be to directly tie self-efficacy states to stereotype threat conditions through an experimental design study investigating student experiences in simulated classroom activities, such as completing Tutorials in Introductory Physics or answering clicker questions in a Peer Instruction format, under control and stereotype threat conditions.

The research on stereotype threat is also useful for illustrating the aspects of the physics-learning environment that may have contributed to the poor self-efficacy states women experienced. These include student characteristics and environmental characteristics, with environmental characteristics including the immediate social and physical environment and the more distant cultural environment in which it is nested. Student characteristics including social identity and self-efficacy can influence their susceptibility to stereotype threat (Schmader, 2002; Deemer et al., 2014). Immediate environmental characteristics include: working solo and how tests are administered. When student groups consisted of only one female participant that female tended to do worse than women in groups with multiple female participants (Sekaquaptewa and Thompson, 2003; Inzlicht and Ben-Zeev, 2000). In upper-division mathematics courses tests that were presented as assessments of ability undermined women's performance such that men and women performed similarly (Good et al., 2008). When students were told that these tests were free of gender bias women outperformed men. In terms of the broader culture, cultures with greater gender egalitarianism will have fewer gender stereotypes and less negative impacts on women's performance (Lippa et al., 2010). One factor that manifests itself in the students, the immediate environment and the larger culture are beliefs about the fixedness of intelligence, as opposed to intelligence being malleable (Dweck and Leggett, 1988). Students who perceive intelligence to be more fixed are more susceptible to stereotype threat (Aronson et al., 2002). In immediate environments in which interventions focus on intelligence being malleable students are less susceptible to stereotype threat (Aronson et al., 2002). In terms of the larger culture, the extent to which academic domains ascribe to a fixed intelligence theory predicts the number of women in those domains (Leslie et al., 2015). The causes of women's poor self-efficacy states in physics likely dwell in each of these aspects, or across many of them. Thus, future research on students' experiences in physics will need to investigate a broad range of variables, just as stereotype threat research has.

7.2 Implications for instruction

The negative effect of physics instruction on women's self-efficacy that I found in the focal IE physics course was consistent with physics being one of the smallest and least diverse STEM majors (National Science Foundation, 2012). Since self-efficacy directly effects students' interest in pursuing STEM majors (Luzzo et al., 1999) and is related to students' choice of STEM majors (Betz and Hackett, 1983; Marra and Bogue, 2009), harming women's self-efficacy in physics likely decreases the likelihood of women selecting a physics major. Furthermore, self-efficacy directly effects cognitive performance (Bouffard-Bouchard, 1990), partially explains gender differences in physics conceptual knowledge (Kost-Smith, 2011) and predicts students' performance in introductory physics courses (Kost-Smith, 2011; Sawtelle et al., 2012a). Subsequently, it seems likely that harming women's self-efficacy will harm their ability to gain the knowledge that will support them in their future learning, and may also decrease women's opportunities to pursue physics degrees by potentially decreasing the number of women that pass introductory physics courses. Therefore, it stands to reason that self-efficacy is a significant factor in the underrepresentation of women in physics.

Given the likely importance of self-efficacy in the underrepresentation of women in physics some instructors may want to assess the impact of their courses on students' self-efficacy. Instructors could use a self-efficacy trait instrument to measure shifts in students' self-efficacy traits across instruction. This could be a complete instrument that would provide information on many aspects of self-efficacy such as the PSES or a smaller collection of questions on a broader instrument such as the problem solving confidence questions on the CLASS. Given that I found the negative impact on women to occur in the physics-learning environment, physics instructors may alternatively want to use a formative measure of students' self-efficacy states. This formative measure could be accomplished by asking students how confident, successful, skillful or in control they feel in their ability to learn the topic at hand. These questions could be asked in class as a clicker question or with some other polling mechanism, or they could be asked on assignments.

The largest difficulty in addressing inequities in self-efficacy is that so little is known about the characteristics of the situation that lead to very low or very high self-efficacy states. Though, it seems obvious that if students perceive the activities to be far too challenging then they will likely experience very low self-efficacy states. Therefore, it could also be helpful to monitor the difficulty of the learning tasks, and a formative self-efficacy measure could be used for this purpose. Similarly, there are many students who are interested in pursuing physics majors, but who have not had the opportunity to develop the skills and knowledge necessary to succeed in calculus-based introductory physics courses (Whitten et al., 2003). When these students enter a physics course they are being confronted with a much greater challenge than their peers. Therefore it is likely that the self-efficacy states they experience are much lower than those of their better prepared peers. Given that many women start college physics courses with lower conceptual knowledge and less expert attitudes, departments wishing to increase the diversity of their physics majors may want to develop alternative pathways into physics majors. These alternative pathways can balance the difficulty of the material with the skills students have to create opportunities for students to experience high self-efficacy states in the process of learning physics. These high self-efficacy states would likely support students in learning the physics content and in developing an interest for pursuing a physics degree. While I am inferring that self-efficacy is an important aspect for the effectiveness of these alternative pathways; there is evidence of that effectiveness. Alternative pathways are a common feature of physics departments with diverse populations of students (Whitten et al., 2003) and there is at least one documented case of alternative pathways dramatically improving women's outcomes in college physics (Brahmia, 2008).

7.3 General Implications for research

Looking broadly at the research agenda in physics learning there are several general areas where investigations using self-efficacy could be informative. The relationship between self-efficacy and gender inequities in this study is an indication that it may be informative to investigate gender differences in physics settings other than introductory college physics courses using self-efficacy. For example, investigating self-efficacy states of female students in secondary education could inform why high schools with a strong STEM curriculum and integrated extracurricular activities reduce the gender gap in STEM interest by 25% (Legewie and Diprete, 2014), since interest in STEM degrees is affected by self-efficacy (Luzzo et al., 1999). Exploring self-efficacy in upper-division undergraduate physics education and graduate physics education may also be informative because the differences between departments that support women in succeeding and those that don't seem to largely result from departmental culture (Whitten et al., 2003) and many women leave STEM majors after completing their first year in spite of being fully capable of success (Seymour and Hewitt, 1997). The quasi-longitudinal data that these studies would provide could situate the experience of women in introductory college physics courses within the broader experiences of women in physics education. This would move away from comparing men to women, which can position women as deficient, and move towards focusing on maximizing opportunities for individual excellence (Gutierrez, 2008).

In this study both self-efficacy states and traits consistently measured gender differences favoring male students. However, the effect size of the state difference was much larger than the effect size of the trait difference indicating that the state measure was a much more sensitive measure of gender differences. This contrasts with findings for state and trait anxiety in mathematics. Common gender differences in anxiety traits in mathematics (Hyde et al., 1990) were expected to represent gender differences in anxiety states in mathematics. However, Goetz et al. (2013) found that there likely were no gender differences in anxiety states experienced while learning mathematics and that the trait differences arose due to difference in retrospection. Further evidence for the complex relationship between affective states and traits has also been found for control and value. Bieg et al. (2013)

found that trait measures for control and value were not consistently better nor consistently worse predictors of students' affective state experiences than the state measures for control and value. Since the relationships between states and traits have received little attention and no discernible pattern has arisen within the studies I have mentioned, future studies are warranted. Furthermore, claims from work using only one of these methods should be conservative if they extend to the domain of the other measure.

7.4 Directions for extending the present work

I found that women frequently experienced very low self-efficacy states while learning physics. Unfortunately making specific recommendations on how to address these apparently detrimental experiences is difficult because little is known about the details of these experiences or the characteristics of the learning environment that are most influential in their occurrence. One approach to identify these characteristics and build a causal model for their influences would be to use case studies combining the ESM with complementary measures. For example, video data of students' experiences could be collected concurrently with the ESM data. The ESM could be used to both identify students who tended to have either very high or very low self-efficacy states in physics and to identify specific experiences that students reported either very high or very low self-efficacy states in. Analysis similar to Sawtelle et al. (2012a) could be used to identify students' opportunities for developing self-efficacy within these video clips. Furthermore, Sawtelle and colleagues' work could also be expanded to identify opportunities for harming self-efficacy. These opportunities could provide fine-grained detail of environmental characteristics that occur for different self-efficacy states. Video data would also inform what it means to have these experiences in terms of students' behavior during and following the experience. Interviews with students could expand upon and validate findings from the ESM and video analysis. In combination with data from trait surveys this research design could be used to develop a model for the development or destruction of self-efficacy traits in physics that would likely be applicable to other learning environments.

A second approach for investigating the details of students' self-efficacy state experiences is to increase the resolution of the ESM by collecting a much larger set of students' experiences. Increased resolution would allow investigating self-efficacy states within finer grained aspects of instruction. For example, enough ESM samples could be used to identify if there were differences in self-efficacy states between specific components of either Peer Instruction or Tutorials in Physics. Either of these approaches would benefit from collecting data in multiple forms of physics instruction since comparisons between the different instructional environments would support building a causal model of the relationships between environment, self-efficacy states and self-efficacy traits.

7.5 Conclusion

Interactive engagement is an extremely successful form of physics instruction (Hake, 1998; Madsen et al., 2013). However, while student learning has been dramatically increased by these research-based teaching practices, less advancement has been made in terms of students' affective outcomes (Kost et al., 2009b; Madsen et al., 2015; Kost-Smith, 2011). The poor self-efficacy states students experienced in physics and the subsequent negative impact on their self-efficacy traits that I found in this study locate the negative effect on self-efficacy within the physics-learning environment. Further understanding the aspects of the physics-learning environment that are causes of this detrimental effect and mitigating those causes are necessary steps in developing instruction that supports students in attaining self-efficacy in physics. Instruction that supports the development of self-efficacy would further support student learning of physics concepts and would support students to choose and succeed in physics majors.

REFERENCES

Adams, W., Perkins, K., Podolefsky, N., Dubson, M., Finkelstein, N., and Wieman, C. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, 2(1).

Adams, W. K., Wieman, C. E., Perkins, K. K., and Barbera, J. (2008). Modifying and Validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry. *Journal of Chemical Education*, 85(10):1435.

Aronson, J., Fried, C. B., and Good, C. (2002). Reducing the Effects of Stereotype Threat on African American College Students by Shaping Theories of Intelligence. *Journal of Experimental Social Psychology*, 38:113–125.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., and Brown, J. (1999). When White Men Can't Do Math: Necessary and Sufficient Factors in Stereotype Threat. *Journal of Experimental Social Psychology*, 35:29–46.

Asher, H. (1983). *Causal Modeling.* SAGE Publications.

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

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

Bandura, A. (2006). Guide for constructing self-efficacy scales. In Pajares, F. and Urdan, T., editors, *Self-Efficacy Beliefs of Adolescents*, pages 307–337. Information Age Publishing.

Berk, R. (2003). *Regression Analysis: A Constructive Critique: A Constructive Critique.* Advanced Quantitative Techniques in the Social Sciences. SAGE Publications.

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.

Betz, N. E. and Schifano, R. S. (2000). Evaluation of an Intervention to Increase Realistic Self-Efficacy and Interests in College Women. *Journal of Vocational Behavior*, 56(1):35–52.

Bieg, M., Goetz, T., and Hubbard, K. (2013). Can I master it and does it matter? An intraindividual analysis on control-value antecedents of trait and state academic emotions. *Learning and Individual Differences*, 28:102–108.

Bieg, M., Goetz, T., and Lipnevich, A. (2014). What students think they feel differs from what they really feel: Academic self-concept moderates the discrepancy between students' trait and state emotional self-reports. *PloS one*, 9(3):e92563.

Bledow, R. (2013). Demand-Perception and Self-Motivation as Opponent Processes: A Response to Bandura and Vancouver. *Journal of Management*, 39(1):14–26.

Bouffard-Bouchard, T. (1990). Influence of self-efficacy on performance in a cognitive task. *Journal of Social Psychology*, 130:353–363.

Brahmia, S. (2008). Improving Learning for Underrepresented Groups in Physics for Engineering Majors. *AIP Conference Proceedings*, pages 7–10.

Brewe, E., Kramer, L., and O'Brien, G. (2009). Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS. *Physical Review Special Topics - Physics Education Research*, 5(1):1–5.

Brewer, D. S. (2009). *The Effects of Online Homework on Achievement and Self-efficacy of College Algebra Students.* PhD thesis, Utah State University.

Cavallo, A., Rozman, M., and Potter, W. (2004). Gender Differences in Learning Constructs, Shifts in Learning Constructs, and Their Relationship to Course Achievement in a Structured Inquiry, Yearlong College Physics Course for Life Science Majors. *School Science and Mathematics*, 104(6):288–300.

Chemers, M. M., Hu, L., and Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology*, 93(1):55–64.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* Academic Press, New York.

Crouch, C. H. and Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, 69(9):970.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety.* Jossey-Bass Publishers, San Francisco.

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.

De Muth, J. E. (2014). *Basic Statistics and Pharmaceutical Statistical Applications.* CRC Press.

Deci, E., Koestner, R., and Ryan, R. (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, 125(6):627–668.

Deemer, E. D., Thoman, D. B., Chase, J. P., and Smith, J. L. (2014). Feeling the Threat: Stereotype Threat as a Contextual Barrier to Women's Science Career Choice Intentions. *Journal of Career Development*, 41(2):141–158.

Duit, R. and Treagust, D. (2003). Conceptual Change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, 25:671–688.

Durrheim, K. and Tredoux, C. (2014). *Numbers, Hypotheses Conclusions: A Course in Statistics for the Social Sciences.* Juta and Company Ltd.

Dweck, C. S. and Leggett, E. L. (1988). A Social-Cognitive Approach to Motivation and Personality. *Psychological Review*, 95(2):256–273.

Etkina, E. and Van Heuvelen, A. (2007). Investigative Science Learning Environment - A Science Process Approach to Learning Physics. In Redish, Edward F. an Cooney, P., editor, *Research-Based Reform of University Physics.*

Fencl, H. S. and Scheel, K. (2005). Engaging students: An examination of the effects of teaching strategies on self-efficacy and course climate in a nonmajors physics course. *Journal of College Science Teaching*, 35(1):20–24.

Ferrell, B. and Barbera, J. (2015). Analysis of students' self-effiacy, interest, and effort beliefs in general chemistry. *Chemistry Education Research and Practice*, 16:318–337.

Girden, E. R. (2014). *Evaluating research articles from start to finish.* Sage Publications, Thousand Oaks, CA.

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., and Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science*, 24(10):2079–87.

Goetz, T., Frenzel, A. C., Stoeger, H., and Hall, N. C. (2009). Antecedents of everyday positive emotions: An experience sampling analysis. *Motivation and Emotion*, 34(1):49–62.

Good, C., Aronson, J., and Harder, J. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses? *Journal of Applied Developmental Psychology*, 29:17–28.

Gutierrez, R. (2008). A "gap gazing" fetish in mathematics education? Problematizing research on the achievement gap. *Journal of Research in Mathematics Education*, 39.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1):64–74.

Hall, R. M. and Sandler, B. R. (1982). *The classroom climate: A chilly one for women.* PSEW, AAC, Washington DC.

Hammer, D. (1996). More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research. *American Journal of Physics*, 64(10):1316–1325.

Hazari, Z., Tai, R. H., and Sadler, P. M. (2007). Introductory University Physics Performance: The Influence of High School Physics Preparation. *Science Education*, 91:847–876.

Hektner, J. M., Schmidt, J. A., and Csikszentmihalyi, M. (2007). *Experience Sampling Method: Measuring the Quality of Everyday Life.* SAGE Publications, Thousand Oaks, CA.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., and Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: A meta-analysis. *Psychology of Women Quarterly*, 14:299–324.

Inzlicht, M. and Ben-Zeev, T. (2000). A Threatening Environment: Why Females Are Susceptible to Experiencing Problem-Solving Deficits in the Presence of Males. *Psychological Science*, 11(5):365–371.

Jackson, J., Hill, P., and Roberts, B. (2012). Misconceptions of Traits Continue to Persist: A Response to Bandura. *Journal of Management*, 38(3):745–752.

Jackson, J. E. (2003). *A Users Guide to Principle Components.* Wiley, Hoboken, NJ.

Johnson, L. S. (2009). School Contexts and Student Belonging: A Mixed Methods Study of an Innovative High. *The School Community Journal*, 19(1):99–118.

Koller, O., Baumert, J., and Schnabel, K. (2001). Does Interest Matter? The Relationship Between Academic Interest and Achievement in Mathematics. *Journal of Research in Mathematics Education*, 32(5):448–470.

Kost, L., Pollock, S., and Finkelstein, N. (2009a). Characterizing the gender gap in introductory physics. *Physical Review Special Topics - Physics Education Research*, 5(1):010101.

Kost, L., Pollock, S., and Finkelstein, N. (2009b). Unpacking gender differences in students' perceived experiences in introductory physics. In *PERC Proceedings*.

Kost-Smith, L., Pollock, S., and Finkelstein, N. (2010). Gender disparities in second-semester college physics: The incremental effects of a "smog of bias". *Physical Review Special Topics - Physics Education Research*, 6(2):1–17.

Kost-Smith, L. E. (2011). *Characterizing, Modeling, and Addressing Gender Disparities in Introductory College Physics.* PhD thesis, University of Colorado.

Lawson, A. E., Banks, D. L., and Logvin, M. (2007). Self-Efficacy, Reasoning Ability, and Achievement in College Biology. *Journal of Research in Science Teaching*, 44(5):706–724.

Lazarus, R. and Folkman, S. (1984). *Stress, appraisal and coping.* Springer, New York.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences*, 19(3):355–365.

Legewie, J. and Diprete, T. A. (2014). The High School Environment and the Gender Gap in Science and Engineering. *Sociology of Education*, 87(4):259–280.

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

Lindstrom, C. and Sharma, M. (2011). Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter? *International Journal of Innovation in Science and Mathematics Education*, 19(2):1–19.

Lippa, R., Collaer, M., and Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Archives of Sexual Behavior*, 39:990–997.

Lupien, S. J., Maheu, F., Tu, M., Fiocco, a., and Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and cognition*, 65(3):209–37.

Luzzo, D. A., Hasper, P., Albert, K., Bibby, M., and Martinelli, E. (1999). Effects of self-efficacy-enhancing interventions on the math/science self-efficacy and career interests, goals, and actions of career undecided college students. *Journal of Counseling Psychology*, 46(2):233–243.

Madsen, A., McKagan, S. B., and Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics - Physics Education Research*, 9(2):020121.

Madsen, A., Mckagan, S. B., and Sayre, E. C. (2015). How physics instruction impacts students' beliefs about learning physics: A meta-analysis of 24 studies. *Physical Review Special Topics - Physics Education Research*, 11(1):1–19.

Mangels, J. A., Good, C., Whiteman, R. C., Maniscalco, B., and Dweck, C. S. (2012). Emotion blocks the path to learning under stereotype threat. *Social Cognitive and Affective Neuroscience*, 7(2):230–241.

Marra, R. M. and Bogue, B. (2009). Women Engineering Students and Self-Efficacy: A Multi-Year, Multi-Institution Study of Women Engineering Student Self-Efficacy. *Journal of Engineering Education*, (January).

Mazur, E. (1997). *Peer Instruction: A users manual.* Prentice Hall, Upper Saddle River NJ.

McDermott, L. and Shaffer, P. (2002). *Tutorials in Introductory Physics.* Prentice Hall, Upper Saddle River NJ.

Miyake, A., Kost-smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., and Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330:1234–1237.

Mujtaba, T. and Reiss, M. J. (2013). Inequality in Experiences of Physics Education: Secondary School Girls' and Boys' Perceptions of their Physics Education and Intentions to Continue with Physics After the Age of 16. *International Journal of Science Education*, 35(11):1824–1845.

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):30–38.

National Commission on Mathematics and Science Teaching (2000). Before It's Too Late.

National Science Foundation (2012). Women, minorities, and persons with disabilities in science and engineering.

Nissen, J. M. and Shemwell, J. T. (2014). Measuring affective experience in the midst of STEM learning. *Proceedings of the Eleventh International Conference for the Learning Sciences*, 1:142–149.

Osborne, J., Simon, S., and Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25:1049–1079.

Osborne, J. and Dillon, J. (2008). Science Education in Europe: Critical Reflections. A report to the Nuffield Foundation.

Otero, V., Pollock, S., and Finkelstein, N. (2010). A physics department's role in preparing physics teachers: The Colorado learning assistant model. *American Journal of Physics*, 78(11):1218.

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review*, 18(4):315–341.

Pekrun, R. and Stephens, E. J. (2010). Achievement emotions in higher education. In Smart, J. C., editor, *Higher Education: Handbook of Theory and Research*, pages 257–306. Springer Netherlands, Dordrecht.

Perkins, K. K., Adams, W. K., Pollock, S. J., Finkelstein, N. D., and Wieman, C. E. (2004). Correlating Student Beliefs With Student Learning Using The Colorado Learning Attitudes about Science Survey. *PERC Proceedings*, 790:61–64.

Pietsch, J., Walker, R., and Chapman, E. (2003). The relationship among self-concept, self-efficacy, and performance in mathematics during secondary school. *Journal of Educational Psychology*, 95(3):589–603.

Pintrich, P. R., Marx, R. W., and Boyle, R. A. (1993). Beyond Cold Conceptual Change: The Role of Motivational Beliefs and Classroom Contextual Factors in the Process of Conceptual Change. *Review of Educational Research*, 63(2):167–199.

Rathunde, K. (2003). A Comparison of Montesorri and Traditional Middle Schools. *The NAMTA Journal*, 28(3):12–52.

Redish, E., Saul, J., and Steinberg, R. (1998). Student expectations in introductory physics. *American Journal of Physics*, 66(3):121–224.

Rodriguez, I., Brewe, E., Sawtelle, V., and Kramer, L. H. (2012). Impact of equity models and statistical measures on interpretations of educational reform. *Physical Review Special Topics - Physics Education Research*, 8(2):020103.

Roster, N. (2006). *The effects of inquiry-based teaching on attitudes, self-efficacy, and science reasoning abilities of students in introductory biology courses at a rural, open-enrollment community college.* PhD thesis, Oklahoma State University.

Ryan, R. and Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, 55(1):68–78.

Rydell, R. J., Shiffrin, R. M., Boucher, K. L., Loo, K. V., and Rydell, M. T. (2010). Stereotype threat prevents perceptual learning. *PNAS*, 107(32):14042–14047.

Sanders, W. L. and Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8(3):299–311.

Sawtelle, V., Brewe, E., Goertzen, R. M., and Kramer, L. H. (2012a). Identifying events that impact self-efficacy in physics learning. *Physical Review Special Topics - Physics Education Research*, 8(2).

Sawtelle, V., Brewe, E., and Kramer, L. H. (2010). Positive Impacts of Modeling Instruction on Self-Efficacy. In *PERC Proceedings*, pages 289–292.

Sawtelle, V., Brewe, E., and Kramer, L. H. (2012b). Exploring the Relationship Between Self-Efficacy and Retention in Introductory Physics. *Journal of Research in Science Teaching*, 49(9):1096–1121.

Schmader, T. (2002). Gender Identification Moderates Stereotype Threat Effects on Women's Math Performance. *Journal of Experimental Social Psychology*, 201:194–201.

Schmidt, J. A. and Shumow, L. (2012). Change in Self-Efficacy in High School Science Classrooms. In Britner, S., editor, *Self-Efficacy in School and Community Settings*, pages 53–73. Nova Science Publishers.

Sekaquaptewa, D. and Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*, 39:68–74.

Semsar, K., Knight, J. K., Birol, G., and Smith, M. K. (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in Biology. *CBE Life Sciences Education*, 10(3):268–78.

Seymour, E. and Hewitt, N. (1997). *Talking about leaving: Why undergraduates leave the sciences.* Westview, Boulder CO.

Shernoff, D., Csikszentmihaly, M., Schneider, B., Shernoff, E. (2003). Student Engagement in the High School Classroom from the Perspective of Flow Theory. *School Psychology Quarterly*, 18(2):158–176.

Sinatra, G. M. (2005). The "Warming Trend" in Conceptual Change Research: The Legacy of Paul R. Pintrich. *Educational Psychologist*, 40(2):107–115.

Singh, K., Granville, M., and Dika, S. (2002). Mathematics and Science Achievement: Effects of Motivation, Interest, and Academic Engagement. *The Journal of Educational Research*, 95(6):323–332.

Stewart, J., Griffin, H., and Stewart, G. (2007). Context sensitivity in the force concept inventory. *Physical Review Special Topics - Physics Education Research*, 3:1–11.

Taasoobshirazi, G. and Carr, M. (2008). Gender Differences in Science: An Expertise Perspective. *Educational Psychology Review*, 20(2):149–169.

Tabachnick, B. and Fidell, L. (2001). *Using Multivariate Statistics.* A&B Interactive. Allyn and Bacon.

Thayer, R. E. (1967). Measurement of activation through self-report. *Psychological Reports*, 20:659–662.

Thayer, R. E. (1996). *The origin of everyday moods: Managing energy, tension and stress.* Oxford University Press, New Yor, NY.

Thornton, R., Kuhl, D., Cummings, K., and Marx, J. (2009). Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory. *Physical Review Special Topics - Physics Education Research*, 5(1):1–8.

Thornton, R. K. and Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula. *American Journal of Physics*, 66(4):338–352.

Turpen, C. and Finkelstein, N. (2009). Not all interactive engagement is the same: Variations in physics professors' implementation of Peer Instruction. *Physical Review Special Topics - Physics Education Research*, 5(2):020101.

Villafane, S., 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*, 15:114–127.

Vosniadou, S. and Skopeliti, I. (2014). Conceptual Change from the Framework Theory Side of the Fence. *Science Education*, pages 1427–1445.

Whitten, B. L., Foster, S. R., and Duncombe, M. L. (2003). What works for women in undergraduate physics. *Physics Today*, 156(9).

Williams, T. and Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology*, 102(2):453–466.

Wittmann, M. (2009). FMCE Grading Template.

Zajacova, A., Lynch, S. M., and Espenshade, T. J. (2005). Self-Efficacy, Stress, and Academic Success in College. *Research in Higher Education*, 46(6):677–706.

Zeldin, A. and Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, 37(1):215–246.

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.

BIOGRAPHY OF THE AUTHOR

Jayson Nissen was born in Havre, Montana in 1983. He was raised there and graduated from Havre High School in 2001. He attended Montana State University and graduated in 2010 with a Bachelor's degree in Physics. He then moved to Maine and entered the Physics graduate program at the University of Maine in the Fall of 2010. Jayson Micheal Nissen is a candidate for the Doctor of Philosophy degree in Physics from the University of Maine in May 2016.