6-30-2017

Describing and Mapping the Interactions between Student Affective Factors Related to Persistence in Science, Physics, and Engineering

Jacqueline Doyle
Florida International University, doylejackd@gmail.com

DOI: 10.25148/etd.FIDC001978

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

DESCRIBING AND MAPPING THE INTERACTIONS BETWEEN STUDENT AFFECTIVE FACTORS RELATED TO PERSISTENCE IN SCIENCE, PHYSICS, AND ENGINEERING

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

Jacqueline Doyle

2017

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This dissertation, written by Jacqueline Doyle, and entitled Describing and Mapping the Interactions Between Student Affective Factors Related to Persistence in Science, Physics, and Engineering, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Zahra Hazari

_____
Joerg Reinhold

_____
Laird Kramer

_____
Geoff Potvin, Major Professor

Date of Defense: June 30, 2017

The dissertation of Jacqueline Doyle is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2017

ABSTRACT OF THE DISSERTATION

DESCRIBING AND MAPPING THE INTERACTIONS BETWEEN STUDENT AFFECTIVE FACTORS RELATED TO PERSISTENCE IN SCIENCE, PHYSICS, AND ENGINEERING

by

Jacqueline Doyle

Florida International University, 2017

Miami, Florida

Professor Geoff Potvin, Major Professor

This dissertation explores how students' beliefs and attitudes interact with their identities as physics people, motivated by calls to increase participation in science, technology, engineering, and mathematics (STEM) careers. This work combines several theoretical frameworks, including Identity theory, Future Time Perspective theory, and other personality traits to investigate associations between these factors. An enriched understanding of how these attitudinal factors are associated with each other extends prior models of identity and link theoretical frameworks used in psychological and educational research. The research uses a series of quantitative and qualitative methodologies, including linear and logistic regression analysis, thematic interview analysis, and an innovative analytic technique adapted for use with student educational data for the first time: topological data analysis via the Mapper algorithm.

Engineering students were surveyed in their introductory engineering courses. Several factors are found to be associated with physics identity, including student interest

in particular engineering majors. The distributions of student scores on these affective constructs are simultaneously represented in a map of beliefs, from which the existence of a large "normative group" of students (according to their beliefs) is identified, defined by the data as a large concentration of similarly minded students. Significant differences exist in the demographic representation of this normative group compared to other students, which has implications for recruitment efforts that seek to increase diversity in STEM fields. Select students from both the normative group and outside the normative group were selected for subsequent interviews investigating their associations between physics and engineering, and how their physics identities evolve during their engineering careers.

Further analyses suggest a more complex model of physics and engineering identity which is not necessarily uniform for all engineering students, including discipline-specific differences that should be further investigated. Further, the use of physics identity as a model to describe engineering student choices may be limited in applicability to early college. Interview analysis shows that physics recognition beliefs become contextualized in engineering as students begin to view physics as an increasingly distinct domain from engineering.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

Chapter I: Introduction

The President's Council of Advisors on Science and Technology has argued for increasing the number of STEM graduates by approximately one million over the next decade when the report was issued in 2012, in order to maintain economic competitiveness, growth, and quality of life in the United States (National Academies, 2007, 2010; President's Council of Advisors on Science and Technology, 2012). The shortage of STEM professionals is particularly pertinent to the fields of physics and engineering where fewer women, African Americans, and Hispanics graduate than what is commensurate with their population sizes (NRC, 2013). The President's Council's primary suggestion to achieve this goal was to increase undergraduate retention of STEM majors; 48 percent of students who entered STEM fields at the start of the 2003-2004 academic year seeking their bachelor's and 69 percent of those seeking their associate's degree had left by spring 2009 (Chen & Soldner, 2013). While these rates are comparable with other fields like humanities, health sciences, and business, they nevertheless reveal a massive loss of majors which, if it could be reduced by as little as 10%, would result in hundreds of thousands of additional students graduating in STEM fields. Therefore, an understanding of which factors are related to or lead to increased persistence (and thus reduced attrition) is key to achieving this goal of more graduates. Further, because "identification with a group or community of STEM professionals may overshadow many other factors in determining persistence" (President's Council of Advisors on Science and Technology, 2012), the current work maintains a focus on identity in particular among several theoretical frameworks.

Introductory-level university physics courses (both algebra-based and calculus-based) are taken by students in a wide range of STEM majors, only a small fraction of which are physics majors. Instead, these courses serve the undergraduate STEM population as a whole and provide some physics instruction for students with a wide variety of career intentions. One large sub-population of students taking introductory college physics is engineering majors, many of whom will use physics-related ideas throughout their studies and will pursue careers in the physical sciences/engineering.

A wide variety of theoretical frameworks either directly address student persistence, engagement, and retention, or have been linked to them in prior research. Student identification with physics, as described in the identity framework of Hazari et al. (Hazari, Sonnert, Sadler, & Shanahan, 2010), Carlone (Carlone & Johnson, 2007), etc. has been found to be a strong predictor of student persistence in physics, and intentions related to a career in physical science (Godwin, Potvin, Hazari, & Lock, 2016). Other affective factors have been separately studied in the context of student science-related performances. For example, a students' *sense of belongingness* has been linked to persistence in their college program and their performance (Freeman, Anderman, & Jensen, 2007; Pittman & Richmond, 2008). Also, the personality traits of *grit* and *conscientiousness* have been consistently associated with academic success and persistence (Duckworth, Peterson, Matthews, & Kelly, 2007; Trapmann, Hell, Hirn, & Schuler, 2007). The Big Five personality traits (McCrae & John, 1992) have also been linked to academic motivation (Clark & Schroth, 2010; Komarraju, Karau, & Schmeck, 2009). And foremost, student identity as a science, physics, or engineering person has been linked with performance, retention, and eventual career choice in a STEM field

(Carlone & Johnson, 2007; Godwin et al., 2016; Hazari et al., 2010; Plett, Hawkinson, Vanantwerp, Wilson, & Bruxvoort, 2011). However, many of these studies have focused, for theoretical or practical reasons, on a single affective factor in any one study, rather than exploring the relative role of several at one time, though there are some exceptions which examine a handful of related factors at one time (e.g., Grit and the Big Five (Duckworth et al., 2007)).

In this dissertation, I examine the attitudes of engineering students, with a focus on their physics identities and related attitudinal constructs. Physics identity has been previously found to be a critical predictor of engineering-related career choices at the precollege-to-college transition (Godwin et al., 2016). Specifically, in a nationally-representative study of college freshmen, three factors were found to be predictive of engineering choice in college: students' precollege physics and math identities, and their *agency beliefs*: beliefs in the power of science and engineering to impact one's life and the world around oneself. Unlike other domains, students who pursue engineering majors in college often have few direct engineering experiences or course-taking (Katehi, 2009), so the importance of identities in other related domains—physics and math—is increased. Once students gain a number of direct engineering experiences—say, by taking college engineering courses—then the importance of a physics or math identity to their engineering pursuits may diminish over time. At the start of college, these other domain identities remain quite relevant, which is why the current study focuses on early college experiences.

Chapter 1 introduces the research topic and provides background and motivation for the conducting these studies. I then introduce the research questions featured in each

chapter. I describe the theoretical frameworks informing the current work and the affective constructs from each that were measured and analyzed in the subsequent chapters. The chapter finishes with a description of the survey used to collect the initial student data.

Chapter 2 addresses the first two research questions through multiple linear regression analysis. I present two regression analyses: the first looks at which affective constructs are significantly related to students' physics identities, while the second analysis includes interactions with students' interest in particular engineering majors to examine whether the pattern of significance is different for various groups.

Chapter 3 answers the third and fourth questions by introducing topological data analysis, a new method in education research used to construct a representation of the space of affective beliefs. I combine it with traditional statistical analyses (proportion tests, logistic regression, and various tests of difference in means) to understand the representation and look for significant effects in both attitudes and the representation of traditionally-underrepresented demographics.

Chapter 4 builds on the results of the previous chapter by qualitatively analyzing interviews from individuals selected using the results from Chapter 3 to answer the final two questions. Interviews were coded by thematic phenomenological analysis, and the results are presented and discussed.

Chapter 5 finishes the dissertation by reflecting on the findings of each chapter *in toto* and in combination and discussing implications and directions for future work.

Research Questions

This dissertation seeks to answer the following questions throughout its chapters:

Chapter 2:

1. For the introductory engineering students at the four collaborating institutions, how are various attitudinal factors associated with students' physics identity beliefs?
2. How are the associations identified in Research Question 1 mediated by students' interests in various engineering disciplines?

Chapter 3:

3. How are students distributed in the space of affective beliefs?
4. What demographic differences exist between students holding normative beliefs and those with non-normative beliefs?

Chapter 4

5. How do students' perceived connections between engineering and physics change as they become more experienced in engineering?
6. How does the nature of students' physics recognition beliefs change over time?

Background and Literature Review

Increasing the diversity in engineering education has been a priority of educators and education researchers for at least the past 30 years. Despite years of research and reform, the enrollment of demographically diverse individuals in undergraduate engineering degree programs has not substantially improved. In much work that has studied diverse student experiences, an approach is often taken to divide students on the basis of singular (or a small set of) demographic identifiers (e.g., Black or White; male or

female; etc.). These categorizations often serve to bin students and generalize findings for women or underrepresented minority students in a way that seeks to highlight the issues faced by underrepresented groups and/or identify ways to support such students effectively. However, one limitation of this general approach is that it often ignores the multitude of identities and holistic experiences of individuals that combine uniquely for every person. That is, such a traditional approach to understanding diversity does not take into account the rich and nuanced differences in individuals' experiences. Further, this approach of binning students into predefined demographic categories may not faithfully account for the true spectrum of motivations, attitudes, and goals of individual students since people with a variety of affects may be "binned" together as a presumed-homogenous group, thus missing out on a more nuanced and faithful understanding of students, as demographic diversity does not necessarily have a one-to-one relationship with affective diversity.

Examining the multi-faceted aspects of student identities can provide a more holistic understanding of students' attitudes and beliefs than examining just one particular dimension of students' identities. Individuals have multiple overlapping identities that comprise their affiliations, experiences, attachments, and social engagement. Foregrounding just one of these identities in an analysis potentially limits the richness of understanding a person as a whole and how their multiple identities impact how they are positioned and position themselves in the world.

Intersectionality

An approach to understanding multiple overlapping identities has its roots in intersectionality theory. Originating from critical legal studies (Crenshaw, 1989, 1991), the theory examines how multiple intersecting identities form interacting layers of oppression in society. Kimberlé Crenshaw (1989) first put forward this understanding of how identities intersect from her experience studying case law. In one case, *Emma DeGraffenreid et al. v. General Motors Assembly Division* (1977), a woman of color, Emma DeGraffenreid, was fired from General Motors. She and four other Black women brought legal suit against the company citing discriminatory labor practices. In the company, white women did one set of jobs (mostly secretarial) and white men did another set of jobs (management). Additionally, Black individuals were hired in the hands-on labor jobs while white individuals did the clerical or office jobs. The issues for Black women were compounded. Jobs for Black individuals were "men's jobs", and the jobs for women were "white jobs". Black women faced double challenges when applying for positions within the company. When the case came to court, the judge dismissed the case citing that the company had representative numbers of both Black employees and female employees. The court would not allow the claimants to combine racism and sexism into one suit. Because Emma could not demonstrate that the discrimination she faced was along purely racial or gender lines, she could not prove her claim. This injustice allowed the intersections of both race and gender to be ignored and prompted Crenshaw to develop the theory of Intersectionality.

Intersectionality theory provides a way to identify and examine the relationship between individuals' multiple identities and structures of power. In her work, Crenshaw

identified different variations in experience for Black women. Sometimes they had similar experiences to Black men or to white women. Other times, they faced additive or multiplicative effects ("double discrimination"), whereas in other cases, they had particular experiences specific to their status as Black women. It is important to note that some members of disadvantaged groups also hold, in part, privileged identities (e.g., middle-class Blacks, White women in STEM). These variations of experiences reveal that although much of the literature on intersectionality has been theorized from the standpoint of those who experience multiple dimensions of disadvantage, this framework can also inform how privileged groups are understood (Cole, 2009).

The present study utilizes intersectionality theory in a new way to better understand the underlying attitudes and beliefs of students. Rather than pursuing a critical analysis of power and positionality, I instead use it as a guiding principle in examining multiple intersections of students' attitudes, beliefs, and identities to more faithfully understand the students who are pursuing college engineering and what underlying attitudes might be privileged within engineering culture. Approaching research with a mind towards intersectionality provides a different, but complementary, way to understand the nuanced differences and similarities among engineering students. I acknowledge that my focus is on the intersections of student identities and not on a critique of power and positionality within the existing social structures of engineering programs. This approach enables an understanding of underlying attitudes and beliefs, influenced by college engineering students' incoming attitudes, that shape students' experiences within engineering, reify engineering culture, and promote or deter an individual's persistence in engineering.

Identity framework

Identity is a framework of analysis (Chachra, Kilgore, Loshbaugh, Mccain, & Chen, 2008; Gee, 2000) used to study student engagement, belonging, and persistence in STEM, including physics, mathematics, and engineering (Cass, Hazari, Cribbs, Sadler, & Sonnert, 2011). Very broadly, one's identity describes how they see themselves and interact with the world. One can hold many different identities, corresponding to different spheres of life, activating each identity when it is relevant.

In the context of my study, physics identity can be thought of as the extent to which someone sees themselves as a "physics person" (Lock, Castillo, Hazari, & Potvin, 2015); likewise, someone with a strong math identity sees themselves as a "math person", and someone with a strong engineering identity sees themselves as an engineer. In the framework developed by Hazari et al, it is conceptualized as a quasi-trait—something which is relatively stable but which can change over time as a result of experiences (Cribbs, Sadler, Hazari, Conatser, & Sonnert, 2013; Hazari et al., 2010; Potvin & Hazari, 2013). A student's identity is constructed of three sub-constructs (Carlone & Johnson, 2007; Godwin, Potvin, & Hazari, 2013; Hazari et al., 2010). *Performance/Competence beliefs*, originally constructed as two separate factors (Hazari et al., 2010) which were experimentally indistinguishable in repeated measurements of students in high school or early college (Potvin & Hazari, 2013), describes a student's belief in their ability to succeed at physics both in terms of understanding the content, and in terms of their performance (e.g., exams). *Recognition beliefs* describe students' beliefs that others, including parents, instructors, and peers, recognize them as a physics person (in the case of physics identity, for example). *Interest*, which was not present in Carlone and

Johnson's original construction of science identity (Carlone & Johnson, 2007) but later emerged in discussions of domain-specific identity (Hazari et al., 2010), describes a student's interest and enjoyment in learning about the subject and doing related tasks.

I use the quantitative identity framework developed by Hazari et al. (Hazari et al., 2010) to describe science identity and physics identity, which has been replicated in engineering (Godwin, 2016; Godwin, Potvin, & Hazari, 2013) and math (Cribbs, Hazari, Sonnert, & Sadler, 2015; Godwin et al., 2016). The overall measure of Physics Identity is constructed from the three sub-constructs, Physics Performance / Competence, Physics Recognition, and Physics Interest, which are combined in an unweighted average to give an overall score. Math Identity is similarly constructed of three domain-specific sub-constructs, measured with similar items that are framed in terms of math instead of physics (Cribbs et al., 2015; Godwin, Potvin, Hazari, & Lock, 2013). However, the relationship between these two identities has not yet been fully explored, even though they have been used together as predictors of other outcomes (e.g., engineering identity or interest in pursuing a career in engineering (Godwin, Potvin, & Hazari, 2013; Godwin et al., 2016)).

A note about social cognitive career theory

Social cognitive career theory (SCCT) (Lent, Brown, & Hackett, 1994) has been used in engineering education research for studying career choice (e.g., (Carrico & Tendhar, 2012; Sheu & Bordon, 2017)). Social cognitive career theory combines aspects of social cognitive theory (Bandura, 1977, 1997, 1999). Instead of using this framework to shape the current analysis, I instead chose to focus on the identity framework presented

by Hazari et al., because it better predicts engineering career choice (Godwin et al., 2016) than SCCT alone. Furthermore, SCCT does not account for recognition beliefs, which have been found to be integral to the identity framework and on related career choices, and instead focuses primarily on performance/competence beliefs, which prior work with the identity framework has shown has only an indirect effect on career choice, mediated by interest and recognition beliefs (Godwin et al., 2016).

Other salient theoretical constructs

Belongingness

Belongingness is a measure of how accepted, comfortable, and welcome a student feels in their engineering classroom and program, which contributes to academic engagement and achievement (Freeman et al., 2007; Pittman & Richmond, 2008). In the survey used in this dissertation, this factor is domain-specific to engineering. Example items include: "I feel welcome in engineering," and "I feel supported in my engineering class." This construct was developed by the research team for the InIce survey, following prior literature. Originally envisioned as several factors constructed from many more items, the pilot factor analysis showed that a single overall factor was appropriate, as the distinctions between hypothesized sub-constructs were not present.

Achievement Goal Theory

Performance Approach and *Mastery Approach* are drawn from Achievement Goal Theory, and describe why a student engages in behaviors related to their achievement (Dweck & Leggett, 1988). Students who take a performance approach engage in behaviors to display their competence to others (example items: "Proving to

my peers that I am a good student", and "Getting a better grade than other students in this class") while for students with a mastery approach the focus is on developing competence and understanding (example items: "Really understanding this course's material" and "Feeling satisfied that I got what I wanted from this course"). Related to these two is *Work Avoidance* (Dowson & McInerney, 2001), in which the student's goal is to minimize the amount of effort required in order to pass the requirements (example items: "Getting a passing grade with as little studying as possible" and "Not having to work too hard in this class"). The combination of these three factors influences how students approach problem-solving, learning, and their education as a whole.

Kaplan and Flum (Kaplan & Flum, 2010) connected these approaches to generalized identity formation and argued that the mindsets and approaches of Achievement Goal Theory are related to the mindsets and approaches students use when forming their identities. They raised questions of whether identity formation styles inform which achievement goal mindset a student employs in a particular situation.

Expectancy-Value Theory

Expectancy is drawn from Expectancy-Value Theory (EVT) (Eccles et al., 1983; Eccles & Wigfield, 2002), and describes how well someone expects to do on a task, in the present. An expectation of success is informed by a students' socialization, including gender and cultural stereotypes, and past performances on similar tasks. Example questions measuring this construct include "I expect to do well in this engineering course" and "I am confident I can do an excellent job on the assignments in this engineering course." Notably, these questions are a measure of students' expectation of

their academic success in a class, as opposed to an engineering program as a whole, their college experiences, or their later careers. This construct is related to but distinct from self-efficacy (Bandura, 1997). According to expectancy-value theory, expectation of success at a particular task (including broad definitions of a task like "pass a course") combined with the student's subjective task value (a combination of what they gain from doing the task and what it will cost them) influences their choice of actions and overall performance (Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2007). Self-efficacy, like expectancy, addresses student perception of success at a task, but the scope of what constitutes a task is smaller and more focused on the present (e.g., solving a particular kind of math problem right now, versus passing a math class). Expectancy also overlaps with performance/competence beliefs in a particular domain, but again differ in scope. In the case of this survey, a student's Expectancy describes their expectation of success in this particular engineering course and its assignments; their Engineering Performance/Competence beliefs describe their ability to do engineering more generally. While performance/competence beliefs may be related to classroom participation, they represent a broader set of beliefs not tied to specific classroom or other contexts.

Future Time Perspective

Connectedness, *Instrumentality*, *Value*, and *Perceptions of Future* are all aspects of Future Time Perspective (FTP) theory (González, Fernández, & Paoloni, 2016; Husman & Lens, 1999; Kirn, Faber, & Benson, 2014; Simons, Vansteenkiste, Lens, & Lacante, 2004) which expands existing motivation theories to explicitly include time considerations in values and goal setting.

Connectedness is a measure of the perceived interconnectedness of the present and future, in general (example items [negatively-coded]: “I don’t like to plan for the future”, “It’s not really important to have future goals for where one wants to be in five to ten years.”).

Perceptions of Future describe how certain a student is that they are going to have a future career in engineering, and how positively they view that future (example item: “I want to be an engineer”).

Instrumentality is a measure of how connected or useful one feels their current tasks are for one’s future career and success. Perceived instrumentality is a context-specific measure and relates to one’s emerging identity (example item: “I will use the information I learn in this engineering course in the future”). In other words, what the value of the current task (i.e., taking and passing their engineering course) is to their future lives. Instrumentality has been associated with student performance; students with a positive perceived future and high instrumentality have higher motivation and performance for tasks related to that future, while students with a negative perceived future and high instrumentality see decreases in both motivation and performance (Simons et al., 2004). Here, I measure Instrumentality as it relates to a future career as an engineer.

Value is a statement about the worth of the future as compared to that of the present (example item: “Long range goals are more important than short range goals”). Value, as used in FTP here, is a distinct idea from that used in EVT, which is more similar to the Instrumentality construct (specifically, Instrumentality is a measure of Utility Value the explicitly considers time).

Kirn, Faber, and Benson (2014) describe how students with particular combinations of these FTP constructs fashion their identities in distinct ways. Students with high Connectedness, Perceptions of Future, and Instrumentality (called "sugar cone students" in this earlier work) had clear and detailed ideas of what they wanted to do and be in the future, as well as clear paths to achieve that future. Sugar cone students were able to envision possible futures containing both positive outcomes (the person they wished to become) as well as outcomes they wished to avoid, negative futures closely related to their ideal future (e.g., a student who wants to become an anesthesiologist, and doesn't want to become a surgeon of doctor; both the ideal and the avoided futures are similar in kind).

Grit

Grit is defined as perseverance and passion for long-term goals (Duckworth et al., 2007), and has been associated with success such as job retention and scholastic achievement (Duckworth & Quinn, 2009; Eskreis-Winkler, Shulman, Beal, & Duckworth, 2014). A person's grit can be divided into two sub-constructs. Example items from *Persistence of Effort*, or perseverance for long-term goals, include "I have overcome setbacks to conquer and important challenge" and "I finish whatever I begin". *Consistency of Interest* describes the student's passion and commitment to long-term goals. Example items include (negatively coded) "I have difficulty maintaining my focus on projects that take more than a few months to complete," and "My interests change from year to year."

Agency Beliefs

Agency beliefs refer to a student's perception of their ability to change their world through their everyday actions and life goals (Basu, Calabrese Barton, Clairmont, & Locke, 2009; Godwin, Potvin, & Hazari, 2013; Turner & Font, 2003), and have been previously connected to the decision to enter engineering and interest in various engineering fields (Potvin et al., 2013). These studies investigated both "Personal" and "Global" Agency Beliefs and found that Personal Agency Beliefs were positive significant predictors of decisions to enter college for science or engineering (Godwin, Potvin, & Hazari, 2013; Godwin et al., 2016). The questions measuring Personal Agency Beliefs were included in the survey as a measurement of *Science Agency Beliefs*, in contrast with a related set of questions, which were similarly phrased, but concerning engineering. For example, an item from the Science Agency beliefs factor is "Science is helpful in my everyday life," whereas a similar item from the *Engineering Agency Beliefs* factor is "Engineering can improve our society". Both talk about the impact of science or engineering, but Science Agency Beliefs are focused on the student (with "I" and "me" phrases), while Engineering Agency Beliefs are somewhat more externally focused on engineering, though still interested in how it relates to the student and their ability to affect the world. In prior research (Godwin, Potvin, & Hazari, 2013; Godwin et al., 2016) these were studied in tandem with physics identity to predict students' choice to go into engineering, but the constructs were non-interacting in that model.

The "Big Five" Psychological Traits

The "Big Five" Psychological Traits describe a five-factor model of personality that has solidified through several decades of research (Judge & Ilies, 2002; McCrae & John, 1992; Zillig, Hemenover, & Dienstbier, 2002). These five traits are Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. *Neuroticism* describes the tendency to show poor emotional adjustment in the form of stress, anxiety, and depression, and has alternatively been positively and negatively associated with student GPA (Noftle & Robins, 2007; Trapmann et al., 2007). *Extraversion* represents the tendency to be sociable, outgoing, and positive. *Agreeableness* describes tendencies to be kind, gentle, trusting, trustworthy, and warm. *Conscientiousness* describes the ways in which individuals are dutiful, orderly, deliberate, and self-disciplined; Conscientiousness has been consistently positively associated with academic success at the high school and college levels (Dumfart & Neubauer, 2016; Rimfeld, Kovas, Dale, & Plomin, 2016; Trapmann et al., 2007). High scores *Openness to Experience* are associated with people who are creative, flexible, curious, and unconventional. The Big Five have also been associated with student motivation, whether extrinsic or intrinsic (Clark & Schroth, 2010; Komarraju et al., 2009; Ryan & Deci, 2002). Neuroticism and Extraversion have been associated with extrinsic motivations to succeed academically, while Openness has been associated with high intrinsic motivation to know and experience stimulation. Conscientiousness has been associated with both kinds of motivation, and Agreeableness has been negatively associated with disengagement from learning (i.e., work avoidance).

Grit and Conscientiousness are highly correlated with each other, and persistence has been identified as a major facet of Conscientiousness in studies probing the underlying factor structure of that personality trait (MacCann, Duckworth, & Roberts, 2009). However, other studies (Eskreis-Winkler et al., 2014; Rimfeld et al., 2016) have shown that while Grit remains a significant predictor of several life outcomes while controlling for Big Five personality traits (including and especially Conscientiousness), it explains a small additional amount of variance.

Relationship Between Theoretical Frameworks

The spectrum of theoretical frameworks employed in the current work is broad, though some of these frameworks are partially overlapping. In part because the theories from which I drew these constructs were developed independently, the concepts described by each are not necessarily unique to that theoretical framework. For example, expectancy-value theory and the identity framework overlap in that both theories describe a person's belief about their ability to accomplish something. However, as described above, Expectancy (from expectancy-value theory) and Performance/Competence beliefs (from the identity framework), while related, are still distinct, and worth considering together. Similarly, the three identity constructs used (physics identity, math identity, and engineering identity) have each been strongly associated with each other in prior research, as discussed above.

Grit and the "Big Five" psychological traits have been studied in relation to each other for years, due to the correlations between grit and conscientiousness, and have each been linked time and again to persistence and success in academic settings. They also

overlap with some of the future-oriented constructs of future time perspective theory; connectedness and value both describe beliefs which are related to the dutifulness and industriousness facets of conscientiousness.

Though correlations between these constructs exist, and hint at a connected theoretical space that might encompass them all, each theory has individual differences that make it unique from the others and independently worthy of consideration in the overall combined analysis because of its potential for increased explanatory power. In terms of an entire space of affect, the chosen theories cluster in a relatively small space, as each has been chosen because of its association with academic performance, persistence, etc. Further, the affective constructs are all quantitatively characterized, which limits their ability to accurately describe small, nuanced differences between students in favor of better describing broad patterns. The result is a relatively broad brush with which to describe students' affect mostly as it relates to their academic lives.

Chapter II: Survey Development and Deployment

In this chapter, I describe the process by which the survey used in this work was developed and deployed[1].

Survey Development

The theoretical constructs used in this work are latent variables which are cannot be directly measured. Proxy measurements can be directly made with related questions; the overall trend of those items can stand as a proxy for the latent variable. These proxy measurements are assessed by running a factor analysis on the measured questions, and determining which questions load onto which factors. Each factor described by the factor analysis corresponds to a particular theoretical construct, and each question is given a "loading" by the analysis which corresponds to how strongly the responses to that question correspond to the overall factor. In mathematical formalism, given n-many sets of d random variables $x_n = \{x_{1,n}, \dots, x_{d,n}\}$, with overall means $\mu = \{\mu_1, \dots, \mu_d\}$, a factor analysis with k-many factors seeks to solve the equation $\boldsymbol{x} - \mu = \boldsymbol{L}\,\boldsymbol{F} + \varepsilon$, where $\boldsymbol{x}$ is the $d \times n$ matrix of observed variables, $\boldsymbol{L}$ is a $d \times k$ matrix of loadings, $\boldsymbol{F}$ is a $k \times n$ matrix of factors values for each observation, and ε is a $1 \times n$ vector of uncorrelated errors which are independent of $\boldsymbol{F}$. Thus, through $\boldsymbol{L}$, a particular observation x_n can be converted into a list of numbers F_k describing the scores for that observation on each factor or latent variable. (For more information, see e.g., Graffelman, 2012).

[1] This material is based upon work supported by the National Science Foundation under Grants No. 1428689 and 1428523. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Items for the chosen theoretical constructs were taken from previously developed surveys. The expected factor structure was established with an exploratory factor analysis (EFA) using a promax rotation[2] on data from a pilot survey. An exploratory factor analysis was used because many of these questions had not yet been used together with the population under study. This rotation was chosen to maximize interpretability of the factors since inter-correlation was expected between several factors. For example, the identity sub-constructs are known to be well-correlated and interrelated, so forcing those factors to be orthogonal (i.e., with another rotation choice) would reduce the ability of that factor to accurately describe the underlying construct. Results were used to shorten the survey by eliminating poorly performing items. Items with low loadings onto their factor were removed. As a first pass, items needed to have a loading of higher than 0.4 on corresponding factors; subsequent passes increased this cutoff on a factor-by-factor basis depending on the number of questions remaining in the factor. In the end, each factor was measured with 3-5 items which performed best. For newly-developed questions, items which had loadings that were split between multiple factors were particularly targeted for removal to improve factor interpretability, and no such items remained in the final survey.

Personality tests designed to measure the "Big 5" psychological traits can be hundreds of items long. To reduce survey fatigue, the survey tried to measure these constructs with as few questions as possible without affecting reliability. Starting with a

[2] Promax rotation allows the resulting factors to be correlated, as opposed to forcing them to be orthogonal in a varimax rotation.

50-item instrument from Goldberg (1992), the number of items for each factor was reduced to five by choosing the items with the highest loading in a five-factor EFA, as described above. Credé et al. have shown (2012) while two-item measures of these psychological traits have reduced reliability, the reliability quickly increases with just a few more items. Thus, while the measurement of students' psychological traits may not have the nuance to separate into the various facets of each trait (i.e., six facets of a trait cannot be accurately measured with only five items), the measure of the overall trait can still be considered valid.

The factor analysis revealed 26 theoretical constructs underlying the questions about attitudes and beliefs, drawing from a variety of affective theories as discussed above. Some of these constructs were developed by the research team, for this project or in prior work, and others were drawn from the literature as being relevant to engineering student academic success, performance, learning, retention, and STEM career choice. The numeric results factor analysis establishing this structure and item loadings are included in Chapter 3.

The demographic questions at the end of the survey were developed in large part by the research team, or adapted from either the National Survey of Student Engagement (NSSE) or the Sustainability and Gender in Engineering (SaGE) surveys (Fernandez et al., 2016). Questions were constructed to be as broadly reaching and inclusive as possible; i.e., a "select all that apply" response structure was used for questions about ability/disability status, race and ethnicity, gender identity, sexuality, parental/guardian gender identity(s), and family occupations, and more inclusive response options were provided than, for example, a simple gender binary.

Students' current major was asked as an open-ended, fill-in-the-blank question. These open responses were then cleaned by hand to remove unnecessary variations while retaining as much information as possible. For example, responses of "ME", "Mech. E", and "Mechanical Engineering" were all interpreted to mean "Mechanical Engineering" for subsequent analysis. In all, 23 unique majors were provided with an additional 49 unique combinations (i.e., two or more majors simultaneously reported), though the majority of responses (54.4%) fell into one of three well-populated majors[3]: 24.9% of students responded that they were "General Engineering" majors, 15.3% responded with "Mechanical Engineering", and 14.3% responded with "First-year Engineering". The next most popular response was "Civil Engineering", with 5.96%, significantly lower than the three most common response categories.

In addition, students were asked directly to asses their current interest in several different majors, each on an anchored scale from 0 (not at all interested) to 6 (extremely interested). The majors included all of the engineering majors offered at the four participating institutions[4], as well as "Other STEM-related Degree" and "Other non-STEM-related Degree".

The final version of the survey consisted of 22 (multi-item) questions, including several affective constructs (described above), their current major, their career interests,

[3] For example, "Mechanical Engineering" and "General Engineering" were two popular categories. "Mechanical Engineering and General Engineering" (if the student wrote both on their survey) was considered a unique response and had many fewer responses.

[4] For a full list, see the final survey in Appendix.

and demographic factors. Affective items used anchored scales (on 0 to 6 scales), while demographic questions were all select-all-that-apply.

Survey Deployment

The survey was developed in Spring 2015 by a four-institution collaboration between Florida International University (FIU); University of Nevada, Reno (UNR); Clemson University; and Purdue University as part of the Intersectionality of Non-normative Identities in the Culture of Engineering (InIce) project. Questions measuring student affect were drawn from previously completed survey studies performed by the grant PIs (Godwin, 2016; Godwin, Potvin, & Hazari, 2013; Hazari et al., 2010; Kirn & Benson, 2013; Potvin et al., 2013; Potvin & Hazari, 2013) or from instruments developed and discussed in the literature (Duckworth & Quinn, 2009; Goldberg, 1992; Husman, Lynch, Hilpert, & Duggan, 2007). These questions were revised and pared down following a piloting of the survey during Spring 2015 at three of the institutions. The pilot survey had 537 respondents (223 from UNR, 78 from Purdue, and 236 from Clemson). See Appendix on page 148 for the final survey version.

At the beginning of the Fall 2015 semester, engineering students were surveyed at the four participating institutions. Surveys were administered between August 15th and September 14th. Students were recruited because of their enrollment in each institution's introductory engineering classes and were surveyed during class time with paper & pencil instruments during the first two weeks of the semester, before students had significantly progressed into their courses.

Student participation was voluntary and anonymous, though at the end of the survey students were asked to provide a contact e-mail address if they were willing to participate in follow-up interviews at a later date. As the survey was given during class time with nothing else to distract the students, the participation rate was high (average response rate of 70.7%, with the response rate at each institution being over 65%). In all, 2916 responses were collected (514 from UNR, 1104 from Purdue, 1050 from Clemson, and 298 from FIU[5]). A confirmatory factor analysis of the survey data confirmed that the factor structure from the pilot survey persisted.

[5] FIU had an undergraduate engineering population of approximately 2,800 students. Purdue had approximately 7,640 undergraduate engineering majors. Clemson had approximately 1160 general engineering majors, which all first-year engineering students take before later specializing. UNR has approximately 2610 undergraduate engineering students. Thus, though the numbers of students at each institution are not equal, the sample sizes reflect the relative sizes of the student populations of interest at each institution.

Chapter III: Attitudes associated with Physics Identity

Introduction

In this chapter, I reintroduce the two research questions that will be investigated herein and then give a brief overview of the statistical methodology which will be used to conduct the analysis. I present the results of the factor analyses and first linear regression model, followed by a discussion of those results. Then, I present a series of additional linear regressions, a discussion of those results, and an overall discussion of the implications of these findings. I conclude with a discussion of the limitations of this research.

Motivating the search for discipline-specific effects

Engineering is a diverse set of fields that deal with a wide variety of subjects and contexts, and different engineering disciplines can be appealing to different people; the sort of person who wants to become a mechanical engineer is not necessarily the same person who wants to become a chemical engineer. With this in mind, I wanted to investigate whether a model of which attitudinal factors associated with physics identity changed with the addition of discipline-specific effects. Specifically, it was of interest to understand whether a student's interest in a particular major mediates the effect of other affective factors discussed in the previous. Prior research has shown differences between various engineering disciplines with regards to how students' intentions to pursue a career in that discipline are associated with particular factors, including Physics Identity, Math Identity, and Science Agency Beliefs (Potvin et al., 2013). Whereas prior work used each factor independently (i.e., factors were used simultaneously as predictors in a

model, but one factor was not used to predict another), this section extends the analysis to investigate similar discipline-specific effects while simultaneously accounting for multiple affective factors.

Research Questions

I investigate the following research questions in this chapter:

1. For the introductory engineering students at the four collaborating institutions, how are attitudinal factors associated with students' physics identity beliefs?
2. How are the associations identified in Research Question 1 mediated by students' interests in engineering disciplines?

Answering these questions may help to illuminate some of the connections between previously-independently-considered factors which have been studied in relation to student choice, success, and persistence in STEM. Knowing about these associations can help guide future research towards more nuanced and sophisticated explanatory models, and clarify new effects by better controlling for previously known results. And answering the second research question can provide additional depth and nuance to the findings from the first question if it turns out that the sort of engineering being considered can drastically change how these factors interact with each other.

Methodology

The analytic methodology for this chapter and the development of the InIce survey are related. The goal is to determine the association between several theoretical constructs and the one of primary interest, physics identity. The factor structure of these constructs was established and confirmed as described in Chapter 1; the factor loadings

for each item in these factors is discussed below. To examine the association between several factors and Physics Identity, I created a linear model predicting Physics Identity as a function of the other factors using multiple linear regression. That model was iteratively improved by removing factors which were found to be non-significant to create a final primary model of the associations between attitudes and physics identity.

Students were surveyed near the beginning of their engineering program, and a plurality had not yet declared a major beyond "First Year Engineering" or "General Engineering", as is typical for the two largest engineering programs studied. Over 40% of the students responded in this way to a survey item (Q11) probing this. However, included in the survey was a question (Q14) asking students to "Please rate your interest in the following majors" with several response categories, each on an anchored scale from zero ("Not at all") to six ("Very much so"). See the full InIce survey in the appendix for the full wording of questions Q11 and Q14 (page 155). Associations between the student responses to "What is your current major?" and these interest items are high; the correlation between a student's interest in a major and their declared major ranged between 0.5085 and 0.7462, with a mean correlation of 0.6337, which can be interpreted as evidence for concurrent criterion-related validity of Q14 as a proxy for students' major. See Figure 1 for more details.

Figure 1 – Polyserial correlation between having a declared major and interest score.
Q11 (vertical axis) probed students' declared major, and their interest in each major was probed by Q14 (horizontal axis). Minimum correlation of a major with itself (diagonal terms) was 0.5085, with a mean correlation of 0.6337. The largest off-diagonal terms were -0.371, between mechanical engineering and a declared major in environmental/ecological engineering, and 0.322, between interest in agricultural biological / biosystems engineering and a declared bioengineering / biomedical engineering major. Abbreviations are explained in **Table 1**.

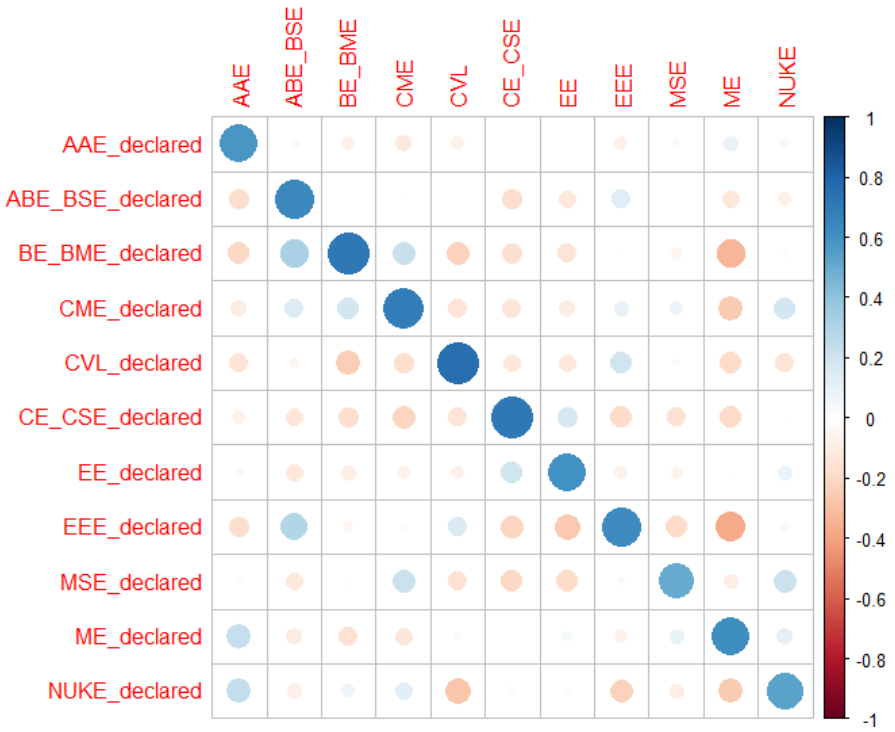
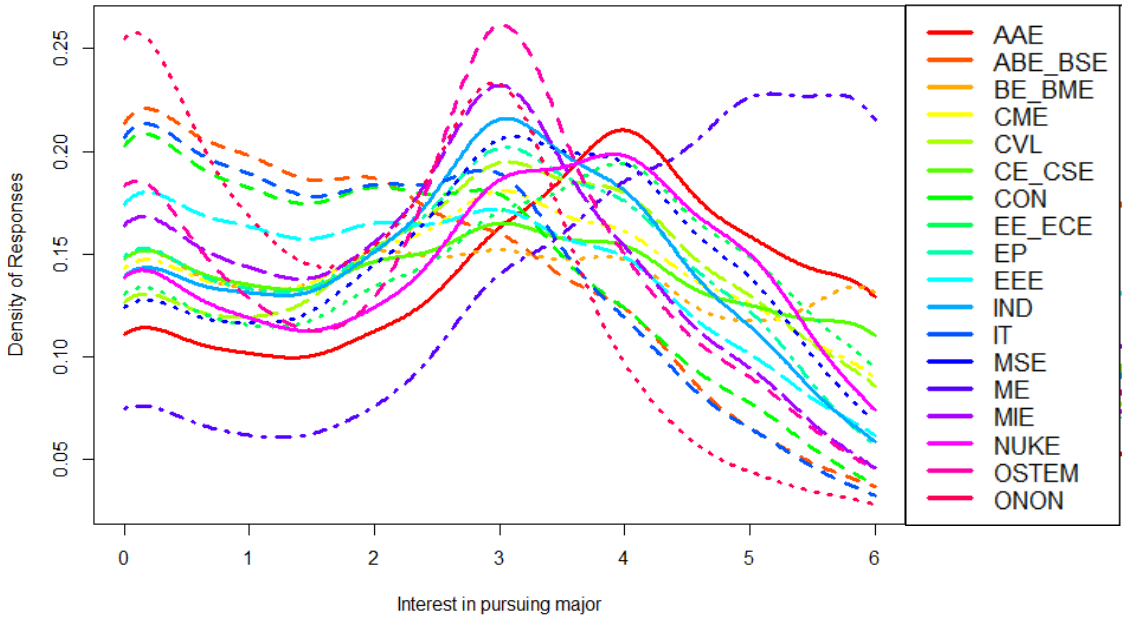


Table 1 - Abbreviations used for majors.
The chosen abbreviations are specific to this dissertation, and do not always reflect the canonical nomenclature.

Abbreviation	Full Name
AAE	Aero/Astronautical Engineering
ABE_BSE	Agricultural and Biological / Biosystems Engineering
BE_BME	Bioengineering / Biomedical Engineering
CME	Chemical Engineering
CVL	Civil Engineering
CE	Computer Engineering
CON	Construction Management Engineering
EE	Electrical Engineering
EP	Engineering Physics
EEE	Environmental / Ecological Engineering
IND	Industrial Engineering
IT	Information Technology
MSE	Materials Engineering / Material Science and Engineering
ME	Mechanical Engineering
MIE	Multidisciplinary / Interdisciplinary Engineering
NUKE	Nuclear Engineering
O-STEM	Other STEM-related Degree

Responses to the question on major interest were typically bimodal, with one peak being near zero (meaning, students declaring no interest in that major), and another in the 3-4 range (those students with significant interest). See Figure 2 for distributions of interest in each major. The only major which broke this trend was Mechanical Engineering, which had a significantly higher fraction of highly interested responses, with three times as many students answering each of 5 or 6 as compared to 0. However, this information matches with the information about declared majors from Q11, in which Mechanical Engineering was overrepresented compared to any other discipline, being the second highest number after only "General Engineering". Therefore, one would expect that a higher proportion of students would show an interest in Mechanical Engineering.

Figure 2 - Density estimates for responses to Q14
For each of the identified majors. Major names were abbreviated for space. Abbreviations for majors are explained in **Table 1**.



The primary model was then extended to multiple parallel models, each corresponding to the addition of another regressor describing student interest in a

particular engineering major. Each interest was added to the model, tested, and then removed before adding the next interest (i.e., the extended models each had only a single additional predictor compared to the primary model). The list of interests was drawn from Q14 in the survey and represented all of the engineering major choices available at the four institutions administering the survey.

Results of the Primary Model

The survey questions to measure physics identity and their loadings on these factors are described below, followed by the questions and loadings for the other attitudinal factors included in this dissertation.

Table 2 - Factor Loadings for Physics Identity sub-constructs

Survey Item	Factor			Variance Explained
	Performance/ Competence	Recognition	Interest	
My parents see me as a physics person.		0.776		23.9%
My instructors see me as a physics person.		0.840		
My peers see me as a physics person.		0.923		
I've had experiences in which I was recognized as a physics person.		0.714		
Others ask me for help in physics.		0.567		
I am interested in learning more about physics.			0.801	14.7%
I enjoy learning physics.			0.874	
I find fulfillment in doing physics.			0.674	
I am confident that I can understand physics in class.	0.927			25.0%
I am confident that I can understand physics outside of class.	0.903			
I can do well on exams in physics.	0.840			
I understand concepts I have studied in physics.	0.728			
I can overcome setbacks in physics.	0.467			
Total Variance Explained				63.6%

Table 3 - Factor loadings for Belongingness

Survey Item	Factor	Variance Explained
	Belongingness	
I feel comfortable in engineering.	0.837	64%
I feel I belong in engineering.	0.825	
I enjoy being in engineering.	0.818	
I feel comfortable in my engineering class.	0.837	
I feel supported in my engineering class.	0.727	
I feel that I am part of my engineering class.	0.748	
Total Variance Explained		64%

Table 4 - Factor loadings for constructs from Grit

Survey Item	Factor		Variance Explained
	Consistency of Interest	Persistence of Effort	
My interests change from year to year.	0.634		28.1%
I have been obsessed with a certain idea about a project for a short time but later lost interest.	0.885		
I often set a goal but later choose to pursue a different one.	0.905		
I have difficulty maintaining my focus on projects that take more than a few months to complete.	0.627		
Learning science has made me more critical in general.	0.624		
Engineering can improve our society.		0.791	29.3%
Engineering will give me the tools and resources I need to make an impact		0.805	
Engineering can improve our quality of life.		0.921	
I see engineering all around me.		0.678	
Engineering allows me to think deeply about problems.		0.564	
Total Variance Explained			57.4%

Table 5 - Factor loadings constructs from Achievement Goal Theory

Survey Item	Factor			Variance Explained
	Performance Approach	Work Avoid	Mastery Approach	
Doing better than the other students in this class on exams.	0.926			29.6%
Proving to my peers that I am a good student.	0.546			
Doing better than the other students in the class on assignments.	0.959			
Getting a better grade than other students in this class.	0.934			
Getting a passing grade with as little studying as possible.		0.847		23.5%
Getting through the course with the least amount of time and effort.		0.963		
Not having to work too hard in this class.		0.839		
Knowing more than I did previously about these course topics.			0.754	18.0%
Really understanding this course's material.			0.889	
Feeling satisfied that I got what I wanted from this course.			0.656	
Total Variance Explained				71.1%

Table 6 - Factor loadings for constructs from Expectancy-Value Theory and FTP
The table is split between two pages; the factor analysis was done with all five factors simultaneously.

	Factor			
Survey Item	**Expectancy**	**Connectedness**	**Perceptions of Future**	**Variance Explained**
I expect to do well in this engineering course.	0.741			15.3%
I am certain I can master the skills being taught in this engineering course.	0.809			
I believe I will receive an excellent grade in this engineering course.	0.951			
I am confident I can do an excellent job on the assignments in this engineering course.	0.909			
Considering the difficulty of this engineering course, the teacher, and my skills, I think I will do well in this engineering course.	0.829			
*I don't think much about the future.		0.783		10.8%
*I don't like to plan for the future.		0/801		
*It's not really important to have future goals for where one wants to be in five to ten years.		0.579		
*One shouldn't think too much about the future.		0.710		
*Planning for the future is a waste of time.		0.672		
I am confident about my choice of major.			0.618	10.3%
Engineering is the most rewarding future career I can imagine for myself.			0.849	
My interest in an engineering major outweighs any disadvantages I can think of.			0.823	
I want to be an engineer.			0.816	
Total Variance Explained				53.7%

Table 6, continued

Survey Item	Factor: Value	Factor: Instrumentality	Variance Explained
The most important thing in life is how one feels in the long run.	0.531		
It is more important to save for the future than to buy what one wants today.	0.581		
Long range goals are more important than short range goals.	0.784		9.0%
What happens in the long run is more important than how one feels right now.	0.802		
It is better to be considered a success at the end of one's life than to be considered a success today.	0.492		
I will use the information I learn in my engineering course in the other classes I will take in the future.		0.728	
I will use the information I learn in this engineering course in the future.		0.877	8.2%
What I learn in my engineering course will be important for my future occupational success.		0.691	
Total Variance Explained			53.7%

Table 7 - Factor loadings for constructs from Agency Beliefs

Survey Item	Factor: Science Agency Beliefs	Factor: Engineering Agency Beliefs	Variance Explained
Learning science will improve my career prospects.	0.634		
Science is helpful in my everyday life.	0.885		
Science has helped me see opportunities for positive change.	0.905		28.1%
Science has taught me how to take care of my health	0.627		
Learning science has made me more critical in general.	0.624		
Engineering can improve our society.		0.791	
Engineering will give me the tools and resources I need to make an impact		0.805	
Engineering can improve our quality of life.		0.921	29.3%
I see engineering all around me.		0.678	
Engineering allows me to think deeply about problems.		0.564	
Total Variance Explained			57.4%

Table 8 - Factor loadings for constructs from the "Big 5" Psychological Traits

	Factor					
Survey Item	Extraversion	Neuroticism	Agreeableness	Openness to Experience	Conscientiousness	**Variance Explained**
*Am quiet around strangers	0.867					
*Keep in the background	0.831					
Talk to a lot of different people at parties	0.659					12.1%
Am the life of the party	0.601					
*Don't talk a lot	0.798					
Have frequent mood swings		0.707				
Get irritated easily		0.696				
Get stressed out easily		0.648				11.2%
Change my mood a lot		0.801				
Get upset easily		0.800				
Have a soft heart			0.656			
Sympathize with others' feelings			0.896			
Am interested in people			0.512			9.8%
Feel others' emotions			0.790			
Make people feel at ease			0.458			
*Do not have a good imagination				0.719		
Have excellent ideas				0.784		8.8%
Have a vivid imagination				0.817		
Am full of ideas				0.494		
*Often forget to put things back in their proper place					0.741	
*Make a mess of things					0.724	8.4%
*Avoid my responsibilities					0.520	
*Leave my belongings around					0.734	
Total Variance Explained						50.3%

* indicates an item which was reverse coded

Table 9 - Factor loadings for constructs related to Math Identity

Survey Item	Factor			Variance Explained
	Performance /Competence	Recognition	Interest	
My parents see me as a math person.		0.775		20.5%
My instructors see me as a math person.		0.690		
My peers see me as a math person.		0.899		
I've had experiences in which I was recognized as a math person.		0.669		
Others ask me for help in math.		0.552		
I am interested in learning more about math.			0.802	15.4%
I enjoy learning math.			0.892	
I find fulfillment in doing math.			0.735	
I am confident that I can understand math in class.	0.893			23.8%
I am confident that I can understand math outside of class.	0.885			
I can do well on exams in math.	0.810			
I understand concepts I have studied in math.	0.721			
I can overcome setbacks in math.	0.445			
Total Variance Explained				59.7%

Table 10 - Factor loadings for constructs related to Engineering Identity

Survey Item	Factor			Variance Explained
	Performance/ Competence	Recognition	Interest	
I will feel like an engineer in the future		0.453		19.6%
I am interested in learning more about engineering.		0.844		
I enjoy learning engineering.		0.899		
I find fulfillment in doing engineering.		0.750		
My parents see me as an engineer.			0.744	15.9%
My instructors see me as an engineer.			0.847	
My peers see me as an engineer.			0.560	
I have had experiences in which I was recognized as an engineer.			0.451	
I am confident that I can understand engineering in class.	0.859			25.5%
I am confident that I can understand engineering outside of class.	0.942			
I can do well on exams in engineering.	0.855			
I understand concepts I have studied in engineering.	0.751			
Total Variance Explained				61.1%

To investigate the relationship between physics identity and the other attitudinal factors, I performed a linear regression testing for association between physics identity and associated factors. The model was first tested as a blockwise regression (i.e., inserting all factors as predictors) then using reverse elimination to remove non-significant predictors. At each iteration, the regressor with the highest non-significant p-value (closest to 1) was removed and the regression repeated. Significance values were corrected for multiple comparisons with a Holm-Bonferroni correction. Table 11 summarizes the regression estimates in the final model (empty rows signify non-significant regressors that were removed in the final model). VIF statistics (measuring

collinearity of regressors) for the final model were all below 2.0, suggesting the adjusted R^2 is not inflated and the regressors are not collinear.

Table 11 - Linear model of physics identity by attitudinal factors.
Bolded lines highlight significant positive associations. Italicized lines highlight significant negative associations. Lines which are empty were found to be not significant and were removed from the final model (i.e., the final model consists only of the factors with non-empty lines in the following table).

	Estimate	**Beta**	**Std. Error**	**Signif.**
(intercept)	0.010	0.000	0.188	
Belongingness	**0.236**	**0.182**	**0.029**	*******
Performance Approach				
Mastery Approach				
Work Avoid				
Expectancy	**0.098**	**0.076**	**0.026**	******
Connectedness	*-0.065*	*-0.060*	*0.018*	**
Instrumentality	*-0.094*	*-0.060*	*0.031*	*
Perceptions of Future	**0.116**	**0.103**	**0.025**	*******
Value				
Grit: Persistence of Effort				
Grit: Consistency of Interest				
Engineering Agency Beliefs	**0.115**	**0.064**	**0.038**	*****
Science Agency Beliefs	**0.210**	**0.163**	**0.025**	*******
Neuroticism				
Extroversion				
Agreeableness				
Conscientiousness				
Openness	**0.091**	**0.083**	**0.019**	*******
Math Identity	**0.176**	**0.138**	**0.023**	*******
*** p < 0.001, ** p < 0.01, * p < 0.05				
Multiple R^2: 0.253			Adjusted R^2: 0.250	

The Physics Identity factor was significantly and positively associated to Belongingness ($p<0.001$), Expectancy ($p<0.01$), Perceptions of Future ($p<0.001$), Engineering Agency Beliefs ($p<0.05$), Science Agency Beliefs ($p<0.001$), Openness ($p<0.001$), and Math Identity ($p<0.001$) factors, meaning that students who indicated high scores on these factors also had high scores, on average, in their Physics Identity. The largest effects (in both raw estimate and standardized beta) were Belongingness, Science Agency Beliefs, and Math Identity; for these, the difference between the highest possible

score (on a scale from 0 to 6) and the lowest score amount to a difference of 1.0-1.3 in the Physics Identity outcome for each.

Physics Identity was significantly and *negatively* predicted by Connectedness ($p<0.01$) and Instrumentality ($p < 0.05$), meaning students indicated higher scores on this factor had *lower* scores on their Physics Identity measure. For the negative predictors, the difference between having the highest possible score (on a scale from 0 to 6) and the lowest possible score amounted to a predicted difference of approximately 0.39 or 0.57 in the Physics Identity measure.

Overall, the model explains 25% of the variance in the measured physics identity scores, a moderate effect.

Discussion and Interpretation of the Primary Model

Math Identity has been previously studied in relation to Physics Identity development(Godwin, Potvin, Hazari, et al., 2013), so its presence in this regression is unsurprising. Notably, it remains one of the strongest effects, but is smaller than either Belongingness or Science Agency Beliefs (both of which were themselves significantly associated with Math Identity).

The Agency Beliefs factors describe a student's perception of the importance of science or engineering in their lives in a variety of positive ways. At this stage of their education, most students have not had significant exposure to many experiences which might be described as related to engineering or engineering contexts, as opposed to science (or physics in particular). For many students, science is a more familiar and commonly seen context in their lives, while engineering is perhaps less contextualized in

the present, but is something for the future (perhaps in their perceived future). It is not surprising that beliefs about the ability of science to have a positive impact on the world would be more closely associated with identifying as a physics person than similar beliefs about the same ability of engineering, which may be seen as somewhat separate from the sciences. Both have been used to predict choice of career, but primarily as simultaneous predictors independent of physics identity (Godwin et al., 2016). The strong association found here suggests that a more complex interaction of agency beliefs and identity may better describe how these factors interact and are associated with engineering identity or engineering career choice.

On the surface, that Belongingness (in engineering) strongly predicts Physics Identity can be understood as a reflection of the fact that Physics Recognition beliefs are the most important of the three sub-constructs in the measure of Physics Identity. Recognition from peers and teachers is important to identity development (Pittman & Richmond, 2008) and is associated with feeling that one belongs in their community. However, if Belongingness is predicted with the three sub-constructs of Physics Identity as regressors, then Performance/Competence ($p<0.001$) and Interest ($p<0.001$) are significantly and positively associated with Belongingness, and Recognition is not. See Table 12 for details on these associations.

Table 12 - Linear regression predicting Belongingness with Physics Identity

	Estimate	**Beta**	**Std. Error**	**Signif.**
(intercept)	3.441	0.000	0.059	***
Performance / Competence	**0.253**	**0.341**	**0.021**	*******
Recognition	0.014	0.021	0.017	
Interest	**0.053**	**0.079**	**0.017**	******

*** p < 0.001, ** p < 0.01, * p < 0.05 Adj. R^2: 0.1694

One explanation of this relationship is that Belongingness in an engineering program depends on the acceptance of one feels from their peers, who see classmates who are highly capable as being valuable additions to the classroom. So, while being recognized by peers is important to feeling like one belongs, it appears that this facet is less important in the face of perceived judgment from those same peers on one's proficiency and competence. This dependence on competence may be a consequence of studying feelings of belongingness in an engineering context while associating them with physics identity subconstructs; technical ability and interest in physics may be seen as useful or important skills for many engineering students because the physics content is central to much of their engineering education.

Expectancy describes how the student sees their future success in this class, whether or not they thing they will succeed. Like the Belongingness construct, I argue that this association with Physics identity can be best understood in terms of how it relates to the student's Physics Performance/Competence beliefs. If a student feels they can do well in physics, then their likelihood of success in their introductory engineering course (which has strong connections and overlaps with physics content in many areas) is much higher, and so a belief in one would be associated with belief in the other. Of course, the associations present in this correlational regression analysis does not imply causality. I hypothesize that in fact the causality may be actually reversed, and that high Performance/Competence beliefs lead to higher expectations of success. As a quasi-trait, Physics Identity is stable over medium time periods (Potvin & Hazari, 2013). On the other hand, Expectancy is a judgment about expectations of success for a very particular task (in this case, succeeding in an engineering class). Therefore, I expect that incoming

Physics Identity beliefs inform their expectations of success in an environment where physics competence is relevant. High levels of Expectancy can certainly influence future Physics Identity as predictions of success are either validated, thereby increasing a student's belief in their ability to do physics tasks, or repudiated, and their beliefs in their own ability are diminished. However, as an interaction between a longstanding identity and a short-timescale belief measured at the same point in time, identity beliefs may inform the expectations of success in the moment, but this would need to be further studied in another analysis.

Of the "Big 5" Personality Traits, only Openness to Experience was found to be a significant predictor of Physics Identity, with a small effect size. Facets of Openness to Experience include imagination, intellectual curiosity, and a willingness to experiment, all of which are traits and behaviors which may be highly valued and promoted in the framing of the physics community so its presence as a significant predictor is perhaps not unexpected. Prior research has shown an association between intrinsic motivations (such as "inventing new things" and "developing new knowledge and skills") with physics identity (Hazari et al., 2010), and that relationship repeats here in a similar fashion. While the intrinsic motivations in that prior work were oriented in the future, Openness to Experience describes the current affective state of the student, and so may hypothetically be a precursor trait to both identity and intrinsic motivations.

A coherent explanation for the significant negative predictors—Instrumentality and Connectedness—is more nuanced. High scores in each of these factors indicate that the student has some strong sense of a specific future for themselves. Connectedness speaks about personal preference for making plans for the future, having goals, and

thinking about what they want to do, and Instrumentality describes how important they see their current class to this future. While these actions can be positive, when asking a student in an engineering class these strong plans are most likely for a concrete path or specific goals in engineering. While Physics Identity is a strong predictor of choice and persistence in engineering, these associations are all founded on measurements of early-college identity, before most students have had many authentic engineering experiences or is deeply involved with the culture of engineering. Less research has focused on the evolution of this association through a student's college experiences, but (Zavala & Dominguez, 2016) have shown a marked decrease in students' perceptions of the relevance of Math and Physics to their engineering education and careers by their third semester of college engineering studies. Thus, while many students may have unclear ideas of exactly what it means to do engineering, those who are disposed towards planning for the future (i.e., those with high Connectedness and Instrumentality scores) may have a clearer picture, and so associate themselves less with physics, and more with engineering. The students with high scores on these factors may be more fully committed to an engineering-related future, where their focus on a specific future and their courses relevance for that future narrows their identification with identities that are not perfectly aligned with how they see themselves in the future.

In contrast to this, however, there is a positive association between Perceptions of Future, a factor describing how students imagine a positive future for themselves through a career in engineering, and Physics Identity, which appears opposite to the explanation above. After all, if a student has a positive concept of a future in engineering, isn't that the same as making plans for a particular future? I argue no; for students without many

authentic engineering experiences in their history, their concept of "a future in engineering" at the beginning of college may be more nebulously formed than for a student who tends to make specific plans for the future, and thus may be more informed of the realities of what engineering entails. In this case, the student might fall back on other related identities (e.g. math, science, and physics, which have been shown in prior research to be strongly associated with identifying as an engineer) to shape their idea of what a future in engineering would look like and entail.

In summary, I found the largest associations with Physics Identity to be feelings of belongingness in engineering and one's engineering class (Belongingness), beliefs in the ability for science to have a positive effect on the world (Science Agency Beliefs), and seeing oneself as a math person (Math Identity). Secondary, weaker associations were found with how students view the future and its relationship to their current educational trajectory (Perceptions of Future and Expectancy), along with a sense of imagination and intellectual curiosity (Openness), plus a belief in the ability of specifically engineering to have a positive effect on the world (Engineering Agency Beliefs). The largest effects are all things which have been previously well-studied in tandem with Physics Identity, though now in a combined form, and with added nuance.

The inclusion of the future-pointing affective constructs from Future Time Perspective add an additional dimension to the discussion of physics identity, especially in the context of Hazari's quantitative framework (2010). Namely, not only is physics identity a time-variant quantity, which was previously understood, but also that student perceptions of the future (in the general sense, not strictly in terms of the named attitudinal factor) are associated to their identity in the present. While one's identity is

shaped by one's experiences, these experiences are colored by expectations and hopes about what the future will look like, and whether that future contains a congruous identity as the one being formed in the present.

Using these measures of students' interest, I expanded the primary regression analysis to consider interest in various majors. Models were considered in parallel; each engineering major interest was incorporated into the model separately, for a total of 17 additional models, which are summarized in Table 13, and the p-values of these associations were manually corrected with a Holm-Bonferroni factor to account for the fact that so many hypotheses were tested in parallel.

Table 13 - Summarized expanded models
Statistical significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, after being corrected for multiple comparisons. Blank cells in the table represent terms which were not statistically significant. For Major Interest, + represents a positive estimate, while – indicates a negative estimate. Bolded Adjusted R^2 values indicate an increase of at least 1% (absolute) over the original model.

	Belongingness	**Expectancy**	**Connectedness (-)**	**Instrumentality (-)**	**Perceptions of Future**	**Engineering Agency Beliefs**	**Science Agency Beliefs**	**Openness to Experience**	**Math Identity**	**Interest in MAJOR**	**Adj. R^2**
Primary	***	**	**	*	***	*	***	***	***		0.250
AAE	***	***	*		***		***	**	***	+***	0.289
ABE_BSE	***	***	**	*	***	*	***	***	***		0.247
BE_BME	***	***	**	*	***	*	***	***	***	-***	0.254
CME	***	***	**	*	***		***	***	***		0.244
CVL	***	***	**	*	***	*	***	***	***	+*	0.251
CE_CSE	***	***	**	*	***	*	***	***	***		0.249
CON	***	***	**	**	***	*	***	***	***	+***	0.253
EE	***	***	**	*	***	*	***	***	***	+***	0.258
EP	***	**		*	***	*	***	***	***	+***	0.351
EEE	***	***	**	*	***	*	***	***	***		0.245
IND	***	***	**	*	***	*	***	***	***	+**	0.255
IT	***	***	**	*	***	*	***	***	***		0.248
MSE	***	***	**	*	***	*	***	***	***	+***	0.252
ME	***	**	**	*	***		***	**	***	+***	0.281
MIE	***	***	***	**	***	*o*	***	***	***	+***	0.262
NUKE	***	**	**	*	***	*	***	***	***	+***	0.263
O-STEM	***	***	***	**	***	*	***	***	***	+**	0.251

Results and Discussion of the Secondary Models

Of the majors probed in Q14, student interests in the following majors were found to be significantly and positively associated with Physics Identity:

- (AAE) Aero / Astronautical Engineering ($p<0.001$)
- (CVL) Civil Engineering ($p<0.05$)
- (CON) Construction Management Engineering ($p<0.001$)
- (EE_ECE) Electrical Engineering / Electrical and Computer Engineering ($p<0.001$)
- (EP) Engineering Physics ($p<0.001$)
- (IND) Industrial Engineering ($p<0.01$)
- (MSE) Materials Engineering / Material Science and Engineering ($p < 0.01$)
- (ME) Mechanical Engineering ($p<0.001$)
- (MIE) Multidisciplinary / Interdisciplinary Engineering ($p<0.001$)
- (NUKE) Nuclear Engineering ($p<0.001$)
- (O-STEM) Other STEM-related degree ($p<0.05$)

On the other hand, interest in Bioengineering / Biomedical Engineering ($p<0.001$) was significantly and *negatively* associated with Physics Identity.

The majors for which student interest did not include a statistically significant effect were Agricultural and Biological / Biosystems Engineering, Chemical Engineering, Computer Engineering / Computer Science Engineering, Ecological and Environmental Engineering, and Information Technology.

Adding "interest in pursuing this major" to the regression tended to affect the resulting adjusted model in one of a few broad ways.

A. For the first group, defined by a statistically significant association between interest and physics identity coupled with a moderate to large increase in

explained variance in the model. The majors in this group are aerospace engineering, mechanical engineering, nuclear engineering, engineering physics, and multidisciplinary/interdisciplinary engineering.

B. For the second group, there are statistically significant associations, but with a smaller effect; the added interest term increased the variance explained by only a small amount compared to the original model. The engineering majors in this group are civil engineering, construction management engineering, industrial engineering, electrical engineering, and material science engineering. This group also included interest in "other STEM-related degree".

C. The third group consists of all the other majors which showed no statistically significant positive association with physics identity and showed no improvement in the variance explained. The majors in this group are bioengineering / biomedical engineering, which actually showed a statistically significant negative association with physics identity, agricultural / biosystems engineering, chemical engineering, environmental/ecological engineering, computer engineering, and information technology.

Differences with the Primary Model

The original model explained 25% of the variance in Physics Identity scores. Adding major interest to the model improved this value for a handful of majors (i.e., those in the Group A) by at least 1% more, up to 10% for engineering physics.

I found no statistically significant difference in the regression coefficients for any of the original factors between the primary model and the models with an added Interest

term, even though the calculated significance value for the new estimate may indicate that one of the factors is no longer statistically significant. For example, when adding Interest in Aerospace Engineering to the model, after correcting for multiple comparisons the terms for Instrumentality and Engineering Agency Beliefs are no longer statistically significant. Prior to the additional term, they were each significant at the $p<0.05$ level; after, Instrumentality had a p-value of 0.052, and Engineering Agency Beliefs had a p-value of 0.068. However, when investigating whether there was a statistically significant difference between the estimates for these (and all the other) factors had changed between models, I was unable to reject the null hypothesis that the estimates were the same between models ($p>0.10$ for all comparisons).

In summary, the model was improved incrementally, but not in a significant step up. No significant differences were seen in the associations between the affective factors in the primary model after including interest in particular engineering majors. While this may be a result of the differences being too small to distinguish, all differences between effect sizes were less than 0.035, which is a small difference. Therefore, the answer to the second research question of whether there is a difference in the associations between physics identity and related factors after controlling for interest in an engineering major appears to be no.

Implications and Directions for Future Work

Because introductory physics classes often serve many students to provide a background in physics knowledge, understanding how students of various majors may see themselves as related (or not) to physics could help improve their experiences,

increasing both student affect and performance. The goal of increasing physics identity among engineers is not meant to pull them away from other interests, but rather to tap into the benefits related with such an identity in order to make their interaction with physics more rewarding, both in terms of increased knowledge gains and increased affect.

It is worth noting that using physics identity as a proxy indicator for increased interest, persistence, and performance is not always appropriate. The negative associations between physics identity and the future time perspective constructs of Connectedness and Instrumentality suggest that engineering students with a strong and specific sense of their future tend to identify less with physics, even though these students are more likely to have the motivation and interest to persist in their engineering programs (Kirn et al., 2014). The hypothesized decoupling of physics identity and engineering identity is further investigated in Chapter 5.

Primarily, these results suggest the need for a more complex and flexible model of physics identity as applied to engineering students, particularly through the model of engineering identity. The structural model of engineering identity proposed by Godwin et al. (2016) is a simple and effective model, but may need to be revised and expanded to include other conceptions of what it means to be an engineering. The expansion may include other domain identities in addition to physics (e.g., chemistry, biology, computer science), as well as additional affective constructs from future time perspective theory. The goal of such an expansion is not to reconstruct the identity framework itself, but rather to establish where and how connections exist between the identity framework and other frameworks that have been studied in education and psychology research.

The lack of association between interest in broad categories of engineering and physics identity, a construct underpinning the construction of engineering identity in quantitative models and theory (Godwin, 2016; Godwin et al., 2016; Katehi, 2009), suggests missing explanatory variables, and that the engineering identity model is not accurately capturing what it means to identify as an engineering for all engineering students. Further work investigating this relationship and future extensions to the model could extend the analysis in this chapter to look at Engineering Identity as a predicted construct, as well as examine additional domain identities that may be more relevant to particular engineering disciplines.

Limitations of this Study

The schools sampled for student data were not randomly selected but were chosen because they were the four universities of the members of the research collaboration. All four schools are large public research institutions (three are R1, one is R2). Their populations are not fully representative of the U.S. engineering student population, the college student population, or overall population of the country. Within the participating schools, the survey had high response rates, over 70% of the population of interest at the four schools in the Fall 2015 cohort. Thus, these results are well-representative of the schools from which they are drawn, but nevertheless, should not be assumed to be fully generalizable to all engineers or engineering programs.

Though the use of Interest in a Major as a proxy for Member of a Major was justified with concurrent criterion-related validity testing, it is still a potential limitation on the interpretation of the results. Discussion and implications drawn about, e.g.

“mechanical engineers”, are actually only able to say something about “people who expressed high interest in mechanical engineering”, which broadens the scope to include more than just students who are in that major. This broadened scope may be a strength and a weakness of the analysis.

The regression analyses are correlational in nature, and the study design was a cross-sectional, non-experimental one. Combined with a lack of time-series or longitudinal data, these limitations prevent definitive conclusions from being drawn about the causality of these associations.

Chapter IV: Topological Mapping of Student Affective Factors

Introduction

In this chapter, I create a map[6] of the space of affective constructs previously discussed in Chapters 1 and 3. I begin by introducing the theoretical motivations for this new analytical methodology, then describe the new technique of topological data analysis (TDA) and how it will be applied to the research data.

I finish by discussing several results that can be gathered from the resulting map, including the presence of a large "normative" group defined by the data as characterizing the most popular set of beliefs, as well as a limited number of moderately-populated deviations from this profile. I also describe differences between the normative group and the students who were assigned to no group in terms of traditional demographic markers.

In this chapter I investigate the following research questions:

1. How are students distributed in the space of affective beliefs?
2. What demographic differences exist between students holding normative beliefs and those with non-normative beliefs?

Answering these questions can help deepen and extend the understanding of how various attitudinal factors relate to each other that was started in Chapter 3. Further, answering the second question will help clarify how related the concepts of normativity

[6] In this context, a **map** is a two-dimensional representation of a high-dimensional set of data that encodes several levels of relational information between the data. A formal definition of these maps and their construction detailed in this chapter.

in terms of demographics and attitudes are to each other, i.e., if the patterns of concentrations in beliefs are demographic-dependent.

Background

Challenges of Intersectionality in Quantitative Research

To date, much of the quantitative research on diversity in STEM has first binned students by certain demographic categories (e.g., male or female, etc.) and only then examined differences in students' attitudes or beliefs. Such an approach is limited in several ways (Pawley & Slaton, 2015). First, students at the intersections of multiple underrepresented categories often represent a small proportion of any sample of students. These small numbers can result in several problems to analyze quantitatively which results in these students being diminished in importance. Small groups of students can be viewed as "anomalies" not representative of the whole and, hence, dismissed. Additionally, the statistical power to detect differences or understand students at multiple intersections is difficult or impossible to achieve with small datasets. Finally, small numbers of students can be disaggregated from the larger dataset in ways that risk the re-identification of participants and make their responses non-anonymous, which has ethical implications (including violations of standard IRB protocols).

The second issue with quantitative research on diversity is that many statistical techniques rely on various parametric or non-parametric data assumptions (including normality, homoscedasticity, etc.) and often use group averages to compare between groups or minimize the error of models. This approach can result in findings that generalize for fixed demographic categories. As a result, many studies make claims for

"all" women or "all" women of color without recognizing the proper variance and systematic effects within groups and so lose an understanding of the nuance of individuals' experiences. These issues limit the power and interpretability of approaches which bin individuals by researcher-defined categories *a priori* as a way to understand how a diverse population of students navigate engineering.

Another Approach to Understanding Student Diversity: Cluster Analysis

One possible approach that handles the issue of binning students is provided by cluster analysis, an alternative, quantitative method of grouping students, which can use criteria other than factors such as demographics. At its core, cluster analysis uses a similarity measure to determine which data points (e.g., students, or something else) are "close" to each other, which ones are far away, and then grouping the close points together. Groups which are close to each other and far from other things are called clusters.

Groupings of the data are therefore determined by the variance in the data (and choice of clustering algorithm), not by *a priori* imposition. An example of an external grouping imposed on the data would be organizing students by gender. One might argue that such a grouping is "determined by the data", as each student provides their own gender and this information is part of the data set, but the categories to which the students can belong are predetermined by the researchers (which often have unstated and unexamined value judgments present in the choice of categories). Cluster analysis seeks to discover potentially new categories that do not yet have a label and attach one, so that a student could, in addition to being described with a particular gender identity or

ethnicity, also be described as a "member of group A" which carries its own information connotations, emergent from the data.

Topological Data Analysis as a Means of Clustering

Though there are several methods which are characterized under the framework of Topological Data Analysis (Carlsson, 2009), the current study focuses on the so-called "Mapper" algorithm (Singh, Mémoli, & Carlsson, 2007) as the method of choice. Originally designed as a way of describing the topology of point cloud data for image processing, I adapted it to use with human subjects/educational data. Some technical features of Mapper have been modified for ease of implementation in the programming language **R**, though the eventual result is identical.

By "point cloud data" I mean that each data point is represented by a point in some vector space, with numerical values for each dimension. The dimensions can represent, e.g., individual questions, or factors constructed out of multiple questions with a factor analysis. In the current work, these numbers represent the responses of students that are used to cluster them together based on their similarity or closeness.

Topological Data Analysis in InIce

I turned to topological data analysis because I wanted to find a way to conduct a quantitative analysis that respected the intersectional identities of the participants, and made no presuppositions about the sort of structures I would find. I hypothesized there would be one large group and a handful of smaller subgroups of similar density, separated in the space of beliefs; subsequent analysis showed the initial hypothesis was only partly correct, but had a more traditional cluster analysis been used and forced to

produce multiple clusters through choice of parameter, I could have obtained a result which suggested such a pattern existed, even if such a hypothesized distribution was not in fact the best description of the underlying structure. Topological data analysis allows the structure (and subsequently, the number of groups) to emerge from the data, which protects against errors of this sort.

Methodology

In this section, I outline the process by which I selected the subset of affective constructs to map out. I then introduce the Mapper algorithm, a form of topological data analysis which reduces high-dimensional data to a two-dimensional representation showing the how the data are distributed in relation to themselves. From there, I discuss the steps taken to prepare the student survey data for mapping and the researcher choices involved.

Description of InIce Survey

A pilot survey was deployed in Spring 2015 at three of the four institutions and had 537 responses. The results of the pilot were used to confirm the factor structure of the questions and select the questions which best-illuminated the factor in question. The final version of the survey was deployed at all four institutions in the Fall of 2015, and had a total of 2916 responses, distributed similarly to the relative sizes of the engineering student populations at each school. The survey was given to students in each institution's

analogous introductory engineering course, intending to capture a broad cross-section of incoming freshman engineers[7].

Attitudinal Factors

The factor analyses produced 26 theoretical constructs underlying the items analyzed. These constructs were drawn from a variety of theoretical frameworks, including achievement goal theory, expectancy-value theory, future time perspective, grit, the "Big Five" psychological traits, and identity. For more information on these constructs, including which questions loaded into each factor, see page 31.

One difficulty of analyzing high-dimensional data is the so-called "curse of dimensionality" (Bellman, 1957), which describes how, as the number of dimensions increases, the difference in distances between different pairs of points in the sample get smaller, and distance functions become less useful in distinguishing between points. A rule of thumb when trying to detect clusters in d dimensions is that a sample size on the order of $N \sim 2^d$ is required (Formann, 1984). With a sample size of 2916, this corresponds to a dimension of approximately 11.5, or between 11 and 12. To maximize the ability of Mapper to detect interesting features in the data, the number of factors used in the map (and thus the dimensionality of the space) was reduced following a multi-criterion analysis summarized in a decision matrix (see Table 14). Each factor was given a score on four dimensions, determined by how that factor related to the others on that

[7] Included in the survey was a question about which year the student was, and whether they were a transfer student. This was most relevant for the students from FIU, which has a relatively higher fraction of transfer students and students switching into engineering after several years in college.

dimension. To maximize the ability of the map to detect variations in the structure of the data, including subgroups, factor variance and uniqueness were highly weighted in the overall decision. A factor with high variance is more likely to spread the students apart in that dimension, making it easier to detect differences. Uniqueness describes a lack of correlation between that factor and others, in order to maximize the orthogonality of the chosen factors. Figure 3 shows a representation of the correlation matrix between the factors considered in the decision matrix. The figure shows three main groups of correlated factors: the physics identity subconstructs with themselves, the math identity subconstructs with themselves, and the engineering identity subconstructs along with belongingness and some of the future time perspective theory constructs. Factors with a high uniqueness score in the decision matrix had low average correlations to other factors. The factor loadings dimension measured the strength of loading, how closely aligned the questions which formed the factor were related to each other. This dimension was given relatively low priority because the average factor loadings for the factors were on average high (mean 0.7461, ranging from 0.6377 to 0.8824). Theoretical interest is a parameter chosen by the research team, corresponding roughly to how interested the researchers were in including that factor in the final map. Factors relating to identity, future time perspective, and belongingness were ranked highly here, as the results of prior work showing their interrelated importance for engineering students, but this dimension was weighted less strongly than either factor variance or factor uniqueness.

Figure 3 – Correlation plot between factors
Dark blue circles along the diagonal indicate self-correlation (by definition, a factor is always perfectly self-correlated). Factors names were arranged to place more-correlated factors adjacent to each other.

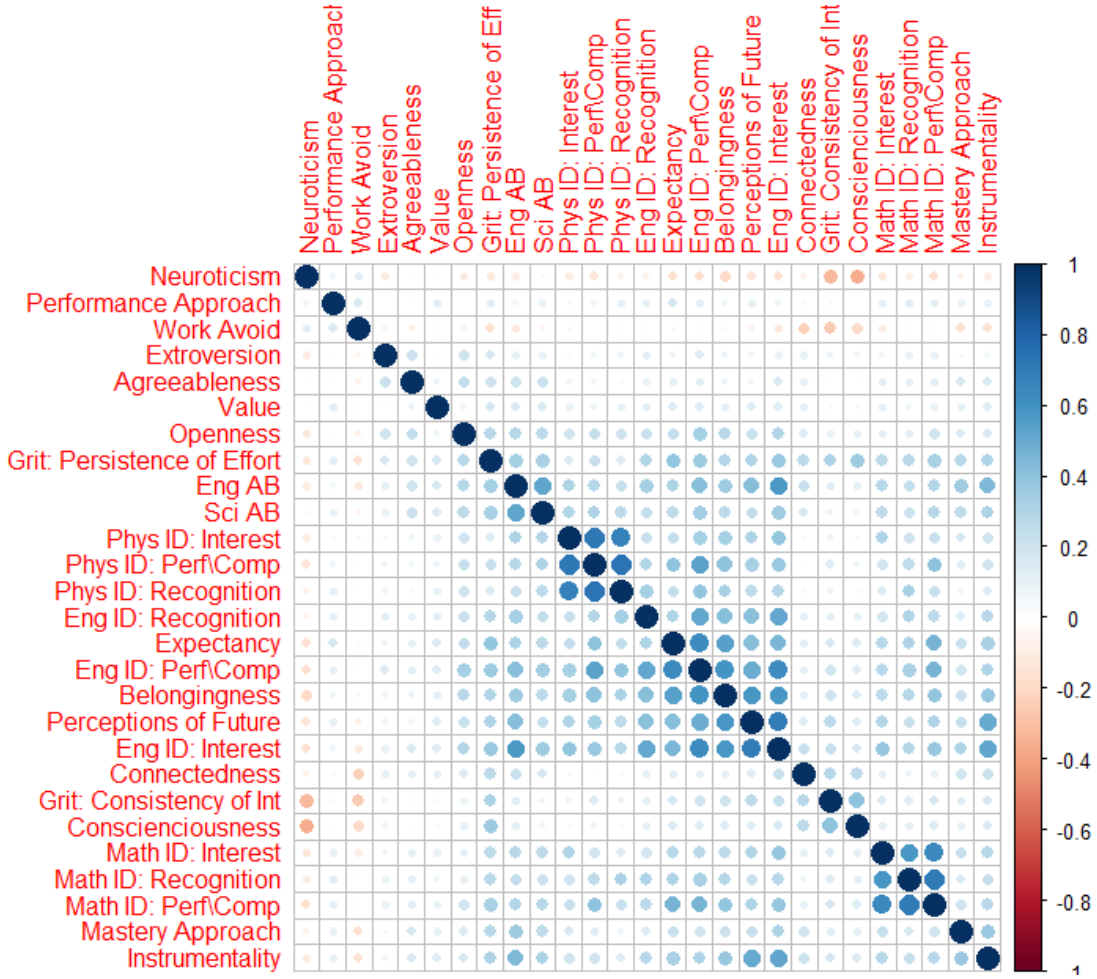


Table 14 - Decision matrix to select a subset of factors

Each factor was scored from one to three on four dimensions, according to how that factor related to the others on that dimension. In each dimension, higher values are better. Variance indicates the relative variance of that factor among the students. Factor loadings indicate the relative average loading of questions in that factor, according to the factor analysis which identified it. Theoretical interest is a parameter decided by the research group, corresponding roughly to how interested the researchers were in seeing that particular factor in the final map. Uniqueness indicates how uncorrelated the factor tended to be from the other factors; factors which were highly correlated with other factors received lower scores. Factors with weighted scores of 100 or greater (bolded) were in the top half of scores and were selected for use in the mapping.

	Variance	Factor Loadings	Theoretical Interest	Uniqueness	Weighted Score
Weights	15	5	10	15	
Factors					
Belongingness	2	3	3	2	105
Performance Approach	2	3	1	3	100
Mastery Approach	1	2	1	3	80
Work Avoidance	3	3	1	3	115
Expectancy	2	3	1	2	85
Connectedness	2	2	3	3	115
Instrumentality	1	2	3	3	100
Value	3	1	3	3	125
Perceptions of Future	2	2	3	3	115
Grit: Persistence of Effort	2	1	2	2	85
Grit: Consistency of Interest	3	1	2	2	100
Engineering Identity: Performance / Competence	2	3	3	2	105
Engineering Identity: Recognition	2	1	3	2	95
Engineering Identity: Interest	1	2	3	2	85
Engineering Agency Beliefs	1	2	2	1	60
Science Agency Beliefs	2	2	2	1	75
Neuroticism	3	2	1	3	110
Extraversion	3	2	1	3	110
Agreeableness	2	1	1	3	90
Conscientiousness	2	1	1	3	90
Openness to Experience	2	2	1	3	95
Physics Identity: Performance / Competence	2	2	2	1	75
Physics Identity: Recognition	3	2	3	1	100
Physics Identity: Interest	1	2	3	1	70
Math Identity: Recognition	2	2	3	1	85
Math Identity: Performance / Competence	2	2	2	1	75
Math Identity: Interest	1	3	3	1	75

The factors with the highest overall scores in the decision matrix were selected to form the basis of the map. These factors were, in descending order of overall score, Value, Work Avoidance, Connectedness, Perceptions of Future, Neuroticism, Extraversion, Belongingness, Performance Approach, Instrumentality, Grit: Consistency of Interest, Engineering Identity: Performance / Competence beliefs, Engineering Identity: Recognition beliefs, and Physics Identity: Recognition beliefs. Because four of these factors were tied for the same score, I decided that it was better to keep the number of factors at thirteen, slightly above the number predicted by the rule of thumb for my sample size, rather than removing all four or making an arbitrary choice between the four after designing and implementing a design matrix to make such a decision with as much objectivity as possible. As a result, I expect that the resulting space will be slightly sparse, as opposed to being overcrowded, but Mapper's ability to handle underpopulation is superior to its ability to handle overpopulation when the dimensions are highly discretized, as is the case with the factors in the InIce survey. I discuss this aspect of the algorithm in greater detail below, on page 77.

Some, but not all, of the factors found to be significant predictors of physics identity in Chapter 3 appear in this list; the choice of factors used to create the map was made independently of the results of that analysis, though the relationships found were considered when determining the "theoretical interest" of each factor. In other words, the goal of making the Mapper map is not to create a picture of the space of attitudes which are all related to physics identity, and using too many of those factors at once would, in fact, reduce the ability of Mapper to resolve differences between groups of students because of the increased collinearity of the basis vectors.

Survey Demographics and Self-Identification

The demographic questions in the survey (Q15-Q22, see Appendix, page 155) were designed to be as inclusive as possible and allow a broad range of self-identification. Because I perform the cluster analysis on data while initially ignoring the demographic information, this facilitates more freedom in how demographic questions are asked. For example, rather than asking a binary gender question, students were provided a range of options and allowed to combine them in whatever fashion accurately reflected their gender identity. Though the vast majority of students (97.4% of responses) responded with one of the two traditional binary options ("Female" or "Male") exclusively, nevertheless 70 (2.6% of responses) students responded in some other fashion, with a total of thirteen other unique combinations of answers.

One issue that arises when increasing the number of categories to which someone can subscribe is a fracturing of the measurement of the population. When groups become highly specified according to several factors, the number of respondents which match these factors exactly can become very small, which threatens classic quantitative analytic techniques that rely on having a large enough N to have acceptable statistical power. The only available solutions included either collapsing categories into a single "Other" category, or throwing out those responses entirely to concentrate on the categories with sufficiently large representation. The first option is distasteful for several reasons, including clearly "Othering" these individuals (Jackson II & Hogg, 2010), and because it collapses the variance in the sample that previously existed because of those responses and by treating them as an indistinguishable category. The second option means the

voices of those who responded in a particular category are formally ignored from analysis, which also reduces sample variance and representativeness.

For compatibility with past work, one would prefer new data collection to be "backward compatible": re-interpretable in a way that is as consistent as possible with previous approaches (even if those approaches have flaws in a modern perspective). Such compatibility would allow us to compare current information and results more easily with the past. However, simply maintaining compatibility is not a sufficient reason to continue poor practices with well-known problems. Survey items can be expanded in such a way that it is a natural extension of previous forms. Doing so allows the researcher to collapse back into previous iterations and thereby be comparable to old data sets. As an example, consider a question which includes 6 options for gender identity, along with another fill-in-the-blank option, where students can select all options which apply to them. If needed, results to this question can be returned to the classic "male/female/other" paradigm by taking every response which marked "male" but not "female", the responses which marked "female" but not "male", and the responses which marked something but didn't include either "male" or "female". A question about race/ethnicity which includes "select all that apply" can be returned to the single selection version by grouping everyone who responded with more than one answer into the NSF category "two or more races/ethnicities".

Requirements to perform TDA using Mapper

Mapper was originally constructed for numerical data (Singh et al., 2007). However, the algorithm in fact only requires a metric space, like most cluster analysis

algorithms. If a coherent definition of pairwise distance between every point can be constructed, then Mapper can construct a map of the topology of that space, but interpreting the resulting map may be more challenging than if the data were embedded in a high-dimensional vector space. A possible alternative distance function involves using the correlation between two sets of responses, which would allow a mixture of numerical and categorical responses to be used in the mapping. Formally, whichever function is used to calculate distances must satisfy certain criteria. A distance function on a given set of points, M, is a function $d: M\ x\ M \rightarrow \mathbb{R}$ that satisfies the following conditions:

1. Non-negative: $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$
2. Symmetric: $d(x, y) = d(y, x)$
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for $\forall y$.

Because Mapper requires a metric space, student responses which are missing values in the dimensions under consideration pose a barrier to analysis. Because each data point must exist somewhere in the space and have a measurable distance to each other point in order to be clustered, the algorithm requires there be no missing values. One approach to address the issue of missing values in the data is to imputing the missing values using a maximum-likelihood estimate (Little & Rubin, 2014), and then analyzing the complete data set. Imputation estimates what a student's response to a question would be if the question had been answered by analyzing the distributions of their other answers, and comparing them to the distributions of responses to those questions across all responses. Missing values are then estimated based on the distribution of how students with similar response patterns on the non-missing questions answered the missing item.

The single response with the maximum likelihood according to this estimate is then selected and filled in. Imputation algorithms iteratively fill the missing values from the questions with the fewest missing responses to most. In addition to facilitating the current analysis, imputation is a best-practices method that also ensures that the distributions in the data are not skewed based on systematic missingness (Little & Rubin, 2014).

In addition to the vector of numbers representing a student's position in this point cloud, *Mapper* requires another number for every data point, a one-dimensional function called a **filter function**. This filter function is used during the iterative clustering process to chunk the data into smaller pieces for analysis, and forms the basis of the shape of the resulting **map**. Singh et al.(2007) define the filter function for a space X as a continuous map $f: X \to Z$ to a parameter space Z which is equipped with a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ for a finite indexing set A, and notes that since f is continuous, the set of $f^{-1}(U_\alpha)$ form an open covering of X, defined as $\overline{\mathcal{U}}$.

In other words, the filter function assigns to each data point a real number in a continuous fashion, which will be later used to iteratively cluster data with similar filter values. One example such function would be a local density estimate[8]. The range of values for the filter function is the broken into a number of overlapping subsets. For example, if the filter (the parameter space Z) ranged from $U_1[0,1)$, it could have three subsets (coverings U_α), $U_1 = [0,0.5], U_2 = [0.25,0.75]$, $U_3 = [0.5,1)$ which together

[8] While the density of the space of the point cloud is in actuality a series of delta functions centered at each point, in the assumption that the data was sampled from an underlying continuous distribution function the density at a point can be estimated using one of several techniques, including maximum likelihood parameter estimation, or non-parametric k-nearest neighbors density estimation.

span the entirety of Z. The coverings of X are the sets of points which are assigned to each range of those filter values. $f^{-1}(U_1) =$ all points in X which were assigned a filter value by f that is in the range [0,0.5]. Notice that if the $\{U_\alpha\}$ overlap, the corresponding sets in $\bar{\mathcal{U}}$ will likewise be overlapping. That is, the same point in X will be a member of multiple sets in $\bar{\mathcal{U}}$.

Choosing an appropriate filter function is key to maximizing the utility of the algorithm, because different functions will result in different maps from the same data that highlight different structures, in the same way that a cylinder looks different if projected from the side (i.e. so it becomes a rectangle) versus the top (so it becomes a circle). Depending on the complexity of the underlying topology, certain filter functions may reveal different aspects of the data. The choice of covering, including the number of covers and the amount of overlap between covers, is another important researcher-driven choice which can affect the shape of the resulting map.

The last requirement to perform TDA using Mapper is a choice of ϵ, which dictates the distance under which two points are considered close enough to be clustered together. For a choice of epsilon, one can construct a Vietoris-Rips complex (de Silva & Ghrist, 2007), defined as follows: given a set of points $X = \{x_\alpha\} \subset \mathbb{R}^n$ in Euclidean n-space and a fixed radius ϵ, the Vietoris-Rips complex of X, $R_\epsilon(X)$, is the abstract simplicial complex whose k-simplices correspond to unordered (k + 1)-tuples of points in X which are pairwise within Euclidean distance ϵ of each other. For the purposes of Mapper, this complex is used to find connected components; all k-simplices with non-empty intersections are connected together into a single connected component. For generalized distance functions and metric spaces (i.e. non-Euclidean spaces), the usage of

the Vietoris-Rips complex can be generalized for this algorithm, and instead consider the sets of points $X_i = \{x | \exists y \in X_i,\ d(x,y) < \epsilon\}$. Each point in X is a member least one X_i, because every point is zero distance from itself, and a member of at most one X_i.

In summary, in order to create a map using the Mapper algorithm, the researcher must choose:

1. The data to be mapped.
2. A distance function or metric for calculating pairwise distances between each data point.
3. A filter function mapping the data to the real numbers, along with a set of coverings, which generally involves a choice of the number of sections to use to form the covering and the percentage by which they should overlap.
4. A distance ϵ to create sets of connected points.

The *Mapper* Clustering Algorithm

For each filter range U_α, the associated points in $f^{-1}(U_\alpha)$ are grouped into clusters. If the connected components of the data were calculated beforehand, then at this step the connected components of $f^{-1}(U_\alpha)$ are found. For ease of calculation in **R**, I instead calculated the connected components at this step for each U_α, using simple linkage agglomerative hierarchical clustering, and cutting the resulting dendrogram at height ϵ. The net result was the same: I had a list of the connected components of X which map into each filter range.

Equipped with these connected components formed from subsets of the original data set, the fact that the filter ranges mapped single points in X into (potentially)

multiple connected components, corresponding to different (but overlapping) filter covers. If the overlap between successive U_α is no more than 50%, then a single point can be mapped to at most two covers. If two covers contain the same point, then the connected components containing that point are overlapping, and are linked together.

The drawing of the map proceeds from this information, using the language of network analysis: nodes and edges (Wasserman & Faust, 1994). Connected components in each filter range identified in the previous steps are drawn, roughly ordered by filter value for simplicity. Literal x- and y-coordinates of nodes in this representation have no interpretation or meaning; the only relational data is conveyed by the edges, and the entire network can be stretched, twisted, etc. without changing this information. However, for clarity, the network is drawn as simply as possible, with minimal self-crossing and entanglement. Once all the nodes have been drawn, any nodes which represent overlapping connected components are joined together by an edge. The resulting network is called a map.

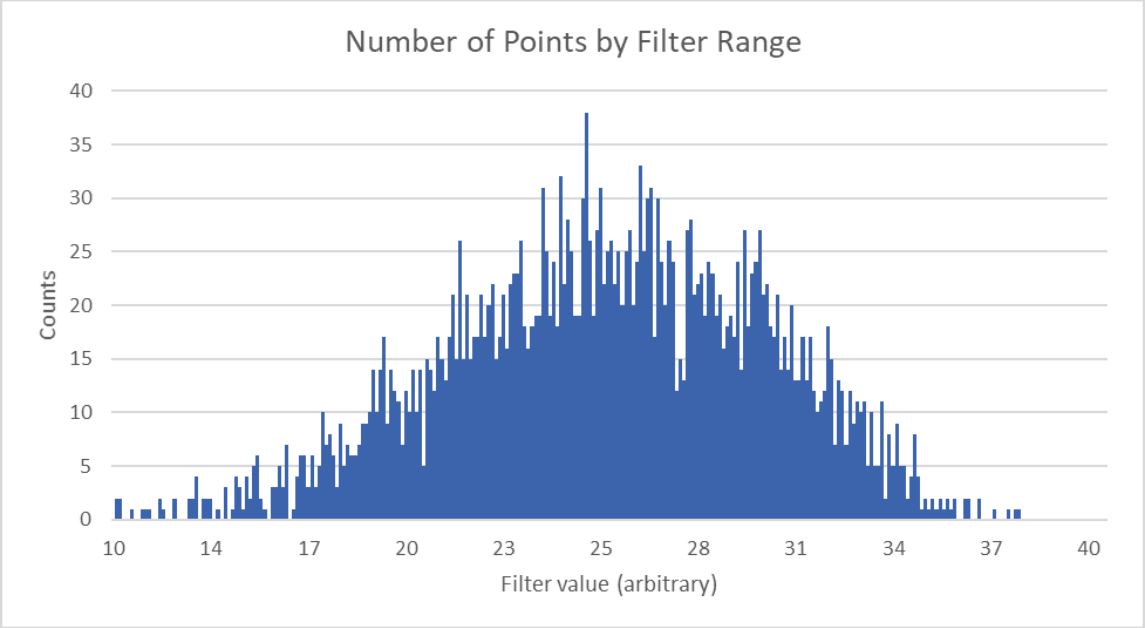
If the data are simply a large cloud of multivariate normal point data, then a final map using a filter such as density will look approximately like a string of overlapping nodes, terminating at lower filter values with brief fragmentation into tiny tails, followed by a cloud of disparate noise (points that are not connected to others in any large structure). Larger tails, or tails which separate from the main chain at lower filter values, are evidence of a more complicated structure. Forks at the higher-density end of the map indicate multiple, separable dense cores in the point cloud.

Chosen Filter Function for InIce Data

Because I am searching for groups of like-minded students in the space of beliefs, I chose a k-nearest-neighbors (knn) density estimate for the filter function. Students with high local density estimates have many other students nearby (who thus have similar beliefs). The density at each point $x \in X \in \mathbb{R}^d$ was estimated as $\tilde{\rho}(x) = \frac{k}{(n\, R_x^d\, c_d)}$ where c_d is the volume of a unit ball in $\mathbb{R}^d$, n is the total number of data points, and k was the number of nearest neighbors to use when calculating R_x^d, the distance in $\mathbb{R}^d$ to the kth nearest neighbor of point x. All terms in this equation except for R_x^d are identical for each point in X, and so can be removed to ease of calculation when creating the filter. Thus, each point was assigned a value inversely proportional to how far away the kth nearest neighbor was from them, with higher filter values corresponding to points with higher local densities; the choice of k = 20 was chosen because it produced a distribution of filter values which relatively smoothly varied over a range. There were not large, well-separated spikes in the histogram of filter values, which would increase the likelihood of significant structures in the map fractionating into substructures because some overlapping filter regions were unpopulated by chance (see Figure 4).

These filter functions will be used by the Mapper algorithm to subset the data for iterative clustering. The clusters of data in each overlapping range of filter values are connected to construct a map of related data.

Figure 4 – Histogram of filter values.
Each bar displays the number of data points which were assigned to each range of filter values when the filter was spanned by 250 covers. The central mass of values is well-populated, with only a few empty ranges at the extreme high and low ends.



Advantages of TDA over other Cluster Analyses

Why use TDA to analyze data, when other cluster analysis techniques exist and are easier to implement? This is especially relevant considering the Mapper algorithm uses another clustering technique (e.g., agglomerative hierarchical clustering) to create connected components. Like other clustering algorithms, Mapper produces a two-dimensional representation of high-dimensional data that would otherwise be difficult-to-impossible to visualize. Further, it provides relational information between different parts of the data, rather than just group memberships for the data points.

In general, TDA provides several benefits, some of which are also provided by the adapted Mapper algorithm. Further, the Mapper algorithm provides some unique

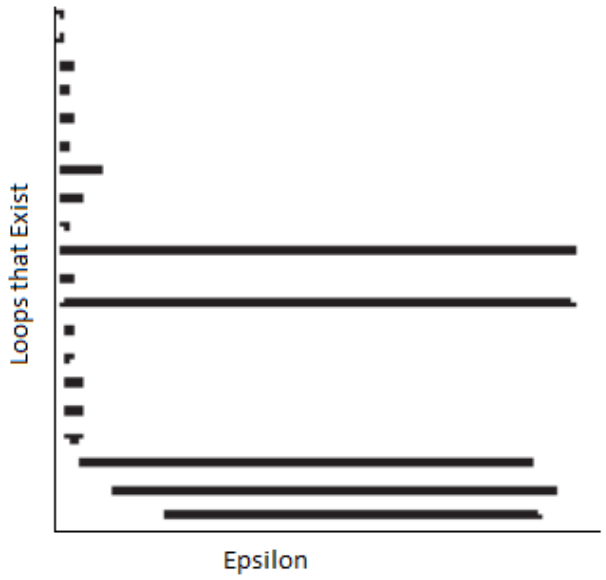
benefits above and beyond TDA in general. Carlsson (Carlsson, 2009) argues researchers would benefit from using TDA when:

1. *Qualitative information is needed*. As an initial step to understanding the data, TDA allows the researcher to obtain knowledge about how the data is organized on a large scale, and identify gross features which can later be further analyzed with other specialized quantitative methods. In the present work, I use Mapper to accomplish the task of identifying significant clusters of students with related attitudes; after identifying these groups, whether significant differences exist between these clusters exist in terms of traditional measures can be studied.
2. *Metrics are not theoretically justified*. The idea of a generalized distance metric, described above, highlights the ability of TDA to handle data in a wide range of formats, including mixed qualitative and quantitative data. Because the topology of a space is invariant to smooth deformations, studying the data in a topological sense protects the researcher from having to choose the perfect metric; an intuitive and coherent measure of similarity is sufficient.
3. *Coordinates are not natural*. An extension of the above idea, discarding the notion that properties of the data must exist in relation to the coordinates in which the data is encoded frees the analysis to potentially uncover additional emergent behavior. The importance of this aspect of TDA depends on how important the coordinates chosen are; if the coordinates are encoding theoretically cohesive and comprehensible information, then there is less reason to ignore them. Fortunately, while TDA and Mapper can work outside

a space of natural coordinates, they are also compatible with them, and so information about these coordinates is not necessarily destroyed in the construction of a map.

4. *Summaries are more valuable than individual parameter choices.* This is an aspect where Mapper, in its current incarnation, departs from generalized TDA. Because Mapper requires a choice of ϵ to create its connected components for mapping, there remains a sensitivity to parameter choice in each map. Carlsson argues that "it is not well understood that it is much more informative to maintain the entire dendrogram of the set…a summary of the behavior of clustering under all possible values of the parameter ϵ at once," and TDA accomplishes this goal with the study of persistent homologies of the space, and encoding this information in barcode diagrams (for example, see Figure 5, which shows the persistence of several loops in the data as ϵ varies).

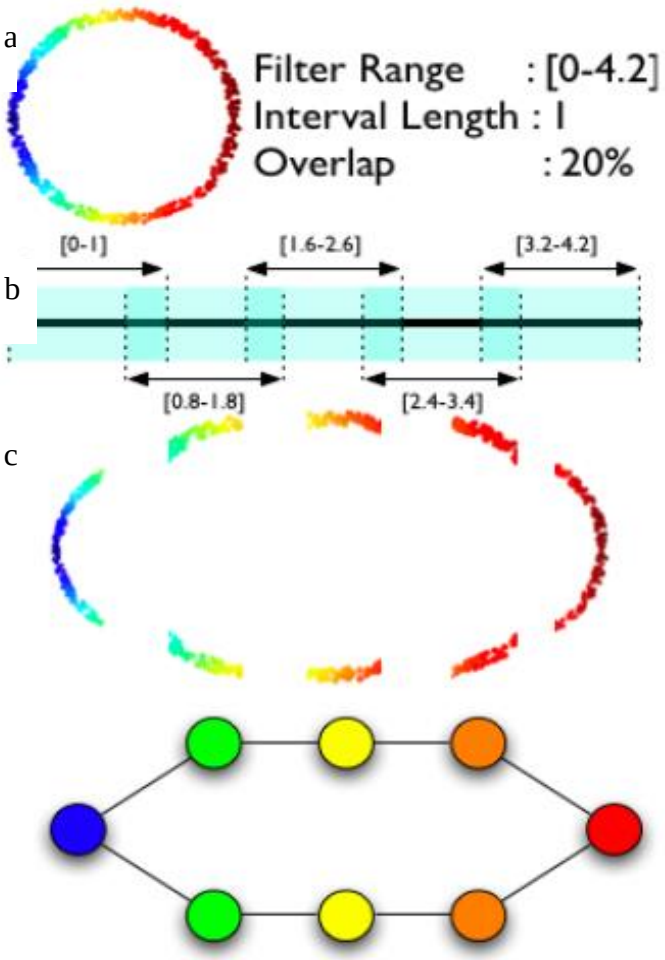
Figure 5 - Example barcode diagram
Persistence of loops in an example data set as ϵ increases. For some value of ϵ, the number of bars above that point on the x-axis says how many loops (holes in a surface) exist in the cover of the data if each point were surrounded by a ball of radius $\epsilon/2$. For this example, there are a number of short-lived loops for small ϵ, along with five persistent loops which suggest real features of the data. Picture taken from (Carlsson, 2009), with modifications.



On the latter point, Mapper differs from generalized TDA in that it seeks to provide a "horizontal" picture while TDA and persistent homology provides a "vertical" picture. In other words, TDA collapses the information about the data, with parameter ϵ, into one vertical slice of its persistence/barcode diagram, and then creates the diagram by integrating across a range of values for ϵ. However, when using Mapper, the goal is to create a picture of the data, not necessarily to detect higher-order features like loops (holes in surfaces) or voids (holes in volumes). That information is encoded primarily in the zero-order barcode diagram of the sets $f^{-1}(U_\alpha)$, which displays the number of connected components, the same information Mapper uses to construct maps. Mapper can still reconstruct and detect higher-order features, like loops, but in a more circuitous fashion (for example, see Figure 6). However, while Mapper takes the connected

components that exist for a particular ϵ and creates a map by filtering the components with its filter function to show how those components are related to a filtering parameter, TDA using barcode diagrams instead shows how the number of connected components varies with ϵ. Both pieces of information are useful, but for visualizing the distribution of data, Mapper is preferable.

Figure 6 - Mapper algorithm being applied to example data.
a) Example data sampled from a noisy circle, plus parameter choices. b) The range of filter values, and overlapping covers U_α. The filter function used was "Euclidean distance from the left-most point in the data". c) The data, partitioned into $f^{-1}(U_\alpha)$ for each filter cover. d) The resulting map, which shows the general shape of the structure and still conveys the presence of one "loop" in the data's distribution. Picture taken from (Singh et al., 2007) and modified.



Challenges of using TDA and Mapper with Quantitative Student Data

Researcher Choices

The main benefit of TDA is that it articulates the "shape" of the data. Introducing several dimensions along which the data is not variant makes it harder for the resolution of the algorithm to show details that exist in more interesting dimensions. Of course, knowing what an interesting dimension looks like is a challenge in and of itself. For this study, I used maximal variance as one of the factors in the Weighted Decision Matrix. If the spread of scores did not look like a narrow normal distribution, it was more likely to produce interesting spread of participants than if everyone fit into a normal curve.

Requirements for the Data

Quantitative Data and Discretization

When working with any data, the range of values each coordinate can take is necessarily finite, due to limits of measurement precision. With survey data, this problem is often exacerbated, particularly in the case of anchored-scale items which are commonly used. Answering an anchored-scale item from 0 to 6 gives a total of seven discrete possibilities for the response to take. Blending this answer with four other questions in a factor analysis increases the number of possibilities to 35 (five total questions, seven possibilities on each), but the range of possible values is still discretized. The fewer possible responses and the more discrete the dimension, the worse Mapper will behave when treating it like a continuous dimension.

Limits on Sample Size N

Traditional quantitative research benefits from having large numbers of data points in the sample to work with. Mapper similarly benefits from substantial numbers; when N is too small then meaningful maps cannot be created because the underlying distribution is undersampled. But if N is increased too much, the discretization of the responses into only certain possibilities creates other potential issues. Consider the case of a point cloud in R^2, with each point occupying some location on a lattice in that space. As the number of points increases, they begin necessarily occupying identical lattice, skewing concepts like local density which rely on assumptions of smooth distributions by creating sharp delta function peaks. This distribution would not be a problem if the underlying density function were in fact constructed of a handful of delta functions, but in most cases the lattice nature of the space is a result of the first issue: discretized response possibilities. Thus, there is a limit to increasing N to boost statistical power, if the corresponding questions don't have a high enough resolution.

Differing Item Scales

Often in educational research, survey data will include questions which return data on completely different scales. A Likert-type question using an anchored scale from 0 to 6, student letter grades, GPAs, and SAT scores all have very different distributions even though these responses can all be considered quantitative data[9]. Whichever distance

[9] In the case of student letter grades, each letter traditionally corresponds to a particular percentage of total points (e.g., an A is 90%+, or in a system with plus and minus letter grades, and A might be 94%+, and and A- might be 90% to 93%. While a Likert scale technically produces ordinal variables, statistical analysis usually assumes the intervals between levels are likewise equally spaced.

function is chosen needs to properly handle dimensions with potentially wildly differing scales and variance. If the data are well-behaved and normally distributed, then the dimensions could be centered and standardized to their respective standard deviations, but not all data are normally distributed.

Correlation between Survey Items and the Distance Metric

Using a Euclidean distance measure, while simple, imposes several assumptions about the data. Among these is a Cartesian metric for the data space, with each dimension orthogonal to the others. However, attitudinal data are often correlated with other data, sometimes to a high degree. Some of this correlation can be collapsed by identifying underlying factor structures, as I did with the attitudinal data to identify latent variables, as discussed previously. However, exploratory (or confirmatory) factor analysis does not guarantee orthogonal factors, as a principle components analysis would, to maximize interpretability (Jolliffe, 2002). As factors become more highly correlated, the distance between points along those factors should become smaller, but this change is not reflected in the metric. As an example, consider the two points (0, 1) and (1,0) in R^2. The distance between these points when the basis vectors for the space are orthogonal is $\sqrt{2}$. As the bases become more collinear and the angle between them (measured in the original orthogonal basis) shrinks, this distance would decrease, and in the limit of the angle between the bases going to 0 then the true distance should likewise approach 0, and the difference between the two points would be a result of measurement error that assigned different values to each question.

As stated on page 72, one benefit to using TDA over traditional cluster analyses is that it is robust to various choices of metric. However, much of that robustness comes

from analyzing the full range of values for ϵ; since Mapper builds a map with a single value, it may be vulnerable to assuming correlated coordinates/factors are in fact orthogonal. The risk can be mitigated by looking at a small range of values of ϵ to ensure stability of the map, but choosing variables which are less correlated also ameliorates this concern.

Because the affective spaces described by the theoretical frameworks employed in this work are so related to each other, some correlation between these frameworks is expected as they describe overlapping concepts. A decision matrix (described above) was used to select dimensions which were minimally correlated with each other in order to minimize the risks of assuming a Euclidean metric (and the associated orthogonal basis vectors) to apply to the vector space of beliefs.

Results

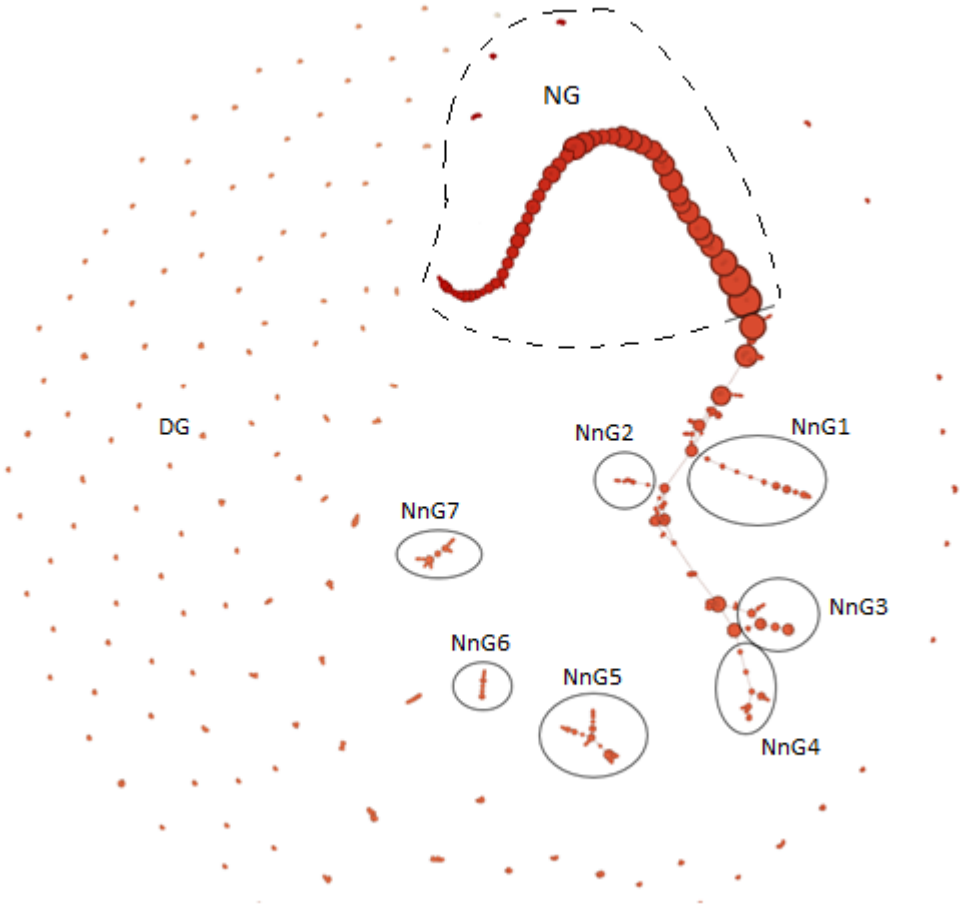
I identify a total of eight distinct groups in the data space, plus a large cloud of "ungrouped" data which were not mapped into a structure because of the relatively large distance between its constituent members. The large group with the highest density is identified as the "normative group" (NG, see **Figure 7**, top)[10]. Seven related groups ("near-normative") were identified by their proximity to the normative group in the map and numbered for identification (NnG1-NnG7). The proximity consisted of structural links for some of the near-normative groups (NnG1-NnG4), and distance in the factor

[10] The small, dark red nodes are clusters of data points with filter values at the extreme high end where not every cover was populated, resulting in them being separated from the main structure of the normative group in the map. However, these points were tested for differences from the larger normative group using two-way permutation tests, and no statistically significant differences in their distributions were found, so they were considered to be part of the normative group for subsequent analyses.

space for others (NnG5-NnG7). On average, the near-normative groups' center (calculated as the centroid of the member data points) were 1.2 units away from the center of the normative group, which motivates their characterization as "near-normative" since this is a comparatively small distance in the entire factor space.

The "cloud" of points which did not coalesce into any large-scale structure (see Figure 7, left side) is collectively named the "disparate group" (DG); the members of this group are spread across the factor space, as opposed to being concentrated in region, and are far enough from one another than each student only clustered with a small number of other students, often zero. To reduce visual noise in Figure 7, only the nodes with at least one link were included in the figure; more data points were too far from any other points to form any clusters or links, and were omitted from the image.

Figure 7 - Map of the InIce attitudinal factors data with highlighted groups.
Nodes (red circles) represent data which has been clustered together in one iteration. Size of the node corresponds (non-linearly) to the number of data points in that cluster. Color represents density in the attitudinal vector space (more red = higher density). Lines between nodes represent links made by the Mapper algorithm between clusters with overlapping membership. Eight features were identified in this map (circled and named), plus an additional "group" for consideration consisting of all the nodes/data points not assigned to another group.



Group Attitudinal Differences

With the groups identified from the Mapper algorithm, differences between the groups (NG and NnG1-7) were assessed. Two-way Fisher-Pitman permutation tests (Berry, Paul W. Mielke, & Mielke, 2002) were conducted between the near-normative and normative group due to the low numbers of data points in some of the near-normative groups, which would make traditional t-tests invalid. Results were corrected for multiple

comparisons with the Holm-Bonferroni method (Abdi, 2010) to account for the fact that thirteen factors were compared between seven pairs of distributions, for a total of 91 statistical tests run in parallel. Mean values of the normative group and statistically significant differences for each near-normative group are presented in Table 15. Because of the nature of the disparate group as a "group of students with no group", I chose not to characterize it in relation to the normative group in a similar fashion using mean factor scores.

Table 15 - Attitudinal differences between groups

Negative values on the right side of the table signify that near-normative group has a lower mean value of that row's factor compared to the normative group. Results are all significant (*: p< 0.05, **: p<0.01, ***: p<0.001); p-values have been corrected for multiple comparisons.

			Near-normative Group (difference from NG)						
	NG	**s.d.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number of students	562	-	30	8	37	26	41	12	24
Value	4.41	0.72	-0.73***	-1.03**			-0.61***		
Work Avoid	2.04	1.04							-1.22***
Connectedness	4.91	0.66		-1.41***	-0.56***				0.66***
Perceptions of Future	5.04	0.64							
Neuroticism	2.19	0.77							
Extraversion	3.00	0.97				-0.68*	0.79***		
Belongingness	4.94	0.66							
Performance Approach	3.97	0.76	-0.68**		-0.78***	-0.60**			
Instrumentality	5.49	0.50							
Grit: Consistency of Effort	3.54	0.77			-0.72***				
Engineering Identity: Performance / Competence	4.64	0.68				-0.50*			
Engineering Identity: Recognition	4.51	0.76		-1.16**		-0.77***		-0.82*	-1.08***
Physics Identity: Recognition	4.07	0.92					-0.71***		

Rather than assuming a uniform distribution of scores, and thus characterizing the low, medium, and high ranges along each factor as 0-2, 2-4, and 4-6 respectively, I adjusted the ranges up slightly to account for the fact that the distributions of attitudinal factors tended to be skewed high on each scale. If all the scores from these thirteen factors are considered as a single distribution, it has a median value of 4 (a full point above the center of the scale), and an interquartile range of 3 to 5. Thus, I define the medium scores to be between the first and third quartile, low to be below the first quartile, and high to be above the third quartile.

With these definitions, I characterize the normative group as having:

- Low (3 or less): Work Avoidance, Neuroticism, and Extraversion
- Medium (more than 3, less than 5): Value, Connectedness, Belongingness, Performance Approach, Grit: Consistency of Effort, Engineering Identity: Performance / Competence, Engineering Identity: Recognition, Physics Identity: Recognition
- High (5 or more): Perceptions of Future, Instrumentality

I found no statistically significant differences in mean values for Belongingness, Perceptions of Future, Instrumentality, or Neuroticism between any of the near-normative groups and the normative group. Other than Neuroticism, these factors were all found to be significantly related to engineering student physics identity in Chapter 3. Notably, the two factors with the highest mean scores in the normative group, Perceptions of Future and Instrumentality, are among those with no significant variation between these groups.

Almost all significant differences from the normative group were negative, i.e., the near-normative groups had lower means on those factors. The two exceptions were NnG7, which had significantly higher average Connectedness, and NnG5, which had significantly higher Extraversion. That nearly all the significant differences were negative may be an artifact of the clustering algorithm when applied to distributions which are skewed. In most cases, this skew may be the result of ceiling or floor effects from the survey question; students could only respond in a limited range of values, so a distribution which may have otherwise looked normal could end up skewed because results which would have been lower than the minimum (maximum) value instead take the minimum (maximum) value.

Engineering Identity: Recognition was the most common factor to show differences from the normative group, differing in four of the seven cases. The differences were not all the same size, and in each case, the other factors for which that group had significant differences were unique among those four if positive and negative differences in Connectedness are considered to be unique changes. The four groups of factors which showed significant differences alongside decreased engineering recognition beliefs are:

- no other significant factors (NnG 6 only differed on the one factor);
- lower Value and Connectedness;
- lower Extraversion, Performance Approach, and Engineering Identity: Performance / Competence beliefs; and
- lower Work Avoidance and higher Connectedness.

Differences in Major Interest between the Groups

Extending the results from Chapter 3, I investigated whether there were significant differences in the interest scores in selected engineering majors between groups (e.g., NG, NnG1-7, DG). As a result of the small numbers of data points in the near-normative groups, the statistical power is insufficient to resolve any but the largest effects, which were not present. The disparate group, on the other hand, has sufficient numbers and represents a potentially interesting set of distinctions.

Two-way Fisher-Pitman permutation tests checked for significant differences in the mean scores of the normative group and the disparate group on interest in the following majors: Mechanical Engineering, Aerospace Engineering, Electrical Engineering, Civil Engineering, Chemical Engineering, Biomedical Engineering, Computer Engineering, and Information Technology (IT). These majors were chosen following the results of Chapter 3; at least two majors were chosen from each tier (groups A, B, and C, see page 48). The results of these tests are shown in Table 16.

Table 16 - Differences in major interest between NG and DG
Differences on selected interests. Differences with superscripts are statistically significant (*: p< 0.05, **: p<0.01, ***: p<0.001); p-values have been corrected for multiple comparisons.

	Difference from NG	95% CI	Effect size
Aerospace	-0.279**	(0.110, 0.447)	0.15
Mechanical	-0.507***	(0.347, 0.665)	0.27
Electrical	-0.139	(-0.030, 0.308)	0.07
Civil	-0.227*	(0.067, 0.388)	0.12
Chemical	-0.276*	(0.104, 0.447)	0.15
Biomedical	-0.057	(-0.123, 0.238)	0.03
Computer	-0.027	(-0.146, 0.200)	0.01
IT	-0.224*	(0.074, 0.375)	0.13

I find statistically significant differences in mean Interest between the normative and disparate groups for Aerospace, Mechanical, Civil, and Chemical Engineering, as

well as Information Technology. The largest effect size was seen in the difference in Interests in Mechanical Engineering; the normative group showed a small-to-medium difference, while the others showed small differences. The disparate group had lower scores for all interests shown in Table 16.

Demographic Differences Between Groups

Four successive logistic regressions were constructed to predict membership in the normative group using students' self-reported demographic information as predictors. Students from the normative group and the disparate group were included in these regressions; those in the near-normative groups were excluded due to low statistical power. The normative group contained 562 students, and the disparate group contained 2040 students. Power analysis with these sub-samples suggests that with 80% power, significant differences in proportions between these groups of effect size 0.14 or larger should be detectable, corresponding to a difference of at least 1% (for "very small" proportions: less than 3% or more than 97%) to at least 7% (for "large" proportions: greater than 35% and less than 65%).

In the first model, the odds ratio of being in the normative group was predicted using student gender identity (Q17). For example, students who responded that they identified as "Female" only had the factor level "Female", a student who responded with "Female" and "Cisgender" would be in the factor level "Female Cisgender". "Male" was chosen as the reference level because it was the most populated level. "Female" was the only statistically significant factor found of 12 non-reference levels; these results are summarized in Table 17.

The second model was similarly built using student race/ethnicity information (Q16) and a reference level of "White". Four factors of a total of 41 unique levels (combinations of responses) that appeared in the data were found to be statistically significantly different from the reference level: "Asian", "Black", "Hispanic", and "White Hispanic"; these results are summarized in Table 18.

The third model added the factors from the first and second test together a combined model. The reference levels for this test were the same as the first two: "Male" and "White". The first factor again had 12 other factor levels, and the second factor had 41 other factor levels. These results are summarized in Table 19.

The fourth model was built using recalculated factors for each student by considering their responses to Q16 and Q17 combined together. A single-factor regression model was then tested using this composite factor. For example, a student who answered "Black" to Q16 and "Female" to Q17 would have the "Black Female" factor, while a student who answered both "Hispanic" and "White" to Q16 and "Male" to Q17 would have the "Hispanic White Male" factor. The reference level for this test was "White Male". Seven factor levels were found to be statistically significant of a total of 87 different combinations of responses to Q16 and Q17 together. Running this regression using only the 6 demographics which created significant effects ("Asian", "Black", "Hispanic", "White", "Female", and "Male") does not significantly change the result, with an average change in the odds ratio of less than 0.002. The largest difference was for the reference factor, which increased by 0.034 (+3.4% likelihood). None of these results were statistically significantly different from the model with 87 factor levels, so the

results of the test which used more authentic student identities are reported. These results are summarized in Table 20.

Table 17 - Odds ratio of membership in NG predicted by gender
All results are significant at (*: p< 0.05, **: p<0.01, ***: p<0.001). Factors not shown are non-significant.

	Odds Ratio	Estimate	Sig.	N
Male	(reference level)	-1.279		1984
Female	0.685	-0.378	**	647
(other factor levels)	-		n.s.	65

Table 18 - Odds ratio of membership in NG predicted by race/ethnicity
All results are significant at (*: p< 0.05, **: p<0.01, ***: p<0.001). Factors not shown are non-significant.

	Odds Ratio	Estimate	Sig.	N
White	(reference level)	-1.164		1564
Asian	0.576	-0.551	**	328
Black	0.484	-0.725	*	99
Hispanic	0.536	-0.622	**	237
White Hispanic	0.420	-0.868	**	95
(other factor levels)	-		n.s.	182

Table 19 - Odds ratio of membership in NG predicted by gender and race/ethnicity
Results are significant at (*: p< 0.05, **: p<0.01, ***: p<0.001). Factors not shown are non-significant. White Hispanic was included because of its significant presence in previous models.

	Odds Ratio	Estimate	Sig.	N
Male	(reference level)	-1.061		1984
Female	0.675	-0.421	**	647
(other gender factor levels)	-		n.s.	65
White	(reference level)	-1.061		1564
Asian	0.587	-0.567	**	328
Black	0.487	-0.702	*	99
Hispanic	0.521	-0.636	**	237
White Hispanic	0.411	-0.889	**	95
(other race/ethnicity factor levels)	-		n.s.	182

Table 20 - Odds ratio of membership in NGG predicted by combined factor
Results are significant at (*: p< 0.05, **: p<0.01, ***: p<0.001). Factors not shown are non-significant. Black Male and White Hispanic Female were included for completeness because of the results of prior models (see Table 17**,** Table 18**,** and Table 19)

	Odds Ratio	Estimate	Sig.	N
White Male	(reference level)	-1.047		1128
Asian Female	0.475	-0.745	*	77
Asian Male	0.531	-0.633	***	242
Black Female	0.095	-2.354	*	31
Black Male	-		n.s.	67
Hispanic Female	0.434	-0.835	*	53
Hispanic Male	0.481	-0.732	**	180
White Female	0.608	-0.497	***	381
White Hispanic Female	-		n.s.	18
White Hispanic Male	0.340	-1.078	**	77
(other factor levels)	-		n.s.	331

All the intercepts for the logistic regressions were significant (p<0.001), and represent the odds that the reference population (in these cases, Male-identified, White-identified, White-identified-and-Male-identified, and White-Male-identified people) being a member of the normative group.

Table 21 - Reference level probabilities
Probabilities signifying odds of being a member of the normative group as a member of the reference level factor for each regression. All results significant at p<0.001. For example, the first regression predicted that a randomly selected student identifying as "Male" has a 27.8% chance of being in the normative group.

Regression on…	Reference Level	Probability of Ref. Level in NG	Estimate
Gender identity	"Male"	0.278	-1.279
Race/ethnicity	"White"	0.312	-1.164
Gender + race, two factors	"White" and "Male"	0.295	-1.061
Gender and race as one factor	"White Male"	0.351	-1.047

The odds ratios in the Table 17, Table 18, Table 19, and Table 20 represent the odds of that demographic group relative to this reference level. For example, in the logistic regression using just gender as a factor, the reference level ("Male") has a 27.8% chance of being in the normative group. "Female" has a 68.5% relative odds, for a total of 0.685*27.8% = 19.0% chance of being in the normative group. Across all models, the probability of membership in the normative group for members of the reference group is approximately similar, near 30%.

The difference in proportions of White Female students and any level other than White Male was not found to be statistically significant; notably, I did not find a statistically significant effect at the p<0.05 level for a difference between White Female and Black Female, despite the seemingly large difference in their values. However, because the relatively low numbers of female-identified students in the sample, this result

may be caused by a lack of statistical power; rather than making comparisons between samples of 2050 and 519, the analysis is limited to 381 (White Female) and just 31 (Black Female). In this case, the normative group had a *single* Black Female student, so the addition or subtraction of just one other student would result in a 100% change in the proportion. Likewise, here was no statistically significant difference between Male and Female students for any of the Race/ethnicity factors other than White within the same Race/ethnicity factor (e.g., no significant differences between Asian Male and Asian Female).

The compounded effect of identifying with multiple underrepresented demographics is hinted at in the comparison between the third and fourth models. The odds ratio for "Black Female", 0.095 appears to be substantially lower in the fourth model than would be predicted by a combination of the factors "Black" (0.487) and "Female" (0.675) from the third model. The combined effect from the third model is predicted as $0.675 * 0.487 = 0.328$, which is greater than 0.095 by at least a factor of three. However, again due to the small numbers of students who identified in this way, the confidence interval on the odds ratio for "Black Female" is fairly wide, ranging from 0.005 to 0.445, which includes the result of the above calculation. Thus, it cannot be concluded that there is a statistically significant difference between the estimates for Black Female students predicted by the two models, even though there appears to be a substantial decrease in probabilities.

As discussed earlier, one of the limitations of applying traditional statistical methods (like regression analyses) by first binning students into categories representing the intersections of their various identities is that students with uncommon identities end

up in groups with very low numbers, and thus the test has little power to resolve any statistically significant differences that may exist. Thus, while there are only a handful of variable combinations which showed up as statistically significant in the analysis, this result should not be taken as strong evidence that no such differences exist, but that with greater sample sizes such differences would be more readily assessed.

Including Other Demographic Variables

I tested models predicting membership in the normative group that included variables for disability status (Q15) and sexuality (Q18), coded in a variety of manners. Whether disability status was a binary "able-bodied" / "not able-bodied" or the full spectrum of eight different responses to Q15, it was never a significant effect, either on its own or in conjunction with other factors in multi-factor models or composite factor models. Likewise, sexuality was not found to be a significant effect either on its own or in conjunction with other factors in multi-factor models, whether it was coded as a binary "straight" / "not-straight", or a spectrum of five different responses.

Disability status had 1872 who responded to only Q15e ("I do not identify with a disability or impairment") and 434 students who responded with something else (but not including students who did not provide any answer to Q15). The difference in proportions was non-significant, with a confidence interval of (-0.031, 0.054), corresponding to effect sizes of -0.13 and 0.07, both of which are small.

Sexuality had 2442 students who responded to only Q18a ("Heterosexual / straight") and 77 students who responded with something else (but not including students who did not provide any answer to Q18). The difference in proportions of straight-

identified students in the normative group and the disparate group was non-significant, with a confidence interval of (-0.012, 0.020), corresponding to effect sizes of -0.18 and 0.07, which are again both small.

Considering this, there is not sufficient evidence to reject the null hypothesis that there is no difference in the proportion of able-bodied / not able-bodied students in the normative group vs the disparate group, and likewise with the proportions of straight / not straight students.

Unlike with the problem of statistical power in analyzing low-N intersections of racial and gender identities, however, the effect sizes excluded by the confidence intervals of the non-significant statistical tests with disability status and sexuality are small. For example, only if disability status had a very small (real) effect would the test have been unable to detect it. The lack of statistical significance suggests there is no difference in proportions for disability status and sexuality, unlike for the uncommon racial and gender identities which suggest that more sampling is needed to be able to say much.

Conclusions and Implications

Variability in the Normative and Near-normative Groups

I found systematic differences on the cluster of traits which define the group of normative attitudinal factors. Each of these near-normative groups differ from the normative group on some distinct factors. That is, the difference between near-normative and the normative group is not consistently along the same factor, but rather different

combinations of factors in each case, suggesting that these near-normative groups indicate ways in which a student is most likely to differ from the normative group.

The variance within the normative group itself is not negligible. From Table 15, it can be seen that many of the differences in means between groups are in fact smaller than one standard deviation of the normative group along that dimension. However, because the map is constructed in a high-dimensional space (thirteen dimensions), it is difficult to consider differences along a single dimension, because those differences can occur in thirteen different directions simultaneously.

These groups and their differences were identified without regard to the demographic responses of the students. Only their responses to the quantitative attitudinal questions were considered in the algorithm. Topological data analysis, therefore, provides an alternative approach to identifying groups of students, driven by affective measures like values and beliefs rather than *a priori* demographic markers like race or gender, in a way which would not have been discoverable with traditional analytic techniques that could not effectively illuminate the underlying structure of the data.

Attitudinal and Demographic Diversity

Students belonging to the demographics which have been traditionally considered "normative" are in fact statistically overrepresented in the normative group of attitudes found by the algorithm. From this fact, there are two interrelated conclusions to be drawn.

The first is that increasing representation in engineering of traditionally underrepresented demographics (URD) is likely to increase the diversity of attitudinal

factors present in the undergraduate student populations. By diversity, I mean specifically variance in these attitudinal factors. Though a randomly student is more likely to be part of the disparate group than the normative group because the disparate group is larger in population, historically-marginalized students are much more likely to be in the disparate group compared to white male students (up to 10 times more likely in the case of black female students). If the distribution of scores on these attitudinal factors for each demographic group persists for subsequent samples, then recruiting more engineering majors from historically marginalized groups would likely increase the number of students who hold "non-normative" attitudes, beliefs, and values.

The second, related conclusion is that increasing the variability in attitudes held by the students by changing recruiting strategies (say, by expanding the discourse on what sorts of careers or career interests one could address with an engineering degree) is likely to increase the demographic diversity of engineering students. Though the majority of students in the normative group exclusively identified as white and male (56%), this was not true of the disparate group, which had only 41% of students identifying exclusively as white and male[11]. Thus, a randomly chosen student with beliefs that would place them in the disparate group is more likely to *not* identify as a white male. Appealing to students on the basis of attitudes and values that are not traditionally held or strongly espoused in engineering lore therefore provides an alternate route to increasing

[11] Difference in proportions between normative group and disparate group is statistically significant ($p<0.001$). The proportion of white male students in the disparate group, taken as a sample from a distribution instead of a population statistic, has a 95% CI of [38.8%, 43.1%], and is statistically significantly different from 50% ($p<0.001$).

representation of historically-marginalized students in engineering: changing the messaging for pre-college students and within engineering programs and classes may attract and retain a broader range of attiudes.

A limitation of all these conclusions regarding student demographics is that all the analyses are concerned with likelihoods. That is, while the normative group is composed of *mostly* white men, they are a simple majority and less than two-thirds of the group membership. Students identifying as white and male are statistically overrepresented in the normative group, but students identifying as other demographic groups are still present. The reverse is true of the disparate group: though historically-marginalized students are statistically more likely to be in the disparate group, white male students are still the plurality, due to the very fact that historically-marginalized students are so underrepresented (as they are in STEM overall).

Therefore, despite showing statistical differences in the likelihoods of appearing in the group structures, one of the major takeaways from this analysis is that the group structure found by TDA and Mapper is not a strong reflection of race or gender. An alternative method—attempting to partition the sample according to these demographics and calling the group of white-identified, male-identified students "normative" and everyone else "non-normative"—would grossly misrepresent the reality of the attitudinal groupings. Instead, by giving priority to attitudes in the analysis, different divisions must be made, and while these divisions are unequal for different racial/ethnic groups and different genders, it is not absolute. Therefore, in applying the findings of this work, educators would be remiss to assume that a student holds normative or non-normative beliefs based on their demographic identifiers.

Limitations of this Study

As mentioned in the background on page 55, quantitative analytic techniques (like the logistic regressions used to assess differences in membership between the normative and disparate groups) are difficult to combine with an intersectional consideration of student identities. This difficulty was reflected in the results of the logistic regressions: the groups which showed statistically significant differences were almost exclusively those groups that would be considered in a standard analysis that combined race and gender, because other, less populated responses did not have the statistical power to produce a significant result if a real effect existed. The one atypical combination, "White Hispanic" (and the related "White Hispanic Male" and "White Hispanic Female") had higher representation than might have been found in another study due to the presence of FIU (a Hispanic-majority institution) in the sample and collaboration.

Unfortunately, the relatively low numbers of female-identified students in the sample (reflective of the overall underrepresentation of women in engineering) creates challenges for digging deeper into several of the results found. Despite the seemingly large difference in odds ratios between different groups of women in the fourth model (see Table 20) the confidence intervals for these estimates were too wide to be able to conclude any sort of statistically significant difference between groups of different races/ethnicities as a result of the low numbers.

Many of the same limitations discussed in the previous chapter (see page 52) are still salient to this analysis. The participating schools were not randomly chosen, and the population of those schools is not fully representative of the undergraduate engineering population of the country. Because this study was mapping the distribution of attitudes of

students from these schools, if the culture of these schools is significantly different from the "average" engineering culture, that fact may be reflected in the map, creating a unique topology that has reduced generalizability to other populations.

Directions for Future Work

As described on page 75, Mapper differs from traditional TDA using persistent homology in the way it represents the data "horizontally" instead of "vertically". Future work which better marries the two ideologies could bring the best of both approaches; a three-dimensional map which shows both the maps from Mapper for each level of ϵ while also linking maps of different parameters together as in a barcode diagram. Plans for future research include implementing such a design, but the scope of that work is beyond the current current. In the meantime, a careful researcher can protect themselves from choosing a value of ϵ which gives them unstable results (i.e., results which would drastically change under small perturbations to the parameter) by creating several maps across a small span of values and confirming that the overall trends remain consistent between maps.

One potential avenue for future investigation of this data would be to further reduce the dimensionality of the data space using a second phase of factor analysis. The data space has already been reduced by way of factor analysis once, but the fact that the near-normative groups are differing in these distinct fashions suggests that there may be additional structures within the data, i.e., meta-factors built out of factors, which would encode the types of deviation from the normative group in the directions of these near-normative groups. These meta-factors would provide additional evidence in support of

using TDA as a technique for analyzing quantitative student data by validating the group structure hinted at by the topological map. Such an analysis would nevertheless rely on a first-pass analysis with TDA to establish the subset(s) of the data worth investigating for meta-structure, as the meta-factors which exist within the data can in theory change entirely from one section of data space to another, according to the underlying distributions and their intersections.

Chapter V: Time-Dependent Characterization of Physics Identity

Introduction

The motivation for this dissertation (see Chapter 1) is to improve recruitment and retention of STEM majors, with a focus on engineering students as they make up a sizable portion of both the enrollment in introductory physics (and so are a relevant population to study in physics education research) and overall numbers of STEM graduates (U.S. Department of Labor, 2015). One avenue for improvement was to take advantage of the benefits prior work has seen associated with strong physics identities, including increased engagement, persistence, and eventual choice of career in a STEM field (Cass et al., 2011; Godwin et al., 2016; Hazari et al., 2010). However, the same benefits of persistence, increased interest, and engagement in STEM as a whole can be achieved without having specifically a physics identity.

In Chapter 3, I discussed a possible explanation for the negative associations between Connectedness and Instrumentality with a student's physics identity; namely, a student with high scores on those factors is future-minded, and so is effectively months or years "ahead" in their program, at least in terms of how they associate engineering with other fields like physics. Underpinning this hypothesis is the idea that engineering students see math and physics as less relevant over time as their college careers progress (Zavala & Dominguez, 2016; Zavala, Dominguez, Millan, & Gonzalez, 2015). Thus, I argued, a student who has a stronger connection to their future may be effectively reaching this point where they less-strongly associate physics and engineering earlier than their peers.

Prior work in Future Time Perspective theory described students with highly detailed, positive pictures of their futures, and plans for how to get there (Kirn et al., 2014); these students had clear ideas of the future person they wished to be, and the steps required to get there, which provided them with the motivation to pursue their goals to completion. Thus, if the negative association of high Connectedness and Instrumentality on a student's physics identity found in Chapter 3 is reflective of an alternative identity which achieves the same positive outcomes, then the fact that the association is negative is not necessarily indicative of a problem. However, exploring possible links between depressed physics identity and a future-oriented framework such as this is needed, and is the focus of this chapter.

I investigate this claim by analyzing interviews with select engineering students from FIU. Prior work within the identity literature has suggested that science identity is relatively stable over short and medium periods of time (Cribbs et al., 2013; Potvin & Hazari, 2013). Thus, rather than quantitatively measuring the physics identity construct at follow-up interviews and doing an analysis of pre-to-post differences, I considered the relative salience of physics to the students' experiences in engineering, changes in their physics identity over time, and the evolution of what constitutes recognition as a physics person as the students progress. I focus particularly on physics recognition beliefs since prior work has shown that recognition beliefs are the most important sub-construct of identity and strongest predictor of self-recognition as a physics or math person (Godwin et al., 2016).

For this study, I investigate the following research questions:

1. How do FIU engineering students' perceived connections between engineering and physics change as they become more experienced in engineering?
2. How does the nature of these students' physics recognition beliefs change over a period of one year following their introductory engineering courses?

Investigating these answers will help broaden the understanding of physics identity as it pertains to engineering majors, a major population in the physical science. The current model of engineering identity shows a strong association between students' engineering identity and their physics identity, but this relationship may be more complicated than originally proposed by Godwin et al. (2016).

Methodology

This study was framed as a mixed-methods sequential, explanatory design. In the preliminary phase, a quantitative mapping of the participants and their peers was drawn from student survey data (see Chapter 4) followed by a set of phenomenological case studies using thematic analysis to develop an understanding of how students' beliefs evolve over time. As a mixed-methods analysis, it draws on both qualitative analyses (specifically, thematic analysis) and quantitative analyses for its conclusions. The quantitative aspects and associated findings were primarily discussed in Chapters 3 and 4.

Case studies are useful when a holistic, in-depth investigation of the data is needed, allowing the researcher to closely examine the data in a specific context (Zainal, 2007). This design was chosen because I wanted to be able to unpack and focus on each student's data, including both their interviews and survey data. The units of study in the case study are the students selected for interviews, with each case being a single student.

Thematic interview analysis involves analyzing interview transcripts for patterns (themes) in the data (Boyatzis, 1998; Fereday & Muir-Cochrane, 2006). Each interview transcript is marked for units of meaning, called codes, which are then grouped into broader categories which are related to the research questions driving the analysis.

As a phenomenological analysis, the phenomenon under study is the evolution of the students' identities and their relationships with physics. I focus on the perceptions and experiences of the students as objects of study, which is complimented by the use of thematic analysis as an interpretive technique. I chose this approach because it allows me to effectively characterize students' relationships with physics in sufficient depth.

To address validity threats to the conclusions drawn from this analysis, I used two main validity measures. First, I triangulated multiple data sources (survey data, interviews) to confirm emergent findings. Second, I implemented peer review of the data analysis, in which I met with other physics education researchers to review evidence for my particular claims (Merriam, 2002).

This study was conducted as part of a larger research project, which involved selecting additional students (not presented here) for interviews on a broader set of topics. I will be analyzing a slice of the overall interview data as it pertains to physics, physics identity, and students' physics recognition beliefs, and focus specifically on how these constructs change in importance and characterization over time for the students.

Choice of Participants

Following the construction of the map of student beliefs and the identification of a large "normative group" (see Chapter 4 for more details), select students were solicited to

participate in semi-structured interviews to further investigate their beliefs. Analysis of the interview content is qualitative in nature, but participant selection was heavily informed and directed by previous quantitative analyses. Participants from the disparate, near-normative, and normative groups were solicited for follow-up interviews. Only students who provided an email address to Q23 of the survey could be contacted due to the anonymity of the survey (e.g., students could voluntarily provide an email if they were willing to participate further in the study). Students were asked to participate in follow-up interviews, and offered a gift card for $25 as compensation for their time after each interview. At the end of each interview, the student was requested to participate in further follow-up interviews; all students in the current analysis participated in both the initial and follow-up interviews.

After finding the mean attitudes of the normative group (see Table 15 in Chapter 4), students in the disparate group were ranked according to their distance[12] from the centroid of the normative group's beliefs. Disparate group members who were distant from the normative group were chosen to be recruited for the interviews, in order to maximize the variability in the attitudes among interview participants, providing broad coverage of the map with a limited number of participants. The furthest members of the disparate group were solicited in the first wave of recruitment emails. Similarly, the most central members of the normative group were solicited first, and more distant members contacted after at least a week's delay. The choice to solicit a particular student was made on the basis of their attitudinal scores, which placed them in either the normative or the

[12] Using a Euclidean metric in $\mathbb{R}^n$.

disparate group, and their distance from the center of the normative group, and not on the basis of any demographic variables other than which school they attended, in an attempt to spread the interviews more or less evenly between participating universities. This purposeful sampling as part of the overall (four-institution) research project ensured adequate representation of each school in the qualitative data, in case a school-dependent effect was present. Because the number of students in the entire data set was unequally distributed between the schools (but in proportion to the relative population sizes at those schools), without such purposeful sampling it would be more likely for one of the smaller populations (FIU or UNR) to be underrepresented or missing simply by chance. However, the analysis in this chapter focuses only on students from FIU who participated in the interviews.

One reason for the specific focus on students from FIU is that the university is a Hispanic-majority institution[13] with a student population representing heritages across the Caribbean and Latin America. Previous work by Zavala et al. (Zavala & Dominguez, 2016; Zavala et al., 2015) showed modest decreases in the reported relevance of math and physics content to engineering through a study of students in a Chilean and a Mexican university. Of the schools participating in the research collaboration, the student body of FIU is most similar (in demographics) to those studied by Zavala et al., thereby maximizing the relevance and transferability of those findings to the current work.

[13] As of Fall 2014, 62.6% of students at FIU identified as Hispanic (Office of Governmental Relations, 2014)

The original InIce survey was deployed between September 2nd and September 14th, 2015, during the first two weeks of the semester. Interview solicitation began in April of 2016., and the first interviews were completed between May 20th and June 8th, 2016, at the end of the spring semester following the deployment of the InIce survey (approximately nine months after survey data collection). The second, follow-up interviews with the same students were completed between November 4th and November 15th, 2016, in the latter half of the following fall semester, approximately one year (14 months) after the initial survey, and approximately six months after the first interview. The interview protocol for the interviews can be found in the Appendix on page 157.

About the Participants

In this section, I begin by describing demographic similarities between the surveyed students. I then summarize their differences on the quantitative affective measures and finish with a brief description of each student individually, including their scores on the physics identity sub-constructs from the InIce survey.

A total of five participants from FIU were interviewed. All five were straight-identified and female-identified students. All but one of the students were born outside the United States in a Caribbean or South American country. Every students' parents/guardians were born outside the United States in a Caribbean or South American Country. All students listed two parents or guardians; one indicated two female parents/guardians, while the others each identified one male and one female parent/guardian. Most parents/guardians had some college education at an associate's or bachelor's level. Three students reported having no disability, one had a disability not

listed (PTSD), and one declined to answer. All but one of the students participating in the interviews were not first-year students, including two who had attended other colleges/universities prior to attending FIU.

In addition to being "far" from the center of the normative group, with an average distance of 4.37 units, disparate group members were also pairwise distant from each other, with an average pairwise distance of 6.04 units (min 4.05, max 8.93). See Table 22 for details. For scale, each of the thirteen dimensions used to create the map spanned a range from 0 to 6, and the maximum possible distance between two points was 21.63. The mean distance between points in the total sample was 5.76. Therefore, in terms of their affective scores, students were as different from each other as they were from the normative group, which was a consequence of the intentional selection of distant students to cover a wide range of beliefs.

Table 22 - Pairwise distances between interview participants and the normative group
Distances in the space of attitudinal factors. The first four participants are all members of the disparate group, while the fifth, Pilar, is a member of the normative group.

	Allison	**Betty**	**Cara**	**Elisa**	**Normative Group**
Allison					4.38
Betty	5.61				3.70
Cara	5.99	5.06			3.70
Elisa	8.93	7.46	7.18		5.70
Pilar	6.18	4.43	4.05	5.47	2.60

Allison[14] was a second-year[15] mechanical engineer, who was born in Peru, and who initially came to FIU as an electrical engineer but switched majors on the first day of orientation. She had high scores on all three physics identity sub-constructs.

Betty was a second-year biomedical engineer who was born in the United States. She had a moderately high score on the physics interest sub-construct, but low values for both physics performance/competence and recognition beliefs.

Cara was a second-year civil engineer who was born in the Bahamas. In the survey, she responded that she was in her fourth year of college, though later interviews elaborated that she was a sophomore at FIU and had previous experience in another school. She had low sub-scores on physics performance/competence beliefs and interest, and a moderate score on physics recognition beliefs.

Elisa was a first-year construction management engineer who was born in Haiti. In the survey, she listed her major as mechanical engineering. She had high sub-scores for physics interest and performance/competence beliefs, and a low sub-score for physics recognition beliefs.

Pilar was a third-year biomedical engineer who was born in Columbia, and a transfer student from another college. At the time of the first interview, she was a mature student who had returned to college, and was 33 years old. She had a moderately high

[14] All names are pseudonyms. Each student was given the option to choose their own pseudonym for use in this project.

[15] At the time of the InIce survey, in September 2015.

sub-score for her physics interest, and moderate to moderately-low sub-scores for both physics performance/competence and recognition beliefs.

See Table 23 and Table 24 for a summary of these student characteristics.

Table 23 - Summary of selected student demographic information
All five students had very similar demographic traits and backgrounds. Some of the salient differences are below.

	Year	Major	Heritage	Born in the US?
Allison	2nd	Mechanical	Peruvian	No
Betty	2nd	Biomedical	Venezuelan	Yes
Cara	2nd	Civil	Bahamian	No
Elisa	1st	Construction	Haitian	No
Pilar	3rd	Biomedical	Columbian	No

Table 24 – Interview participant physics identity sub-construct scores
Scores at the time of the initial survey. Numbers (provided in parentheses) are on a scale from 0 to 6.

	Performance Competence	Recognition	Interest
Allison	High (5.6)	High (6.0)	High (6.0)
Betty	Low (2.8)	Low (2.8)	Moderately-high (4.3)
Cara	Low (2.4)	Moderate (3.4)	Low (2.0)
Elisa	High (5.4)	Low (2.8)	High (5.0)
Pilar	Moderately-low (3.0)	Moderate (3.2)	Moderately-high (4.7)

Choice of Questions in Interview Protocol

The interview protocols were designed for semi-structured interviews averaging approximately 30-45 minutes. A semi-structured interview format was chosen to allow for elaboration and tangents by the student as topics came up, thereby allowing them to fully express whichever thoughts, feelings, and experiences they found relevant. To reduce participant fatigue during interviews, each participant was only asked a subset of the items on the overall protocol. Every student was asked to tell a story about how they got into engineering, a block of questions about how they see engineering and their engineering identity, and questions about belongingness. The other affective constructs

used in the creation of the map had a block of questions associated each of with them, and students were asked questions from blocks on which they showed substantial differences from the normative group, as well as a block in which they were similar to the normative group, as a potential control. Table 25 outlines the blocks of questions each student was asked during the first interview; details for these questions and the entire interview protocol can be found in the Appendix. Additional questions for the second interview were personalized for each participant based on the content of their first interviews, though the same semi-structured protocol remained.

Table 25 - Interview protocol blocks asked to each participant

All students	**Story, Engineering Identity, Belongingness**
Allison	Work Avoidance, Neuroticism, Physics Identity: Recognition
Betty	Physics Identity: Recognition, Extraversion, Grit: Consistency of Interest
Cara	Value, Extraversion, Physics Identity: Recognition
Elisa	Work Avoidance, Neuroticism, Performance Approach, Grit: Consistency of Interest, Physics Identity: Recognition
Pilar	Instrumentality, Perceptions of Future, Connectedness, Physics Identity: Recognition

Results and Analysis

In this section, I explore the themes which emerged from analyzing the interview transcripts regarding physics, physics identity, and physics recognition beliefs in particular. I found evidence for two themes. First, students tended to see physics as less important and less integral to doing engineering by the second interview, compared to the first. While their view on the importance of physics may not have changed, the salience of physics in engineering contexts to them was seemingly decreased. Second, the nature of what constituted physics recognition evolved over time, and tended to move away from traditionally considered recognition events.

Salience of Physics Identity to Students' Engineering Experience

In the first interview, students were asked "what does an engineer do?" and "what skills are important for an engineer?" as broad, open-ended questions. Student self-generated responses focused on the problem-solving aspect of engineering, as well as a need for critical thinking and analytic skills, along with some creativity and interpersonal skills. Initial interviews from students with higher physics recognition beliefs (e.g., Allison, Cara, and Pilar) included a spontaneous mention of physics as an additional important skill for an engineer to have, but by the second interview, explicit mentions of physics as a required skill dropped off, even though the other skills remained prominent

Allison, who had the highest physics identity scores on the initial survey, explicitly called out math and physics as being principles that engineers use in their work; when asked what engineers do, she described it as applying, math and physics principles to the daily world to solve needs (emphasis added):

> Q So what's an engineer? What does an engineer do?
> A Oh, they fix things […] **they apply math and physics principles into the daily world so they apply science into daily world needs.**
> Q Right. So what skills are important for an engineer?
> A 3-D visualization, critical thinking, problem solving, creativity and team work.

While she doesn't call out physics as an important skills for an engineer, from her description of what engineers do it is clear that she sees an important link between the physics and engineering. In the second interview, she describes a similar relationship (emphasis added):

> Q …Tell me your feelings, what is engineering?
> A It's the applied times, right**, the applied version of like physics and math or whatever**, to create and innovate new, well, you've seen machines but now it's
> just basically anything that can be incorporated into the field what they call the soft sciences like the, both soft and hard sciences.

By the start of her third year of college, Allison sees engineering is a distinct form of physics and math. Rather than applying principles from these fields, engineering is itself an applied version of those disciplines.

Betty, a second-year biomedical engineer with low performance/competence and recognition beliefs sub-scores, but a moderately high physics interest score, heavily emphasized the need for analytical skills for an engineer. She was unique among participants in that she noted that different engineers may require different skills. She identified physics as an essential skill for some engineers, but not for her own major (emphasis added).

> B The skills that are important, analytical skills, being able to see trends, at least for me it would be being able to see trends and data. We're very heavy on statistics. For a different engineer I'd say I guess mechanical or electrical, it would be more, I guess, I'm thinking more abstract. **I'm thinking of my friends in electrical engineering. They have to be very strong on physics, too, extremely strong because that is the basis of everything that they will ever do**. And so I'd say being able to pick up on. And one of the things that I remember from first getting used to learning physics was how difficult it was to pick up on certain things a problem was giving to you. So you have to pick up on patterns, you have to be able to analyze where can I get from here to there.

In the second interview, this distinction is no longer present, and she makes no mention of physics at all. Instead, her description remains focused on the analytical skills mentioned in her first interview, and added a mention of interpersonal or managerial skills required for navigating teamwork:

Q Like so what are the, so what are the skills that are important to do engineering?
B Analyzation of a problem. I mentioned problem solving before. I guess that's really broad so. I guess under analyzation, I don't know if this is, I don't know how to concise it, like concisely say it, but being able to determine different methods of how to, how to get to a solution. Even if you're not technically right or wrong, just being able to identify them. I'm sorry, what was the question again? Like what do they do?
Q You're doing great. I mean, I'm just interested in your thoughts on like what does it mean to be an engineer, what do engineers do and what skills are necessary?
B What skills you need, um, what kind of skills do they need, I guess, to get to the method, you really need to have management skills because if I didn't, if my team didn't have management skills I don't know where we'd be. Luckily we have someone that manages us, well, me and another person manage us very well.

Cara, a civil engineer with low physics performance/competence beliefs and interest sub-scores and a higher recognition beliefs sub-score, saw engineering as a way to blend math and physics together. Initially majoring in math, she was interested in doing "something in physics" when she came to FIU[16], and saw engineering as an avenue to accomplish this. Interestingly, physics alone was not seen by Cara as a way of doing math and physics together (emphasis added):

Q What about physics? You said you wanted to do something in physics. Did you think about physics?
C I did, but **I also like math and physics together and I guess engineering uses both of them** but I haven't got a class that's just like physics by itself.

Her conception of engineering requiring math and physics was reiterated when asked what skills are important for engineering (emphasis added):

[16] Despite the low score on the physics interest sub-construct, which was mainly pulled down by a very low response on the item "I enjoy learning physics." She describes later in the interviews a very negative experience learning physics in the Bahamas, and which was improved dramatically at FIU.

Q …You sort of said a little bit of this, but what other skills do you think are important for engineering?
C You have to have math skills, be good on the computer, be good with dynamics. **Physics of course.**

By the second interview, Cara still saw physics skills as required to be an engineer, repeating that engineers need math and physics skills, and further broadened her list somewhat to include personal skills (emphasis added):

Q …What sort of like, what are sort of the characteristics of the things that you need to have to be an engineer?
C I think math skills. **Definitely physics**. Mmm, a little bit of chemistry, maybe. Personal skills.
Q What kind of personal skills?
C Like to work with others because you know, any engineering project you have to work with others. Maybe managerial skills.

Elisa, a construction management engineer with low physics recognition beliefs but high physics performance/competence beliefs and interest sub-scores, also echoed the sentiment shared by others that "engineers solve problems". When asked about which skills are important for engineering, she responded with "analytical, problem solving," with no specific mention of physics or math skills as some of the other participants had. Her first interview stands out in this regard as being the only person to not mention any importance of physics skill to do engineering.

The last participant, Pilar, was a nontraditional student majoring in biomedical engineering. She had moderate scores for all three physics identity sub-constructs, with slightly lower performance/competence beliefs, and higher interest. In her first interview, she describes a broad range of important skills to be an engineer, including "more than a decent grasp of physics…a good grasp of physics". She says (emphasis added):

Q So what skills are important to do engineering?
P You have to be very analytical. Very analytical. Way patient. Very well versed in math or at least well versed, well rounded. **Have a decent grasp of physics, more than a decent grasp of physics. I think you have to have a good grasp of physics.** Chemistry is not bad especially if you're dealing with anything with materials. Being able to maybe look, because there's also creative aspect to engineering, especially if you end up in research and development or something like that. You have to be able to maybe tackle a problem from a different angle, see it from a different perspective. So having a creative edge also helps. And just a lot of patience.

By the second interview, Pilar responded differently when asked what skills are required to do engineering, saying:

P Logical thinking definitely. You do need the math and you do need the technical know-how, but if you, it's more of a way of thinking than it is the technical and the programming and all that. I mean, that helps obviously but those are tools but it's more of a way of thinking that is what engineering is.

The list of necessary skills has sharpened from the broad list provided the first time. Notable differences include no mention of physics (whereas previously "more than a decent grasp of physics" was required), as well as chemistry and patience; instead she focuses on engineering being a combination of math, technical know-how, and a certain way of thinking which uses those tools.

In summary, students were less likely to spontaneously connect physics to engineering when asked about relevant skills for engineering by the second interview, compared to the first. I interpret this decreased articulation as evidence that physics is seen as less salient to engineering over time, and more of a distinct and separate field of study.

Evolution of Physics Recognition Beliefs

Students were also asked to recall an instance in which they felt recognized as a physics person. The meaning of this question was left to the interpretation of the student, consistent with past research in this domain (e.g., (Carlone & Johnson, 2007)). They were similarly asked whether they felt like a physics person, whether there was a time they were recognized as an engineering person, and whether or not they see themselves as an engineer and why.

Allison, who had the highest physics identity scores on the initial survey, said in the first interview that she felt recognized as a physics person because "Everyone always comes to me for help questions, concerns." In other words, she felt recognized by others because they relied on her competence in physics. She elaborated and described how she recognized herself (an important feature of overall physics identity) because of her mastery and competence in the subject, as opposed to just performance:

> A …I also feel like I understand stuff vs. plugging and chugging.
> Q I see. Explain to me the difference.
> A So there's a difference between when you have a formula and you use it than to recognize why you're using that formula. And especially, for example, in dynamics which is one of the biggest ones – you can do plug and chug and you'll get through the class if you're lucky, but if you choose the wrong plug and chug you won't get the right answer because you don't have the physical concept of what's happening in the system.

Her feelings of recognition as a physics person are derived from her performance and competence beliefs in physics, and she feels recognized both by others and herself.

In the second interview, her conception of engineering as applied physics spilled over into how she conceived physics recognition. When asked about a time she felt recognized as a physics person, she interprets the question more broadly to include

engineering, explicitly saying “they’re not physics, they’re applied physics”; in her mind, even though she is doing “physics-related things” it is very distinct to her.

> Q Do you feel recognized as a physics person?
> A Not any more because I haven’t taken physics, at least purely, in such a long time.
> [….]
> Q Have you had any sort of time in, since we talked last, where you have felt recognized as a physics person?
> A Yeah, yeah. [….] I mean like physics is engineering but yeah. Like just other people needing my help on physics-related things but they’re not physics, they’re applied physics.

Because she had not taken “pure” physics in a while (i.e., a physics class, as opposed to physics in the context of an engineering class), she no longer felt recognized as a physics person, even though she was approached for help with physics-related things, as she was at the time of the first interview. One possible explanation for why this difference has developed in her mind is suggested later in the interview, when she describes the experiences of a friend of hers that she met during a summer research experience she completed at another university between the first and second interview. The friend, a physics major, “was really upset she was a physics major” and sought to switch to engineering, but was denied the ability to transfer her classes because the physics and engineering classes were incompatible.

> A Because she was already, she was already in junior year so, and like she asked her department if she could switch and the department said you could but your classes don’t count because like the way that physics approach the classes that we do are just completely, apparently completely different.

This difference in approaches to related subjects between engineering and physics may help instantiate distinctions between the fields as separate such that participating in one is exclusionary from the other.

Betty, a biomedical engineer, started with low scores on her physics performance/competence and recognition beliefs sub-constructs, but a moderately high score on physics interest. When asked in the first interview about times that she felt recognized as a physics person, she likewise focused on when her performance/competence was recognized by way of high grades:

> Q But you didn't have the physics experience in high school so I'm interested sort of in what your feeling is about the recognition you get …
> B … Um, not really in Physics 1 but in Physics 2 I remember getting like one of the best test scores like once or twice so that was definitely, I felt pretty recognized like at that point.

However, by the second interview, she no longer felt recognized as a physics person at all. Her experiences with her boyfriend, a physics major, helped shape her ideas of what "a physics person" is like, and thereby let her define herself in opposition to it because of the differences. She said:

> Q So, and that's sort of the background now so since that time, since we talked the
> last time, do you, do you feel recognized as a physics kind of person?
> B As a physics kind of person? No. I don't think so. Maybe like my means of comparison is kind of weird because I kind of like my boyfriend studies physics and not even here, in California, so he'll talk about quantum mechanics so he'll go all Googly-eyed over quantum mechanics or just some like, some really fluid mechanics or just something that I've learned about but I don't care about the intrinsic, like the specifics of it, like I've learned about it and I'll do the equation and yes like I will, like I don't want to derive anything, that's not, I'll do the problem […] I'll think about it, I'll be like that's how it works. I'll be practical but I don't want to get all theoretical and so oh, my God, this is so – no [….] I don't consider myself a physics person.

With a specific picture of what a physics person "looks like" as a result of her personal interactions with a physics major, she now excluded herself from that identity, explicitly denying self-recognition, because of the differences she saw between the

disciplines. She initially laid a tentative claim to an identity as a physics person, evidenced by her somewhat low scores on the sub-constructs, and recognition beliefs that were contingent on performance in a physics class. This tentative identity was abandoned when brought into contact with what she perceived to be a quintessential example of a physics person.

Cara similarly described in the first interview feeling recognized as a physics person because of her grades.

> Q Do you feel that you get recognized as a physics kind of person?
> C Yeah. I mean I get good grades in physics so yes.
> Q Who has recognized you as a physics person?
> C My professor.
> Q And what did that look like? How did you feel recognized in that context?
> C Um, well, on one of the exams like I did exceptionally well and he just congratulated me so that's where I felt recognized, yeah.

Like Allison, in the second interview she interprets a question explicitly about being recognized as a physics person as also being a question about engineering. For her, recognition as a physics person was simultaneously a validation that engineering was "for [her]" and her physics competence.

> Q …tell me about a time or times that you have sort of felt recognized as a physics kind of person.
> C I think when I first, when I first was in the Bahamas when I took the physics class I really, I didn't like, it was horrible. I didn't really know if I'm cut out for this. But then when I re-took it in FIU and I like understood and I got good grades then I definitely felt like engineering was for me, I could actually do physics.

Cara once again gained recognition as a result of her performance and competence in the subject; performing well led to feelings of recognition, even without an external person to explicitly recognize the achievement. However, this successful physics performance was seen as confirmation of belongingness in engineering.

Elisa, a construction management engineer with low physics recognition beliefs but high physics performance/competence beliefs and interest sub-scores, consistently said she did not feel recognized as a physics person. In the first interview, she qualified her denial by saying "I just started a couple, like two, three weeks ago," suggesting that she thought with more time she would be recognized, but without previous experience in a physics class (she had no high school physics experiences) she would not feel recognized. In the second interview, when asked about whether there were times where she felt recognized as a physics person, or whether she sees herself as a physics person, she repeated her previous statements. The one physics class she took between interviews was not a social environment where she felt recognized by either the instructor or her peers as a physics person, and thus she continued to not feel like a physics person.

Q Were there times where you felt recognized as a physics person?
E Um, I'm not sure how to [….] Uh-uh. Because the classes aren't really one to one. [….] It's more of, uh, , uh, he's teaching and you sit.
Q Right. Right. What about in other, like earlier in your education? Did you ever have those kinds of experiences?
E In physics?
E I, I'm, I took one physics class so I'm not really …
Q … Gotcha. So you wouldn't describe yourself as a physics person?
E No.

Between the initial survey early in her first semester at FIU and the first interview at the end of her first year, Elisa changed majors, from mechanical engineering to construction management. In the first interview, she explained that "by taking my classes I had different interests… like each class I took, other things came to mind" and that she settled on construction because she "wanted to be somewhere…with a career that helps me solve problems and build stuff and build from my own ideas." During the summer, in between the two interviews, Elisa audited a physics course because of time constraints on

her schedule that prevented her from being able to put in the required time to complete it; while she initially registered, she dropped the course early in the summer semester.

Pilar had previously taken physics courses at another college, saying "I got an A but I don't know anything… so I'm learning physics now in [engineering classes]". Despite this perceived lack of knowledge, she responded positively to questions in her first interview about whether she ever felt recognized as a physics person. However, despite the question being explicitly about recognition as a physics person, her affirmative response is in terms of engineering:

> P I mean, I guess because a lot of the people that, that are my extended social circle, are not, they're not in a STEM field or, you know, come from a very detailed math background or anything like that, there's some things that just like oh, wow, really, that I end up knowing that I don't think is something like very outlandish and I end up knowing the answer to and they don't. And it's like oh, wow, you're an engineer. But I just, I don't know. But it's very few. You know, where I know why this is going to go that way, you know, or something falling or don't do that there, don't connect it that way.

She describes her physics recognition experience in terms of "wow, you're an engineer", but earlier in the interview she says that she hasn't felt recognized as an engineer yet, potentially suggesting a difference in her mind between "engineer as someone who does physics" and "real engineer", which may be due to her major as a biomedical engineer.

As a non-traditional student who is significantly older than the other participants, Pilar's responses come from a unique perspective. That is, she has been an adult for substantially longer than the other students, has worked several jobs, some of which were tangentially related to biomedical engineering (according to her perceptions). Therefore, while the other students are just beginning their interaction with "authentic" engineering,

Pilar has had more experience in this regard, and so her first interview may be more similar to the second interview of the other students. This difference in experience may explain why she showed the same interpretation of physics recognition as being recognition as an engineer as Allison and Cara did in their second interviews.

Discussion

Two broad themes emerged from the interviews regarding how students saw themselves in relation to physics and engineering. The first was a decrease in the apparent perceived importance of physics skills to doing engineering between the first and second interview. The second was a change in identification as a physics person to either not identifying at all or interpreting this identification in the context of engineering, and to feel less recognized by others as a physics person as time progressed and the student advanced towards their engineering degree.

Engineering as applied physics, increasingly distinct from physics

A common theme among the students' description in the first round of interviews of the skills required to do engineering were physics skills. Four of the five participants explicitly mentioned physics, while the biomedical engineering student with lowest physics recognition beliefs sub-score (Elisa) made no mention of physics, instead bringing up analytical skills and "seeing patterns". While it is true that physics traditionally requires analytical skill and pattern-sensing, the same is true of the entirety of STEM, so I cannot conclude they were implicitly talking about physics. Instead, it is more likely that physics was just not connected with engineering in their minds.

Considering this, the change in how the three other DG participants describe the required skills in the second interview is telling. Cara, the civil engineer, is the only one to repeat that physics is required to do engineering. She described her experience retaking a physics class as "when I re-took it in FIU and I like understood and I got good grades then I definitely felt like engineering was for me, I could actually do physics," when asked about a time that she felt recognized as a physics person; she connected doing succeeding in physics with feeling like she could succeed at engineering. Allison meanwhile describes engineering as "applied physics", with other parts of her interview suggesting that she sees engineering-as-applied-physics as something distinct from physics itself. For example, when she described giving "help on physics related things but they're not physics, they're applied physics." Finally, Pilar drops all mention of physics from her list of skills required of an engineer, mentioning only "logical thinking…math and… technical know-how". These skills are of course relevant to physics, but they are not unique to physics among STEM fields, and the decline in physics associations among these participants is notable.

Physics identification anchored by performance, shifting to engineering

Interview participants initially reported a wide range of identifications with physics, and feelings of recognition as a physics person. While this may be expected, given the range of quantitative scores associated to the participants (see Table 24), the responses from the survey and the responses to the interview were not in complete agreement, as expected due to the time lapse and intervening engineering experiences that occurred between rounds of data collection.

Allison, Betty, and Cara all reported feeling recognized as a physics person in the context of high performance in their previous physics classes; they described getting good grades and positive interactions with their professors as the primary justification. As second-year students, they all had taken physics courses before participating in the first interview, and so had direct experiences with college physics.

On the other hand, Elisa was a first year student who, at the time of the first interview, was only a few weeks into her first physics course. When asked whether she felt recognized as a physics person, she responded with confusion, because she "just started a couple, like two, three weeks ago." She therefore displays a similar connection between physics recognition beliefs and the formal environment of a college classroom as the previous three students, she just did not feel she had sufficient experience/recognition in a relevant environment at that time.

For these students early in their engineering careers, their conception of physics recognition seems anchored to their performance in an academic setting. Not just that high performance in the form of good grades bring the potential for recognition from their professor and peers, but also that in the absence of such a setting, the idea of being recognized as a physics person seems to be somewhat of a non-sequitur. By the second interview, this connection between academic settings and physics identity was strengthened, while the interpretation of physics recognition became more associated to engineering.

When discussing being recognized as a physics person, both Allison and Pilar frame their response by contextualizing their discussions in engineering. Allison said "I mean like physics is engineering but yeah", while Pilar mentioned her friends saying

"Wow, you're an engineer". Whereas before Allison talked about being recognized as a physics person because of her competence in the physics classroom and her physics classmates coming to her for help, she now qualifies her response as feeling recognition as an engineer.

In Cara's second interview, she also brings up engineering when asked about physics recognition. In her case, however, she is discussing how retaking physics at FIU and succeeding (both in terms of increased understanding and better grades) made her realize engineering was for her. In this case, progressing to the second interview, Cara shows the same anchoring she did in the first interview, with her recognition being contingent on successful performance.

At the time of the second interview, neither Betty nor Elisa feel recognized as a physics student, for different reasons. Elisa maintained her previous position that her lack of experience in physics explained why she did not feel recognized as a physics person. One physics class was not enough to change her physics identity, but notably she seems to believe that the number of classes is more essential to this identity than the quality of those classes. On the other hand, Betty changed from feeling recognized as a result of her performance in physics classes to strongly identifying as not a physics person. Having a close relationship with a physics major (her boyfriend) exposed her to someone who presumably has very high positive physics identity (for example, she talks about him going "googly-eyed over quantum mechanics"), which disrupted her tenuous connection to a physics identity which was anchored entirely on her class performance. Once she was given an idea of what a "real" physics person was like, in the case of her boyfriend, she was better able to define herself in contrast.

Allison also reported having experience with a physics major over the summer between the first and second interviews. Her situation (a friend met during a summer research experience) differed from Cara's (a romantic partner) in several important ways, including the duration and closeness of their relationship. Their reported physics identity at the first interview also differed: Allison identified with physics much more strongly than Cara. Perhaps most importantly, Allison's friend was discontented with physics, and wished to change her major, whereas Cara's boyfriend was very positive in his interactions with physics. Thus, while Cara had a quintessential positive example of a "physics person" to compare herself to, Allison did not, which may explain the differences in their responses to these experiences.

Conclusions and Implications

As the students progressed in their education and were exposed to increasing engineering content and experiences (and, to some extent, physics content), they began to see physics at once increasingly integrated into engineering as well as feeling increasingly distant from it as a distinct domain. Student physics recognition beliefs, a key facet of their physics identity, begin to reveal more of a connection to engineering contexts, validating the model of physics identity as a core predictor of engineering identity in the absence of authentic engineering experiences (Godwin et al., 2016). Being recognized as a physics person becomes less anchored to a formal academic setting in which recognition is conferred by way of grades and recognition from peers and authority figures, and can instead be generated through engineering contexts.

However, in terms of seeing themselves as physics people, the interview participants seemed to universally draw away from identifying with physics. The lower their recognition beliefs at the start, the more they disengaged, even as their conception of physics recognition was modified to include engineering. A switch from explicit identification, present early on but only after accruing "enough" academic physics experiences, to implicit identification-by-proxy occurs as their engineering identity replaces their physics identity as their primary STEM domain identification relevant to their lives.

In summary, I find a twofold conclusion: the validation of physics identity as a predictor of engineering identity for students with less prior engineering experience, and a time-dependent evolution of physics identity as it relates to their engineering identity (and relevant college experiences). Thus, continued focus on physics identity over the long term as a key measure of interest for engineering students may be limited because of students' changing conceptions of what physics means and entails in the context of engineering.

Limitations of this Study and Directions of Future Work

This interview study was conducted with small set of deliberately chosen participants, which may restrict the generalizability of the findings. This generalizability to the greater engineering student population may be especially limited given the interview participants were all women of color with roots in the Caribbean and South America, whereas the majority of undergraduate engineering students in the United States

are male and white (NSF, 2015). This fact provides an obvious direction for future research, conducting a similar analysis on another population to extend the findings.

Only the survey data and two interviews were included for each participant in this analysis, which represents the initiation of a future longitudinal study which is expected to be extended into the future. As such, the differences found between interviews consisted primarily of the two data sources for each student. Though I combined these data with information from the quantitative survey to strengthen my claims, combining the interview data with additional follow-up interviews which are more targeted towards these research questions (as opposed to the generalized goals of the broader research collaboration from which this data was drawn) may provide greater depth of understanding than is possible with the currently-available data sources. Such a longitudinal study would also serve to better answer the question of how FTP constructs relate to the observed changes in student physics identity.

Chapter VI: Conclusions

Introduction and Summary of Findings

In this chapter I summarize the findings of this dissertation and the overall conclusions that can be drawn from the collective results. I follow that with a discussion of the implications to education research and college teaching, followed by some directions for future research in different areas that can build on these results.

In Chapter 3, I found significant associations between several affective constructs and engineering students' physics identities. Among the constructs with significant associations were Future Time Perspective constructs of Connectedness, Perceptions of Future, and Instrumentality. When the regression model was extended to include interest in different engineering majors, I found a tiered pattern of effects on the primary model, broadly corresponding to three different classes of engineering disciplines. The negative association found for two of these factors motivated in part the research in the fifth chapter to further investigate this connection. Evidence of significant associations between theoretical constructs from a variety of frameworks helped motivate the search for a way to represent the distributions of these constructs in relation to each other, spurring on the adaptation of topological data analysis in the fourth chapter.

In Chapter 4, I mapped the space of affective constructs, using a new technique in education research to reduce the thirteen-dimensional space to a two-dimensional representation. I provided several examples of interesting differences that could be found from the resulting map. I found one large group of attitudinally similar students, which I describe as the normative group, and found a small number of ways in which students

tend to differ from this normative profile. I also found evidence that the normative attitudinal group was statistically overrepresented by white male students, and discussed some implications of this finding for STEM recruitment and retention.

In Chapter 5, I analyzed interview transcripts of several FIU students selected for their location in the map generated in chapter four to investigate the evolution of students' physics identities and how they see physics as relevant to engineering. I found evidence that students' conceptions of what counted as physics recognition events changed from being anchored in their performance in physics classes to being incorporated into their engineering recognition beliefs. At the same time, students perceived physics as less salient to their engineering education and careers as they continue to advance in their education.

Summary of Answers to Research Questions

1. For the introductory engineering students at the four collaborating institutions, how are various attitudinal factors associated with students' physics identity beliefs?

I found several statistically significant associations between attitudinal factors and physics identity beliefs. The significant factors were Belongingness, Expectancy, Connectedness, Instrumentality, Perceptions of Future, Science Agency Beliefs, Engineering Agency Beliefs, Openness to Experience, and Math Identity. Of these, two (Connectedness and Instrumentality) were negative, and the rest were positive.

2. How are the associations identified in Research Question 1 mediated by students' interests in various engineering disciplines?

No statistically significant differences were found in the associations identified in the first question compared to models with an added interest term. However, I did find, roughly, three "types" of responses that the model could have to the introduction of an interest term. Which type of response an engineering major belonged to depended on how much the variance explained by the model was improved by the introduction of the additional term describing interest in that major.

3. How are students distributed in the space of affective beliefs?

I found evidence for one large "normative group", surrounded by several "near-normative" groups which differed from the normative group in distinct ways. The characterization of these groups in terms of several affective constructs is given in Table 15. Students in the "disparate group" (i.e., in neither the normative or near-normative groups) were spread across the space.

4. What demographic differences exist between students holding normative beliefs and those with non-normative beliefs?

White-identified and male-identified students were statistically overrepresented in the normative group compared to the proportion of those students in the overall sample, whether considered as independent demographic categories or in combination. Students who identified with other race/ethnicity or gender demographics were statistically less likely to be members of the normative group. I did not find any significant differences between students who identified with a disability and those who did not, nor did I find a significant difference between students who identified as straight and those who did not.

5. How do students' perceived connections between engineering and physics change as they become more experienced in engineering?

Students expressed less of a perceived connection between engineering and physics in their second interviews compared to the first (six months earlier). Physics skills were seen as more distinct from engineering in the second interview, and not necessarily required to do engineering.

6. How does the nature of students' physics recognition beliefs change over time?

Students' physics recognition beliefs became more associated with engineering over time. They tended to interpret questions about being recognized as a physics person in engineering contexts.

Conclusions and Implications

Implications for Education Researchers

Education research often focuses on a single theoretical framework at a time, interpreting results and generating theories in terms of that framework. Unexplained variance in these models is a mix of effects from unexamined constructs along with the error term present in the model. However, as the previous chapters show, the interplay between factors of different frameworks is nontrivial at best.

However, this complexity is also a boon, as it hints at the possibility to enrich the community's understanding of factors influencing STEM students. The intersection of these factors hints that gains in explanatory power are possible using the previously well-studied frameworks. The ability of a composition of different theories to explain relationships reduces the need to develop new and independent theories, as the current theories may prove sufficient in combination to explain far more than they might individually.

The model constructed of factors associated with physics identity extends beyond the framework established by Hazari et al. (Hazari et al., 2010; Potvin & Hazari, 2013). In addition to the model of identity comprised of three sub-constructs, these other constructs can be added as precursor (e.g., as may be the case of personality traits like Openness to Creativity) or outcome (e.g., in the case of Expectancy, as hypothesized in Chapter 3) effects, thus linking the Identity framework to other theoretical frameworks in psychology and education research. . These extensions are not meant as replacements for the parsimonious model already presented, where one's domain identity is constructed from several subconstructs, but rather as ways of identifying how identity connects and overlaps with other well-developed theories.

The different STEM domain identities (science identity, math identity, physics identity, etc.) have proven to be exceptionally useful frameworks for understanding student choice and persistence in college physics, math, and engineering. However, results from Chapters 3 and 5 suggest that these frameworks are not as universally generalizable to all engineering students. The different groupings of majors discussed in Chapter 3 (based on how they interacted with the linear regression on physics identity) highlight the need for different considerations for different engineering disciplines.Particularly as each discipline is not as connected to physics as, for example, mechanical engineering, the one-size-fits-all model for how physics identity is connected to engineering identity may not be appropriate to apply to engineers from fields perceived by students to be different from physics, even if physics is a prerequisite for the program and their later understanding of engineering.

Further, the use of the identity framework, particularly with regards to interdisciplinary connections between identities (e.g., the use of physics identity as a predictor of engineering identity), appears to be constrained in its applicability to the transitionary period between high school and early college. The use of science, math, and physics identities as predictors of engineering identity and choice of engineering career was motivated by the fact that many students did not have many previous authentic engineering experiences at the point where they entered their program (Godwin, Potvin, Hazari, et al., 2013). As the results of Chapter 5 show, however, these relationships may be time-dependent as experiences in college drastically affect how students author their identities.

In summary, the overall implication of this dissertation is the possibility of further understanding of how students' identities (in the context of the Identity framework) are influenced by and continue to influence their affects throughout their college experience.

Implications for Educators and Administrators

While not every engineering student needs to become a physicist, or have the strongest identification with physics, nevertheless many engineering programs require some degree of physics coursework and competency for engineers of all stripes. Despite some specialization of these introductory physics classes, distinctions are often along lines of math ability (e.g., Calculus vs Pre-Calculus) as opposed to the major of the engineer taking the class. Other majors (e.g., Biology, Education) can have physics courses with content tailored to that discipline, such as Physics for Life Sciences (Redish et al., 2014), but little difference exists between the physics classes required of engineers

in different fields at the same university. However, the physics content required, for example, for a biomedical engineer may be very different from that required for an aerospace engineer, and both are different from what is required for an industrial engineer or a computer engineer.

Therefore, there may be substantial benefits to be gained by restructuring the "one size fits all" approach to introductory physics classes in light of the differences found in how engineering interest and physics identity interact. The majority of these benefits may be realizable by creating groups of engineering majors with similar relationships to physics to each other. By making physics something the student sees as integral to their form of engineering, their engineering interest and engagement could transfer to increased engagement in physics.

Future Directions

In addition to the possibilities for future work discussed in the preceding chapters, future research building on the findings of this dissertation can investigate several possibilities.

A structural equation model predicting physics identity (Potvin & Hazari, 2013) could be extended to include additional affective constructs beyond the three sub-constructs of Performance/Competence beliefs, Recognition beliefs, and Interest, using the associations found in Chapter 3. Whether these new associations represent an improvement to the model can therefore be empirically tested by investigating whether significant paths exist between those factors, and to test the overall goodness of fit of this more complex SEM. Likewise, a more comprehensive structural equation model

predicting engineering career choices via engineering identity could be similarly extended, and improvements could be empirically verified.

Additionally, the existence of domain identities in other fields (such as biology, computer science, etc.) should be investigated for their relationship to engineering career interest, including whether or not the same identity framework used in this dissertation is extensible to those fields. Even further work could connect those domain identities to a prediction of discipline-specific engineering identity and career choice. For example, the engineering identity of a biomedical engineer may be more strongly informed by their biology identity as they enter college, rather than their physics identity.

LIST OF REFERENCES

Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of Research Design*, *1*(8), 1–8. https://doi.org/10.4135/9781412961288.n178

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

Bandura, A. (1997). Self-efficacy: The Exercise of Control. *Corsini Encyclopedia of Psychology*, *4*, 71–81. https://doi.org/10.1002/9780470479216.corpsy0836

Bandura, A. (1999). Social cognitive theory: An agentic perspective. *Asian Journal of Social Psychology*, *2*, 21–41. Retrieved from https://www.uky.edu/~eushe2/Bandura/Bandura1999AJSP.pdf

Basu, S. J., Calabrese Barton, A., Clairmont, N., & Locke, D. (2009). Developing a framework for critical science agency through case study in a conceptual physics context. *Cultural Studies of Science Education*, *4*(2), 345–371. https://doi.org/10.1007/s11422-008-9135-8

Bellman, R. (1957). *Dynamic Programming*. *Princeton University Press Princeton New Jersey* (Vol. 70). https://doi.org/10.1108/eb059970

Berry, K. J., Paul W. Mielke, J., & Mielke, H. W. (2002). The Fisher-Pitman Permutation Test: An Attractive Alternative to the F Test. *Psychological Reports*, *90*(2), 495–502. https://doi.org/10.2466/pr0.2002.90.2.495

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. SAGE Publications.

Carlone, H. B., & Johnson, A. (2007). Understanding the Science Experiences of Successful Women of Color: Science Identity as an Analytic Lens. *Journal of Research in Science Teaching*, *44*(8), 1187–1218. https://doi.org/10.1002/tea.20237

Carlsson, G. (2009). *Topology and data*. *Bulletin of the American Mathematical Society* (Vol. 46). https://doi.org/10.1090/S0273-0979-09-01249-X

Carrico, C., & Tendhar, C. (2012). The Use of the Social Cognitive Career Theory to Predict Engineering Students' Motivation in the PRODUCED Program. In *ASEE Annual Conference and Exposition, Conference Proceedings*.

Cass, C. A. P., Hazari, Z., Cribbs, J., Sadler, P. M., & Sonnert, G. (2011). Examining the impact of mathematics identity on the choice of engineering careers for male and female students. *Proceedings of the 41st Frontiers in Education Conference*, F2H–1–F2H–5. https://doi.org/10.1109/FIE.2011.6142881

Chachra, D., Kilgore, D., Loshbaugh, H., Mccain, J., & Chen, H. (2008). Being and Becoming: Gender and Identity Formation of Engineering Students. In *Proceedings of 2008 American Society for Engineering Education Conference*.

Chen, X., & Soldner, M. (2013). STEM Attrition: College Students' Path Into and Out of STEM Fields (NCES 2014-001). *National Center for Education Statistics.*, (Statistical Analysis Report. NCES 2014-001). https://doi.org/http://nces.ed.gov/pubs2014/2014001rev.pdf

Clark, M. H., & Schroth, C. A. (2010). Examining relationships between academic motivation and personality among college students. *Learning and Individual Differences*, *20*(1), 19–24. https://doi.org/10.1016/j.lindif.2009.10.002

Cole, E. R. (2009). Intersectionality and Research in Psychology. *American Psychology*, *64*(3), 170–180. https://doi.org/10.1037/a0014564

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*(4), 874–888. https://doi.org/10.1037/a0027403

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine. *The University of Chicago Legal Forum*, *1989*, 139–67.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. Retrieved from http://www.jstor.org/stable/pdf/1229039.pdf?refreqid=excelsior%3A7f75907910b33063f76cf8bf5601b8c7

Cribbs, J. D., Hazari, Z., Sonnert, G., & Sadler, P. M. (2015). Establishing an explanatory model for mathematics identity. *Child Development*, *86*(4), 1048–1062. https://doi.org/10.1111/cdev.12363

Cribbs, J. D., Sadler, P. M., Hazari, Z., Conatser, D., & Sonnert, G. (2013). The Stability of Mathematics Identity and Its Relationship with Students' Career Choice. *Proceedings of the 35th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, 553. Retrieved from http://www.pmena.org/pmenaproceedings/PMENA 35 2013 Proceedings.pdf

de Silva, V., & Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic and Geometric Topology*, *7*(1), 339–358. https://doi.org/10.2140/agt.2007.7.339

Dowson, M., & McInerney, D. M. (2001). Psychological parameters of students' social and work avoidance goals: A qualitative investigation. *Journal of Educational Psychology*, *93*(I), 35–42. https://doi.org/10.1037/0022-0663.93.1.35

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (Grit-S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Dumfart, B., & Neubauer, A. C. (2016). Conscientiousness Is the Most Powerful Noncognitive Predictor of School Achievement in Adolescents. *Journal of Individual Differences*, *37*(1), 8–15. https://doi.org/10.1027/1614-0001/a000182

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256–273. https://doi.org/10.1037/0033-295X.95.2.256

Eccles, J. S., Adler, T., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, Values, and Academic Behaviors. *Achievement and Achievement Motivation*. https://doi.org/10.1207/s15327752jpa8502

Eccles, J. S., & Wigfield, A. (2002). Motivational Beliefs, Values, and Goals. *Annual Review of Psychology*, *53*, 109–132.

Eskreis-Winkler, L., Shulman, E. P., Beal, S. A., & Duckworth, A. L. (2014). The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology*, *5*(FEB), 1–12. https://doi.org/10.3389/fpsyg.2014.00036

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, *5*(1), 80–92.

Fernandez, T., Godwin, A., Doyle, J., Verdin, D., Boone, H., Kirn, A., … Potvin, G. (2016). More Comprehensive and Inclusive Approaches to Demographic Data Collection. In *Proceedings from ASEE 2016: American Society for Engineering Education Annual Conference & Exposition*. https://doi.org/10.18260/p.25751

Formann, A. K., Ehlers, T., & Scheiblechner, H. (1980). Die anwendung der "Latent-Class-Analyse" auf probleme der diagnostischen klassifikation am beispiel der Marburger verhaltensliste. *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, *1*(4), 319–330. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1982-20091-001&site=ehost-live

Freeman, T. M., Anderman, L. H., & Jensen, J. M. (2007). Sense of Belonging in College Freshmen at the Classroom and Campus Levels. *Source: The Journal of Experimental Education The Journal of Experimental Education*, *75*(753), 203–220. https://doi.org/10.3200/JEXE.75.3.203-220

Gee, J. P. (2000). Identity as an Analytic Lens for Research in Education. *Review of Research in Education*, *25*(1), 99–125. https://doi.org/10.3102/0091732X025001099

Godwin, A. (2016). The Development of a Measure of Engineering Identity. In *123rd American Society for Engineering Education Annual Conference & Exposition* (p. 15). https://doi.org/10.18260/p.26122

Godwin, A., Potvin, G., & Hazari, Z. (2013). The Development of Critical Engineering Agency, Identity, and the Impact on Engineering Career Choices. In *120th ASEE Annual Conference & Exposition* (pp. 1–14). Atlanta, GA. Retrieved from http://www.asee.org/public/conferences/20/papers/6290/view

Godwin, A., Potvin, G., Hazari, Z., & Lock, R. (2013). Understanding engineering identity through structural equation modeling. In *Proceedings - Frontiers in Education Conference, FIE* (pp. 50–56). https://doi.org/10.1109/FIE.2013.6684787

Godwin, A., Potvin, G., Hazari, Z., & Lock, R. (2016). Identity, Critical Agency, and Engineering: An Affective Model for Predicting Engineering as a Career Choice. *Journal of Engineering Education*, *105*(2), 312–340. https://doi.org/10.1002/jee.20118

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

González, A., Fernández, M.-V. C., & Paoloni, P.-V. (2016). Hope and anxiety in physics class: Exploring their motivational antecedents and influence on metacognition and performance. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21377

Graffelman, J. (2012). Factor analysis. In *Encyclopedia of Environmetrics* (Second Edi). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470057339.vab016.pub2

Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M.-C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, *47*(8), n/a-n/a. https://doi.org/10.1002/tea.20363

Husman, J., & Lens, W. (1999). The role of the future in student motivation. *Educational Psychologist*, *34*(2), 113–125. https://doi.org/10.1207/s15326985ep3402_4

Husman, J., Lynch, C., Hilpert, J., & Duggan, M. A. (2007). Validating measures of future time perspective for engineering students: Steps toward improving engineering education. In *ASEE Annual Conference and Exposition, Conference Proceedings*.

Jackson II, R. L., & Hogg, M. A. (2010). *Encyclopedia of identity* (Vol. 1). Sage.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology*, *87*(4), 797–807. https://doi.org/10.1037//0021-9010.87.4.797

Kaplan, A., & Flum, H. (2010). Achievement goal orientations and identity formation styles. *Educational Research Review*, *5*(1), 50–67. https://doi.org/10.1016/j.edurev.2009.06.004

Katehi, L. (2009). *Engineering in K-12 education*. *Statement to Subcommittee on Research and Science Education Committee on Science, U.S. House of Representatives*. Retrieved from https://science.house.gov/sites/republicans.science.house.gov/files/documents/hearings/102209_Katehi.pdf

Kirn, A., & Benson, L. (2013). Quantitative assessment of student motivation to characterize differences between engineering majors. In *Frontiers in Education Conference, 2013 IEEE* (pp. 69–74).

Kirn, A., Faber, C., & Benson, L. (2014). Engineering Students Perception of the Future: Implications for Student Performance. In *American Society for Engineering Education* (pp. 1–51).

Komarraju, M., Karau, S. J., & Schmeck, R. R. (2009). Role of the Big Five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences*, *19*(1), 47–52. https://doi.org/10.1016/j.lindif.2008.07.001

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance. *Journal of Vocational Behavior*, *45*, 79–122. https://doi.org/10.1006/jvbe.1994.1027

Little, R. J. A., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Lock, R. M., Castillo, J., Hazari, Z., & Potvin, G. (2015). Determining strategies that predict physics identity: Emphasizing recognition and interest. In *2015 PERC Proceedings* (pp. 199–202). https://doi.org/10.1119/perc.2015.pr.045

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of Conscientiousness. *Learning and Individual Differences*, *19*(4), 451–458. https://doi.org/10.1016/j.lindif.2009.03.007

McCrae, R. R., & John, O. P. (1992). An Introduction to the Five-Factor Model and its applications. *Journal of Personality*, *60*, 175–215.

Merriam, S. B. (2002). *Qualitative research in practice: Examples for discussion and analysis*. Jossey-Bass Inc Pub.

National Academies. (2007). *Rising Above the Gathering Storm*. *Rising above the Gathering Storm*. https://doi.org/10.17226/11463

National Academies. (2010). *Rising Above the Gathering Storm, Revisited*. https://doi.org/10.5810/kentucky/9780813125763.003.0008

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, *93*(1), 116–130. https://doi.org/10.1037/0022-3514.93.1.116

Pittman, L. D., & Richmond, A. (2008). University Belonging, Friendship Quality, and Psychological Adjustment During the Transition to College. *The Journal of Experimental Education*, *76*(4), 343–362. https://doi.org/10.3200/JEXE.76.4.343-362

Plett, M., Hawkinson, C., Vanantwerp, J. J., Wilson, D., & Bruxvoort, C. (2011). Engineering identity and the workplace persistence of women with engineering degrees. In *Proceedings of 2011 American Society for Engineering Education Conference*.

Potvin, G., & Hazari, Z. (2013). The Development and Measurement of Identity across the Physical Sciences. In *2013 PERC Proceedings* (pp. 281–284). https://doi.org/10.1119/perc.2013.pr.058

Potvin, G., Hazari, Z., Klotz, L., Godwin, A., Lock, R. M., Cribbs, J. D., & Barclay, N. (2013). Disciplinary Differences in Engineering Students ' Aspirations and Self-Perceptions. *American Society for Engineering Education*.

President's Council of Advisors on Science and Technology. (2012). *Report to the President; Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics*. Retrieved from https://eric.ed.gov/?id=ED541511

Redish, E. F., Bauer, C., Carleton, K. L., Cooke, T. J., Cooper, M., Crouch, C. H., & Dreyfus, B. W. (2014). NEXUS / Physics : An interdisciplinary repurposing of physics for biologists. *American Journal of Physics*, *82*(368), 1–12. https://doi.org/10.1119/1.4870386

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True Grit and Genetics: Predicting Academic Achievement From Personality. *Journal of Personality and Social Psychology*, *111*(5), 780–789. https://doi.org/10.1037/pspp0000089

Ryan, R. M., & Deci, E. L. (2002). Overview of self determination theory: An organismic dialectical perspective. In *Handbook of Self-Determination Research* (pp. 3–31).

Sheu, H., & Bordon, J. J. (2017). SCCT Research in the International Context : Empirical Evidence , Future Directions , and Practical Implications, *25*(1), 58–74. https://doi.org/10.1177/1069072716657826

Simons, J., Vansteenkiste, M., Lens, W., & Lacante, M. (2004). Placing motivation and future time perspective theory in a temporal perspective. *European Science Editing*, *38*(2), 35–37. https://doi.org/10.1023/B:EDPR.0000026609.94841.2f

Singh, G., Mémoli, F., & Carlsson, G. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *Methods*, 91–100. https://doi.org/10.2312/SPBG/SPBG07/091-100

Trapmann, S., Hell, B., Hirn, J.-O. W., & Schuler, H. (2007). Meta-Analysis of the Relationship Between the Big Five and Academic Success at University. *Zeitschrift Für Psychologie / Journal of Psychology*, *215*(2), 132–151. https://doi.org/10.1027/0044-3409.215.2.132

Turner, E., & Font, B. (2003). Fostering critical mathematical agency: Urban middle school students engage in mathematics to understand, critique and act upon their world. In *American Education Studies Association Conference, Mexico City*.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R. W., & Davis-Kean, P. (2007). Development of Achievement Motivation. In *Handbook of Child Psychology* (pp. 933–1002). https://doi.org/10.1002/9780470147658.chpsy0315

Zainal, Z. (2007). Case study as a research method. *Jurnal Kemanusiaan*, *9*, 1–6. https://doi.org/10.1177/15222302004003007

Zavala, G., & Dominguez, A. (2016). Engineering Students ' Perception of Relevance of Physics and Mathematics. *ASEE Annual Conference and Exposition, Conference Proceedings*.

Zavala, G., Dominguez, A., Millan, A. C., & Gonzalez, M. (2015). Students' perception of relevance of physics and mathematics in engineering majors. *ASEE Annual Conference and Exposition, Conference Proceedings*, *122nd ASEE*. https://doi.org/10.18260/p.24772

Zillig, L. M. P., Hemenover, S. H., & Dienstbier, R. a. (2002). What Do We Assess when We Assess a Big 5 Trait? A Content Analysis of the Affective, Behavioral, and Cognitive Processes Represented in Big 5 Personality Inventories. *Personality and Social Psychology Bulletin*, *28*(6), 847–858. https://doi.org/10.1177/0146167202289013

APPENDICES

Appendix 1: The InIce survey

INICE

Intersectionality of Non-normative Identities in the Cultures of Engineering

Information Regarding Participation in this Survey

We are interested in your attitudes, beliefs, perceptions, and career expectations in engineering. Please make your best estimate for each item and answer as many questions as possible. There are no right or wrong answers. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself. Please note:

- You must be 18 years or older to participate.
- The survey will take approximately 20 minutes to complete.
- Participation is voluntary. You may withdraw at any time.
- Participation will NOT impact your grade in this course in any way.
- You will be asked for contact information (email) in case we want to follow-up on some of your survey responses. This information is voluntary and will not be shared with any third party.
- If you have any questions or concerns, please contact Geoff Potvin at gpotvin@fiu.edu, or (305)348-7614.
- Participants may contact the Florida International University Office of Research Integrity, located in MARC 270, at (305) 348-8311.

Thank you for your time and insight.

FIU | STEM Transformation Institute
FLORIDA INTERNATIONAL UNIVERSITY

Please continue to the next page ➡

This page intentionally left blank

Please continue to the next page ➡

Part I: Community

1. We would like to know about how you feel that you fit in engineering and belong in your engineering community.

		Not at all	0	1	2	3	4	5	6	Very much so
a.	I feel comfortable in engineering.		O	O	O	O	O	O	O	
b.	I feel I belong in engineering.		O	O	O	O	O	O	O	
c.	I enjoy being in engineering.		O	O	O	O	O	O	O	
d.	I feel comfortable in my engineering class.		O	O	O	O	O	O	O	
e.	I feel supported in my engineering class.		O	O	O	O	O	O	O	
f.	I feel that I am part of my engineering class.		O	O	O	O	O	O	O	

Part II: Engineering Course

2. Students differ in what they want to get out of the courses they take. Use the scale given to rate how important achieving each of the following is to you in this class.

		Very Unimportant	0	1	2	3	4	5	6	Very Important
a.	Doing better than the other students in this class on exams.		O	O	O	O	O	O	O	
b.	Proving to my peers that I am a good student.		O	O	O	O	O	O	O	
c.	Doing better than the other students in the class on assignments.		O	O	O	O	O	O	O	
d.	Getting a better grade than other students in this class.		O	O	O	O	O	O	O	
e.	Knowing more than I did previously about these course topics.		O	O	O	O	O	O	O	
f.	Really understanding this course's material.		O	O	O	O	O	O	O	
g.	Feeling satisfied that I got what I wanted from this course.		O	O	O	O	O	O	O	
h.	Getting a passing grade with as little studying as possible.		O	O	O	O	O	O	O	
i.	Getting through the course with the least amount of time and effort.		O	O	O	O	O	O	O	
j.	Not having to work too hard in this class.		O	O	O	O	O	O	O	

Please continue to the next page ➡

Part III: Motivation

3. The following questions relate to your attitudes and beliefs about your experiences in this course, your engineering major, and your future. Please rate your agreement for each item.

		Strongly Disagree	0	1	2	3	4	5	6	Strongly Agree
a.	I will use the information I learn in my engineering course in other classes I will take in the future.		O	O	O	O	O	O	O	
b.	I am confident about my choice of major.		O	O	O	O	O	O	O	
c.	Engineering is the most rewarding future career I can imagine for myself.		O	O	O	O	O	O	O	
d.	My interest in an engineering major outweighs any disadvantages I can think of.		O	O	O	O	O	O	O	
e.	I want to be an engineer.		O	O	O	O	O	O	O	
f.	I will use the information I learn in this engineering course in the future.		O	O	O	O	O	O	O	
g.	What I learn in my engineering course will be important for my future occupational success.		O	O	O	O	O	O	O	
h.	I do not connect my future career to what I am learning in this course.		O	O	O	O	O	O	O	
i.	My future career determines what is important in this course.		O	O	O	O	O	O	O	
j.	I expect to do well in this engineering course.		O	O	O	O	O	O	O	
k.	I am certain I can master the skills being taught in this engineering course.		O	O	O	O	O	O	O	
l.	I believe I will receive an excellent grade in this engineering course.		O	O	O	O	O	O	O	
m.	I am confident I can do an excellent job on the assignments in this engineering course.		O	O	O	O	O	O	O	
n.	Considering the difficulty of this engineering course, the teacher, and my skills, I think I will do well in this engineering course.		O	O	O	O	O	O	O	
o.	It is better to be considered a success at the end of one's life than to be considered a success today.		O	O	O	O	O	O	O	
p.	The most important thing in life is how one feels in the long run.		O	O	O	O	O	O	O	
q.	It is more important to save for the future than to buy what one wants today.		O	O	O	O	O	O	O	
r.	Long range goals are more important than short range goals.		O	O	O	O	O	O	O	
s.	What happens in the long run is more important than how one feels right now.		O	O	O	O	O	O	O	
t.	I don't think much about the future.		O	O	O	O	O	O	O	
u.	I don't like to plan for the future.		O	O	O	O	O	O	O	
v.	It's not really important to have future goals for where one wants to be in five to ten years.		O	O	O	O	O	O	O	
w.	One shouldn't think too much about the future.		O	O	O	O	O	O	O	
x.	Planning for the future is a waste of time.		O	O	O	O	O	O	O	

Please continue to the next page ➡

Part IV: Career Expectations

4. How important are the following factors for your future career satisfaction?

		Not at all	0	1	2	3	4	5	6	Very much so
a.	Making money		O	O	O	O	O	O	O	
b.	Becoming well known		O	O	O	O	O	O	O	
c.	Helping others		O	O	O	O	O	O	O	
d.	Supervising others		O	O	O	O	O	O	O	
e.	Working with people		O	O	O	O	O	O	O	
f.	Inventing / designing things		O	O	O	O	O	O	O	
g.	Developing new knowledge and skills.		O	O	O	O	O	O	O	

5. How closely do the following describe you?

		Not at all like me	0	1	2	3	4	5	6	Very much like me
a.	I have overcome setbacks to conquer an important challenge.		O	O	O	O	O	O	O	
b.	My interests change from year to year.		O	O	O	O	O	O	O	
c.	I have been obsessed with a certain idea about a project for a short time but later lost interest.		O	O	O	O	O	O	O	
d.	I am a hard worker.		O	O	O	O	O	O	O	
e.	I often set a goal but later choose to pursue a different one.		O	O	O	O	O	O	O	
f.	I have difficulty maintaining my focus on projects that take more than a few months to complete.		O	O	O	O	O	O	O	
g.	I finish whatever I begin.		O	O	O	O	O	O	O	
h.	I am diligent.		O	O	O	O	O	O	O	

6. Please rate the current likelihood of you choosing a career in each of the following fields.

		Not at all likely	0	1	2	3	4	5	6	Extremely likely
a.	Academia (Higher Education)		O	O	O	O	O	O	O	
b.	Engineering (Industry)		O	O	O	O	O	O	O	
c.	Entrepreneurship / Start a Company		O	O	O	O	O	O	O	
d.	Government / Policy		O	O	O	O	O	O	O	
e.	K-12 Education		O	O	O	O	O	O	O	
f.	Law		O	O	O	O	O	O	O	
g.	Medicine / Health		O	O	O	O	O	O	O	
h.	Non-profit / NGO		O	O	O	O	O	O	O	
i.	Other		O	O	O	O	O	O	O	

Please continue to the next page ➡

7. To what extent to you agree or disagree with the following statements?

		Strongly Disagree		PHYSICS				Strongly Agree	Strongly Disagree		MATH				Strongly Agree
		0	1	2	3	4	5	6	0	1	2	3	4	5	6
a.	I see myself as a _____ person.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
b.	My parents see me as a _____ person.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
c.	My instructors see me as a _____ person.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
d.	My peers see me as a _____ person.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
e.	I've had experiences in which I was recognized as a _____ person.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
f.	I am interested in learning more about _____.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
g.	I enjoy learning _____.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
h.	I find fulfillment in doing _____.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
i.	I am confident that I can understand _____ in class.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
j.	I am confident that I can understand _____ outside of class.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
k.	I can do well on exams in _____.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
l.	I understand concepts I have studied in _____.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
m.	Others ask me for help in _____.	O	O	O	O	O	O	O	O	O	O	O	O	O	O
n.	I can overcome setbacks in _____.	O	O	O	O	O	O	O	O	O	O	O	O	O	O

8. To what extent do you agree or disagree with the following statements:

		Strongly Disagree			ENGINEERING			Strongly Agree
		0	1	2	3	4	5	6
a.	I feel like an engineer now.	O	O	O	O	O	O	O
b.	I will feel like an engineer in the future.	O	O	O	O	O	O	O
c.	I see myself as an engineer.	O	O	O	O	O	O	O
d.	My parents see me as an engineer.	O	O	O	O	O	O	O
e.	My instructors see me as an engineer	O	O	O	O	O	O	O
f.	My peers see me as an engineer.	O	O	O	O	O	O	O
g.	I have had experiences in which I was recognized as an engineer.	O	O	O	O	O	O	O
h.	I am interested in learning more about engineering.	O	O	O	O	O	O	O
i.	I enjoy learning engineering.	O	O	O	O	O	O	O
j.	I find fulfillment in doing engineering.	O	O	O	O	O	O	O
k.	I am confident that I can understand engineering in class.	O	O	O	O	O	O	O
l.	I am confident that I can understand engineering outside of class.	O	O	O	O	O	O	O
m.	I can do well on exams in engineering.	O	O	O	O	O	O	O
n.	I understand concepts I have studied in engineering.	O	O	O	O	O	O	O
o.	Others ask me for help in this subject.	O	O	O	O	O	O	O

Please continue to the next page ➡

9. To what extent do you agree or disagree with the following statements:

		Strongly Disagree	0	1	2	3	4	5	6	Strongly Agree
a.	Learning science will improve my career prospects.		O	O	O	O	O	O	O	
b.	Science is helpful in my everyday life.		O	O	O	O	O	O	O	
c.	Science has helped me see opportunities for positive change.		O	O	O	O	O	O	O	
d.	Science has taught me how to take care of my health.		O	O	O	O	O	O	O	
e.	Learning science has made me more critical in general.		O	O	O	O	O	O	O	
f.	Engineering can improve our society.		O	O	O	O	O	O	O	
g.	Engineering will give me the tools and resources I need to make an impact.		O	O	O	O	O	O	O	
h.	Engineering can improve our quality of life.		O	O	O	O	O	O	O	
i.	I see engineering all around me.		O	O	O	O	O	O	O	
j.	Engineering allows me to think deeply about problems.		O	O	O	O	O	O	O	

10. How accurately do the following describe you now?

		Very Inaccurately	0	1	2	3	4	5	6	Very Accurately
a.	Often forget to put things back in their proper places.		O	O	O	O	O	O	O	
b.	Have a soft heart.		O	O	O	O	O	O	O	
c.	Make a mess of things.		O	O	O	O	O	O	O	
d.	Am quiet around strangers.		O	O	O	O	O	O	O	
e.	Keep in the background.		O	O	O	O	O	O	O	
f.	Sympathize with others' feelings.		O	O	O	O	O	O	O	
g.	Am interested in people.		O	O	O	O	O	O	O	
h.	Avoid my responsibilities.		O	O	O	O	O	O	O	
i.	Make people feel at ease.		O	O	O	O	O	O	O	
j.	Talk to a lot of different people at parties.		O	O	O	O	O	O	O	
k.	Leave my belongings around.		O	O	O	O	O	O	O	
l.	Feel others' emotions.		O	O	O	O	O	O	O	
m.	Am the life of the party.		O	O	O	O	O	O	O	
n.	Follow a schedule.		O	O	O	O	O	O	O	
o.	Don't talk a lot.		O	O	O	O	O	O	O	
p.	Have frequent mood swings.		O	O	O	O	O	O	O	
q.	Do not have a good imagination.		O	O	O	O	O	O	O	
r.	Have excellent ideas.		O	O	O	O	O	O	O	
s.	Get irritated easily.		O	O	O	O	O	O	O	
t.	Have a vivid imagination.		O	O	O	O	O	O	O	
u.	Get stressed out easily.		O	O	O	O	O	O	O	
v.	Change my mood a lot.		O	O	O	O	O	O	O	
w.	Am full of ideas.		O	O	O	O	O	O	O	
x.	Get upset easily.		O	O	O	O	O	O	O	
y.	Have a rich vocabulary.		O	O	O	O	O	O	O	

Please continue to the next page ➡

Part V: About You

11. What is your current major? _____

12. What year are you in college?

O 1st Year O 2nd Year O 3rd Year O 4th Year or Higher

13. Are you a transfer student?

O Yes O No

14. Please rate your current interest in each of the following majors

		Not at all interested	0	1	2	3	4	5	6	Extremely interested
a.	Aero/Astronautical Engineering		O	O	O	O	O	O	O	
b.	Agricultural and Biological / Biosystems Engineering		O	O	O	O	O	O	O	
c.	Bioengineering / Biomedical Engineering		O	O	O	O	O	O	O	
d.	Chemical Engineering		O	O	O	O	O	O	O	
e.	Civil Engineering		O	O	O	O	O	O	O	
f.	Computer Engineering		O	O	O	O	O	O	O	
g.	Construction Management Engineering		O	O	O	O	O	O	O	
h.	Electrical Engineering		O	O	O	O	O	O	O	
i.	Engineering Physics		O	O	O	O	O	O	O	
j.	Environmental / Ecological Engineering		O	O	O	O	O	O	O	
k.	Industrial Engineering		O	O	O	O	O	O	O	
l.	Information Technology		O	O	O	O	O	O	O	
m.	Materials Engineering / Material Science and Engineering		O	O	O	O	O	O	O	
n.	Mechanical Engineering		O	O	O	O	O	O	O	
o.	Multidisciplinary / Interdisciplinary Engineering		O	O	O	O	O	O	O	
p.	Nuclear Engineering		O	O	O	O	O	O	O	
q.	Other STEM-related Degree		O	O	O	O	O	O	O	
r.	Other non-STEM-related Degree		O	O	O	O	O	O	O	

15. How do you describe your disability / ability status? We are interested in this identification regardless of whether you typically request accommodations for this disability. ***(Mark all that apply)***

O A sensory impairment (vision or hearing)
O A learning disability (e.g., ADHD, dyslexia)
O A long-term medical illness (e.g., epilepsy, cystic fibrosis)
O A disability or impairment not listed above _____
O I do not identify with a disability or impairment
O A mobility impairment
O A mental health disorder
O A temporary impairment due to illness or injury (e.g., broken ankle, surgery)

Please print your specific disability/ability statuses in the space below. Examples of statuses include: Anxiety, Bipolar Disorder, Auditory Processing Disorder, Blindness, Colorblindness, Dyslexia, PTSD, Use of a mobility aid (e.g., wheelchair), etc. Note, you may report more than one.

Disability Status(es) _____

Please continue to the next page ➡

16. With which racial and ethnic group(s) do you identify? ***(Mark all that apply)***

O American Indian or Alaska Native	O Asian	O Black or African American
O Hispanic, Latino, or Spanish origin	O Middle Eastern or North African	O Native Hawaiian or Other Pacific Islander
O White	O Another race or ethnicity not listed above _____	

Please print your specific ethnicities in the space below. Examples of ethnicities include (for example): German, Korean, Midwesterner (American), Mexican American, Navajo Nation, Samoan, Puerto Rican, Southerner (American), Chinese, etc. Note, you may report more than one group.

Ethnicity(s) _____

17. How do you describe your gender identity? ***(Mark all that apply)***

O Female	O Male	O Genderqueer	O Agender
O Transgender	O Cisgender	O A gender not listed _____	

18. **How do you describe your sexual identity?** ***(Mark all that apply)***

O Heterosexual / straight O Homosexual / gay / lesbian O Bisexual O Asexual
O A sexuality not listed _____

19. How would your parent(s) / guardian(s) describe their gender identities? ***(Mark all that apply)***

Parent / Guardian # 1

O Female	O Male	O Genderqueer	O Agender
O Transgender	O Cisgender	O A gender not listed _____	

Parent / Guardian # 2

O Female	O Male	O Genderqueer	O Agender
O Transgender	O Cisgender	O A gender not listed _____	

20. What was the highest level of education for your parent(s) / guardian(s)?

	Less than high school diploma	High school diploma / GED	Some college or associate / trade degree	Bachelor's degree	Master's degree or higher	Don't know
Parent / Guardian # 1	O	O	O	O	O	O
Parent / Guardian # 2	O	O	O	O	O	O

21. Would you describe the occupation of any of your family members as any of the following professions? ***(Mark all that apply)***

	Parent / Guardian # 1	Parent / Guardian # 2	Siblings	Other Relatives
Engineering	O	O	O	O
Other STEM	O	O	O	O
Non-STEM	O	O	O	O

22. Which category best fits you and your parent(s)' or guardian(s)' background?

	Born in United States	
	Yes	No
Me	O	O
Parent / Guardian # 1	O	O
Parent / Guardian # 2	O	O

23. We may contact some students to ask follow-up questions. All communications will be confidential and your email will NOT be disclosed to any third party.

Your email address _____

You have reached the end of the survey. It is our goal that many educators will benefit from the insights you have provided! Thank you for your time.

Appendix 2: Interview Protocols, First Interview

Notes for Interviewer

Wear casual clothes, something that is similar to what the students would wear in terms of level of dress (you need not wear yoga pants). Open the interview with some casual questions such as, “How is your day going?”, “How has the semester been?”, and/or “Did you go to the game on Saturday?” If they are talking for a bit on these points let them keep going do not cut them off. You want them to talk throughout the interview get them going early.

Avoid bringing in large pieces of technology if possible as they can be distracting/intimidating. Find a space that does not have a formal interview set up (you behind a desk and them on the other side). If the room has more than one seating option let the student pick where they want to sit as that will make them more comfortable (some people don't like having their backs to the door).

Having a second person in the room can be helpful so that you can have time to pause and think or someone else can make sure that you have asked all the questions/all answers given by students are actually clear.

Try to avoid asking the student to speak in a different tone or volume than their natural speaking voice as this may make them feel uncomfortable or inadequate. Instead move the recording device around if needed.

This is a semi-structured protocol. Interview questions will be asked as listed, but additional follow up questions may be included based on individual student responses to probe student answers.

Notes to Give to Interviewee

Before starting the interview frame the interview as a conversation or a dialogue. Inform the student that this is the interview protocol (show them the physical document) and tell them you will ask these questions but you may also asked more to gain increased understanding of their story. Tell them all data will be kept anonymous and that you want them to express their opinion. Stress that there are no right or wrong answers only the story they have to tell is what we are interested in. Some questions may seem repetitive but you want to make sure that you are getting the full depth of the story.

Flow of the Interview

1.) Story

2.) Identity

3.) Belongingness

4.) Particular construct of interest (this will depend on each participant's factor scores)

Control (Similar Variable)

Differentiating Variable

Story

- How did you get into engineering?
  - ***Sit back, wait, and listen***

    Follow-ups if needed:
    - What factors do you think influenced this decision?
  - Have you selected an engineering major?
  - Why did you choose to major in [the type of engineering you decided on]?
    - Did you consider other disciplines?
      - [If yes] What helped you settle on the one you picked?
    - Did you consider other majors outside of engineering?
  - Why did you choose to go to college?
    - Why this college?
  - Have you had an individual or individuals who influenced your choice of engineering?

Identity

- Do you see yourself as an engineer?
  - Why or why not?
  - [If yes] Can you give me some examples of ways in which you see yourself as an engineer?
  - [If no] What would help you see yourself as an engineer?
- What are you impressions of engineering?
- In your words, what is an engineer?
  - What do engineers do?
  - What skills do you view as important for engineering?
- Who can do engineering?
- Do you feel that you can do engineering?
  - Why/why not?
  - Do you feel that you can understand engineering?
  - Do you feel that you can do well in engineering?
- Was there a time when you felt recognized as an engineer?
  - Can you tell me about that experience?
- What engineering experiences, if any, have you had outside of the classroom?

Belongingness

- Does engineering feel like a good fit for you? Why or why not?
- Do you feel like you belong in engineering? How?
- What characteristics of yourself make you like an engineer?
  - What characteristics of yourself make you unlike an engineer?
- Do you think that engineering is a good fit for your [friends in engineering, classmates, etc.]? Why or why not?
- Do people with different backgrounds [-OR- people who grew up differently than you did] feel included in engineering?
- Do people who think differently than you feel included in engineering?

Specific theoretical frameworks to ask for each group	
The Normative Group	Instrumentality, Perceptions of Future, Connectedness
~~<deprecated group>~~	~~Value, Neuroticism~~
Near-normative Group 1	Value, Performance Approach
Near-normative Group 2	Neuroticism, Value and Connectedness
Near-normative Group 3	Performance Approach, Grit: Consistency of Interest
Near-normative Group 4	Performance Approach, Extroversion, Grit: Consistency of Interest
Near-normative Group 5	Extroversion, Physics Recognition, Value
Near-normative Group 6	Instrumentality, Physics ID: Rec
Near-normative Group 7	Connectedness, Work Avoid

Specific Theoretical Frameworks

Refer to the above chart for selecting the following frameworks to discuss with participants. These selections should be made and documented prior to conducting the interview. If another factor besides the ones on the list for the participant's group (above chart) seems to be coming up, there is some leeway to explore other factors during the interview as well.

Perceptions of Future

- What are your goals for the future, ideally?
  - *What are your personal goals for the future?*
  - *What are your career goals for the future?*
  - *Describe where you see yourself in 10 years*
  - How did you develop [insert student vision for the future]?
- Given your knowledge about your field and the current state of your field, what do you think you can realistically be in the future?
- *What are you actively striving for?*
  - *What goals or tasks are you currently pursuing to reach your described future?*
- What do you not want to be in the future?
  - In other words, what jobs, or careers do you know you do not want to pursue?
- Why are you pursuing an engineering degree?
  - How confident are you in your choice of major?

Perceived Instrumentality

- What parts of your education do you see as relevant to your future?
  - What skills are relevant to your future?
  - Do you see what you are learning in your courses as useful to your future? In what ways?
  - What parts of your education do you see as not relevant to your future?

Connectedness

(a general tendency to make cognitive connections between the present and the future)

- Are you taking steps to reach your future goals? If yes, can you describe the steps you are taking?
- Do you spend time planning for the future? Why or why not?
  - Do you think it is important to have goals 5 or 10 years in the future?
- Does the future dictate what actions you take now? How? In what ways?

Work Avoid

- How much work do you dedicate to your classes? Your engineering classes?
- How much time do you spend on tasks related to your classes?
- How do you react when a class takes a lot of time and effort to get the grade that you desire?
  - What about classes that do not take a lot of effort?
  - What do you see as a desired grade for you classes?
- What do you think about classes that do not need much time or effort to get a passing grade?

Neuroticism

- Do you worry a lot about the future or things that might go wrong?
  - [If Yes:] What kind of things do you worry about?
  - [If No:] Why not?
- Can you describe a time when you felt anger or bitterness?
  - [If Yes:] Is it difficult for you to get angry even when it's appropriate?
  - [If No:] What is an example of a time in which you were bitter/angry?

Extraversion

- Would you describe yourself as a leader?
  - [If Yes:] What are some good examples of how you have been a leader?
  - [Follow up:] Do others ever consider you to be overbearing or too demanding when you were a leader? How?
- Are you more distant or reserved than most people you know?
  - [If Yes:] Has this affected the number of friends that you have?

Performance Approach

- Do you consider yourself a good student? How so? Do you feel you need to prove this to your peers?
- Are you a competitive student?
  - In classes, is it important to you to do better than your better than your classmates? What about your assignments? Exams?

Physics Identity: Recognition

- Do you feel recognized as a physics person? Who recognizes you as a physics person?
- Tell me about a time that you felt recognized in physics.

Value

(Participants with high FTP tend to show decreased de-valuing of future goals)

- What are your goals? Which of these goals are most important to you? Why?
- Do you consider the future when assessing what is most important to you? Why or why not? In what ways?
- Do you consider the future when making your rankings of what is was most important to you? Why or why not? In what ways?

Grit: Consistency of Interest

- When you set a goal, do you stick with it?
- What happens when you face challenges towards pursuing your goal?
- Can you give me an example of a time when you stuck with your goals?
    - What about a time when you abandoned a goal?

Career Expectations

- How do you define success?
- How would you know that you had become successful?
- What outcomes would indicate that someone is successful?

Appendix 3: The R Analysis Code

The code used to complete the quantitative analysis in the current work, including lines which ended up being superfluous to the final reported analysis and results, as exploratory analysis and dead ends.

Mapper.R

```
## mapper function

## version 1.2


library(igraph)

library(dendextend)


## example function call

# g = mapper(apsdata, apsweights, filter.method = "population", cluster.method =

"single", N = 20, overlap = .5)


#### this is automatically handled in mapper.plot() now

## colors are set up at the moment for a 256-point palette, 0 to 255. To get the heatmap

colors I was using

## run the following command

#

# palette(rev(heat.colors(255)))

#

## which reverses the normal heatmap, so white is actually cold instead of super-hot like

normal fire.
```

```
## red is a cooler color for heat, anyway, and draws the eye better.


## this doesn't do what I want yet
#setClass("mapper", contains = "igraph")


mapper = function(data, filter.weights,
                  filter.method = "size", # can also select "population"
                  cluster.method = "single", # can also select "average", "complete", "ward"
                  ## cannot do centroid clustering because of the tree-cutting method used, fix later?
                  N = NULL, overlap = 0.5, h.list = NULL, max.k = NULL,
                  simplify = TRUE, set.filter.color = TRUE, filter.color.high = FALSE,
                  distance.matrix = FALSE, set.layout = TRUE, constant.cut = FALSE){
  #### Readme ####
    # data = matrix/frame containing n-many d-dimensional observation coordinates
    # weights = list of n-many weights, the result of a filter
    #           function applied to data
    # filter.method = type of data grouping for how to split the filtered data
    # cluster.method = method passed to heirarchal clustering method
    # N = number of data points in each slice of equalN; should be at least about 20 for good knee calculations
    # overlap = fractional overlap between slices to join them together
```

```
    # h.list = pre-provided cuts, likely from a previous iteration of the program. Must have length N.
    #       Supercedes constant.cut if both are present and h.list is satisfactory
    # max.k = maximum number of allowable clusters, used in NbClust; if a slice has fewer points than this, that slice
    #       will have a max.k equal to the number of points in the slice, minus 1. !Problem ?
    # simplify = whether to remove multiple edges between nodes that can result from overlap, in the graph
    # set.filter.color = whether to assign node colors based on a 255-point color scheme
    # filter.color.high = whether high values should get dark red
    # set.layout = whether to run mapper.layout and store to the layout before returning; sets layout.auto
    # constant.cut = whether to automate the cut-height selection with a single value for all slices
        # defaults to FALSE, but you can enter a number here which will be passed to the cutting



  if(length(h.list) == 0){
    h.true = FALSE
  } else if(length(h.list) > 0 & length(h.list) != N){
```

```
    warning("The provided h.list is of improper length, and will be ignored. Note: h.list must have N-many entries.")
    h.true = FALSE
  } else{h.true = TRUE}

  if(filter.method == "population"){
    g = make_rough_pop_graph(data, filter.weights, cluster.method, N, overlap, max.k, distance.matrix, constant.cut, h.list, h.true)
    hlist = attributes(g)$hlist
    g = connect_rough_pop_graph(g, data, filter.weights, N, overlap, distance.matrix)
  }

  if(filter.method == "size"){
    g = make_rough_size_graph(data, filter.weights, cluster.method, N, overlap, max.k, distance.matrix, constant.cut, h.list, h.true)
    hlist = attributes(g)$hlist
    g = connect_rough_size_graph(g, data, filter.weights, N, overlap, distance.matrix)
  }

  V(g)$members = V(g)$name
  # set the names of each vertex equal to the NUMBER of points making it up
  for(i in 1:length(V(g))){
    g = set.vertex.attribute(g, "membersize", i, length((V(g))$members[[i]]))
```

```
  }

  if(simplify == TRUE){ # removes multiple edges that can come from making the
overlapping connections
    g = simplify(g)
  }

  if(set.filter.color == TRUE){
    ## palette(heat.colors(X)) has X many shades of colors
    ## starting with dark red at the low end and going to
    ## white at the high end

    ## since we want red to be "good", depending on whether the filter
    ## measures something like density (high = good) or distance (low = good)
    ## we want the dark red to be either high or low.

    ## in all cases, the "color" scale will go the same way; we simply reverse the
    ## palette in one situation vs the other

    V(g)$color = 254*scale01(V(g)$filter)+1 # colors proportional to filter values
    # the plus one because for some reason a color of 0 is super bad
    if(filter.color.high == FALSE){
      g$color.high = FALSE
```

```
    g$color.palette = palette(heat.colors(256))
  } else{
    g$color.high = TRUE
    g$color.palette = palette(rev(heat.colors(256)))
    ## since the dark reds are normally low, we reverse the palette
  }
 }
 ## store the mapper creation settings in the object to reference later
 g$filter.weights = filter.weights
 g$filter.method = filter.method
 g$cluster.method = cluster.method
 g$N = N
 g$overlap = overlap
 g$constant.cut = constant.cut
 g$h.list = hlist
 g$layout = mapper.layout(g)
 V(g)$gid = c(1:length(V(g)))

 return(g)

}

mapper.reduce.clutter = function(map, N = 2){
```

```
  attr = attributes(map)
  c = clusters(map)
  M = c$membership
  S = c$csize
  clutter.vertices = which(N >= S[M])
  map = map - clutter.vertices
  attributes(map) = attr
  map$reduce.clutter = TRUE
  ## now we need to slice things like the layout and weights to just include those
corresponding to the Ids we have
  map$layout = map$layout[V(map)$gid,]
  map$filter.weights = map$filter.weights[V(map)$gid]
  return(map)
}

mapper.reduce = function(my.map, N = 1, keep.attr = TRUE){
  ## checks each connected component for the number of unique members
  ## and removes all components with N or less total unique members among nodes
  map.copy = my.map # store a copy so we don't destroy our iterator
  attr = attributes(my.map)
  c = clusters(my.map)
  M = c$membership
  for(i in 1:c$no){
```

```
    if(length(unique(unlist(V(map.copy)$members[M == i]))) <= N){
      clutter.vertices = which(M == i)
      clutter.id = V(map.copy)$gid[clutter.vertices]
      clutter.vertices = match(clutter.id, V(my.map)$gid)
      my.map = my.map - clutter.vertices
    }
  }
  if(keep.attr == TRUE){
    attributes(my.map) = attr
    attributes(my.map)$reduce = TRUE
    ## now we need to slice things like the layout and weights to just include those
corresponding to the Ids we have
    attributes(my.map)$layout = attributes(my.map)$layout[V(my.map)$ids,]
    attributes(my.map)$filter.weights = attributes(my.map)$filter.weights[V(my.map)$ids]
  }
  return(my.map)
}

mapper.filter.size = function(map, N=1){
  newmap = map - V(map)[V(map)$name <= N]
  return(newmap)
}
```

```
mapper.recolor = function(map, N = 256){
  V(map)$color = (N-1)*scale01(V(map)$color) +1
  return(map)
}

mapper.plot = function(map, layout=NULL, new.palette=NULL){
  ## plots the map, and then overlays a set of invisible points on top of the map
  ## to be used with the identify function to read out vertex attributes
  if(length(layout) > 0){ # i.e. we stated a layout
    # layout = layout; "save" time by not actually running this line
  } else if(length(map$layout) > 0) { # i.e. the map has an innate layout
    layout = map$layout
  } else{ # i.e. no layout provided
    warning("no layout provided, generating random.auto layout.")
    layout = layout.auto(map)
  }

  if(length(new.palette) > 0){
    old.palette = palette()
    palette(new.palette) # use the palette stored in the map

    plot(map, layout = layout)
    points(layout, type = "n")
```

```
    palette(old.palette) # reset it to what it was
  } else if(length(attributes(map)$color.palette) > 0){
    old.palette = palette()
    palette(attributes(map)$color.palette) # use the palette stored in the map

    plot(map, layout = layout)
    points(layout, type = "n")

    palette(old.palette) # reset it to what it was
  } else{
    warning("no palette given, using the current environmental palette")
    plot(map, layout = layout)
    points(layout, type = "n")
  }
}

mapper.find = function(index, map, layout, attr = "members"){
  ## type can be either "data", to find index in V(map)$members, or
  ## it can be "vertex", to find index in V(map)

  ## searches through the map to find the items which match index and plots the map with
those vertices
```

```
## highlighted in color (red), everything else is made white


## save the old values for later so we can change them without breaking things
old.palette = palette()
old.colors = V(map)$color
palette(c("white", "red"))
V(map)$color = 1 # whitewash everything, then we paint the vertices we want red


if(attr == "ids"){
  v(map)[index]$color = 2
} else{
  vertex.list = NULL
  mlist = get.vertex.attribute(map, attr, V(map))
  for(i in 1:length(mlist)){
    if(sum(!is.na(match(index, mlist[[i]]))) > 0){
      vertex.list = c(vertex.list, i)
    }
  }
  V(map)[vertex.list]$color = 2
  mapper.plot(map, layout, FALSE)
}
## put the old values back where they were
V(map)$color = old.colors
```

```
  palette(old.palette)
}

mapper.identify = function(map, layout, attr = "members"){
  ## runs the identify function on a previously mapper.plotted map, using the given
layout, and returns
  ## the chosen "attr" attribute of the selected vertices. Labels are generated from the
object "map"
  #
  ## returns the ids of the selected vertices
  labels = get.vertex.attribute(map, attr)
  ids = identify(layout, labels = labels)
  return(ids)
  ## given the idrs, you can call V(map)$attr[id] to get the value of that attribute at that id,
  ## if the labels on the graph are blurry
}

mapper.layout = function(map, layout.type = "layout.auto", norm = TRUE){
  ## returns a two-column list of coordinates for each point in the map according to the
chosen layout type
  ## this layout can be independently saved to ensure a constant plot of the same map,
rather than letting
  ## R and/or igraph give you a different shape each time.
```

```
  ## only layout.auto and layout.reingold.tilford are enabled so far
  if(layout.type == "auto" || layout.type == "layout.auto"){
    layout = layout.auto(map)
    if(norm == TRUE){
      ## the natural return of the igraph maps are normalized to have coordinates between -
1 and 1
      ## this makes the layout identical in terms of the absolute numbers, rather than
relative numbers
      layout = layout.norm(layout, xmin = -1, xmax = 1, ymin = -1, ymax = 1)
    }
  } else if(layout.type == "tree" || layout.type == "layout.reingold.tilford"){
    r = which(V(map)$filter == min(V(map)$filter))
    print(r)
    layout = layout.reingold.tilford(map, root = r)
    if(norm == TRUE){
      ## the natural return of the igraph maps are normalized to have coordinates between -
1 and 1
      ## this makes the layout identical in terms of the absolute numbers, rather than
relative numbers
      layout = layout.norm(layout, xmin = -1, xmax = 1, ymin = -1, ymax = 1)
    }
  }else{
```

```
    warning("only some layouts enabled so far, returning layout.auto")
    layout = layout.auto(map)
    if(norm == TRUE){
      ## the natural return of the igraph maps are normalized to have coordinates between -
1 and 1
      ## this makes the layout identical in terms of the absolute numbers, rather than
relative numbers
      layout = layout.norm(layout, xmin = -1, xmax = 1, ymin = -1, ymax = 1)
    }
  }
  return(layout)
}

make_rough_pop_graph = function(data, filter.weights, cluster.method, N, overlap,
                                max.k, distance.matrix, constant.cut, h.list, h.true){
  labelblank = c(1:length(data[,1])) # for us to slice labels out of for the trees
  g = graph.empty(directed = FALSE) # to add vertices to later as we get clusters
  h = constant.cut
  ## counts down from the highest ??weight in n-many evenly (or as much as possible)
distributed groups
  sortweight = sort(filter.weights, decreasing = TRUE)
  nweights = length(sortweight)
```

```
  #first slice is one-sided
  rule = filter.weights >= sortweight[ceiling((1+overlap)*nweights/N)]
  if(distance.matrix == TRUE){data.section = data[rule,rule]} else{data.section =
data[rule,]}
  labels = labelblank[rule]


  ## do clustering
  if(!h.true){
    if(!constant.cut){
      h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
    }
  } else{ h = h.list[1]}
  if(is.na(h)){
    h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
    # lets you edit single points in place
  }
  hlist = c(h)
  if(distance.matrix){tree = hclust(as.dist(data.section), method = cluster.method)}
  else{tree = hclust(dist(data.section), method = cluster.method)}
  tree$labels = labels
  clusters = cutree(tree, h = h)
```

```
  ## start building the network
  g2 = graph.empty(directed=FALSE) + vertices(labels)
  #plot(g2)
  g2 = contract.vertices(g2, as.vector(clusters))
  ## add a vertex attribute to shade the vertices by the average filter value for the cut
!Problem (assign per vertex)
  for(j in V(g2)){
    names = V(g2)$name[[j]]
    V(g2)[j]$filter = mean(filter.weights[names])
  }
  V(g2)$N = 1

  g = suppressWarnings(g %du% g2) ## add on the new cluster vertices; nothing is joined
yet

  for(i in 2:(N-1)){
    top = sortweight[ceiling((i-1 - overlap)*nweights/N)]
    bottom = sortweight[ceiling((i + overlap)*nweights/N)]

    rule = filter.weights >= bottom & filter.weights <= top
    if(distance.matrix == TRUE){data.section = data[rule,rule]} else{data.section =
data[rule,]}
```

```
  labels = labelblank[rule]


  ## do clustering
  if(!h.true){
    if(!constant.cut){
      h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
    }
  } else{h = h.list[i]}
  if(is.na(h)){
    h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
    # lets you edit single points in place
  }
  hlist = c(hlist, h)
  if(distance.matrix){tree = hclust(as.dist(data.section), method = cluster.method)}
  else{tree = hclust(dist(data.section), method = cluster.method)}
  tree$labels = labels
  clusters = cutree(tree, h = h)


  ## start building the network
  g2 = graph.empty(directed=FALSE) + vertices(labels)
  #plot(g2)
```

```
    g2 = contract.vertices(g2, as.vector(clusters))
    ## add a vertex attribute to shade the vertices by the average filter value for the cut
    for(j in V(g2)){
      names = V(g2)$name[[j]]
      V(g2)[j]$filter = mean(filter.weights[names])
    }
    V(g2)$N = i
    g = suppressWarnings(g %du% g2) ## add on the new cluster vertices; nothing is
joined yet
  } # do all the middle slices

  #last slice is one-sided in the other direction
  rule = filter.weights <= sortweight[ceiling((N-1-overlap)*nweights/N)]
  if(distance.matrix ==TRUE){data.section = data[rule,rule]} else{data.section =
data[rule,]}
  labels = labelblank[rule]

  ## do clustering
  if(!h.true){
    if(!constant.cut){
      h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
    }
```

```
  } else{h = h.list[N]}
  if(is.na(h)){
    h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
    # lets you edit single points in place
  }
  hlist = c(hlist, h)
  if(distance.matrix){tree = hclust(as.dist(data.section), method = cluster.method)}
  else{tree = hclust(dist(data.section), method = cluster.method)}
  tree$labels = labels
  clusters = cutree(tree, h = h)

  ## start building the network
  g2 = graph.empty(directed=FALSE) + vertices(labels)
  g2 = contract.vertices(g2, as.vector(clusters))
  for(j in V(g2)){
    names = V(g2)$name[[j]]
    V(g2)[j]$filter = mean(filter.weights[names])
  }
  V(g2)$N = N
  g = suppressWarnings(g %du% g2) ## add on the new cluster vertices; nothing is joined
yet
```

```
  attributes(g)$hlist = hlist # i think this just vanishes into aether at the moment

  return(g)
}

connect_rough_pop_graph = function(g, data, filter.weights, N, overlap,
distance.matrix){
  ## in hindsight I don't think anything except
  ##    new.edges = vertex.overlap(g, rownames(data), "members")
  ##    g[from = new.edges[1,], to = new.edges[2,]] = TRUE
  ## is required for all of this functionality

  labelblank = c(1:length(data[,1])) # for us to slice labels out of for the trees

  ## counts down from the highest weight in n-many evenly (or as much as possible)
distributed groups
  sortweight = sort(filter.weights, decreasing = TRUE)
  nweights = length(sortweight)

  for(i in 1:(N-1)){ # for N intervals, there must be N-1 overlapping regions

    top = sortweight[ceiling((i - overlap)*nweights/N)]
    bottom = sortweight[ceiling((i + overlap)*nweights/N)]
```

```
    rule = filter.weights >= bottom & filter.weights <= top


    if(sum(rule) > 0){
      labels = labelblank[rule] # we don't need to calculate clustering on them, just grab
overlap labels
      vertex_set = vertex.overlap(g, labels, "name")
      g[from=vertex_set[1,], to=vertex_set[2,]] = TRUE


    }
  }


  return(g)
}


make_rough_size_graph = function(data, filter.weights, cluster.method, N, overlap,
                                 max.k, distance.matrix, constant.cut, h.list, h.true){
  filt.min = min(filter.weights)
  filt.int = (max(filter.weights) - filt.min)/N
  labelblank = c(1:length(data[,1])) # for us to slice labels out of for the trees
  h = constant.cut
  g = graph.empty(directed = FALSE) # to add vertices to later as we get clusters
  hlist = NULL
```

```
  for(i in 1:N){
    bottom = filt.int*(i-1 - overlap) + filt.min
    top = filt.int*(i + overlap) + filt.min

    rule = filter.weights >= bottom & filter.weights <= top
    ## if rule is empty, then no points fall in this range and we just end this iteration
    ## otherwise, behavior depends on 1 vs not-1
    if(sum(rule) > 1){  ## when there's only 1 point, the distance function breaks down on
the vector
      if(distance.matrix ==TRUE){data.section = data[rule,rule]} else{data.section =
data[rule,]}
      labels = labelblank[rule]

      ## do clustering
      if(!h.true){knn
        if(!constant.cut){
          h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
        }
      } else{ h = h.list[i]}
      if(is.na(h)){
        h = ask_h(data.section, method = cluster.method, max.k = max.k, distance.matrix =
distance.matrix)
```

```
  # lets you edit single points in place
 }
 hlist[i] = h
 if(distance.matrix){tree = hclust(as.dist(data.section), method = cluster.method)}
 else{tree = hclust(dist(data.section), method = cluster.method)}
 tree$labels = labels
 clusters = cutree(tree, h = h)

 ## start building the network
 g2 = graph.empty(directed=FALSE) + vertices(labels)
 #plot(g2)
 g2 = contract.vertices(g2, as.vector(clusters))
 ## add a vertex attribute to shade the vertices by the average filter value for the cut
 V(g2)$filter = (top + bottom) / 2
 V(g2)$N = i
 g = suppressWarnings(g %du% g2) ## add on the new cluster vertices; nothing is
joined yet
 } else if(sum(rule) == 1){ # when there's only one point we can just add it directly to
the graph
 labels = labelblank[rule]
 g2 = graph.empty(directed = FALSE) + vertex(labels)
 V(g2)$filter = (top + bottom) / 2
 V(g2)$N = i
```

```
    # no need to contract the single vertex
    g = suppressWarnings(g %du% g2)
   }
  }
  if(length(hlist) < N){
   hlist[N] = NA
  }
  attributes(g)$hlist = hlist
  return(g)
}

connect_rough_size_graph = function(g, data, filter.weights, N, overlap,
distance.matrix){
  filt.min = min(filter.weights)
  filt.int = (max(filter.weights) - filt.min)/N
  labelblank = c(1:length(data[,1])) # for us to slice labels out of for the trees
  for(i in 1:(N-1)){ # for N intervals, there must be N-1 overlapping regions
   bottom = filt.int*(i - overlap) + filt.min # this is the bottom of the /next/ interval from
above
   top = filt.int*(i + overlap) + filt.min
   rule = filter.weights > bottom & filter.weights < top

   if(sum(rule) > 0){
```

```
      labels = labelblank[rule] # we don't need to calculate clustering on them, just grab
overlap labels
      vertex_set = vertex.overlap(g, labels, "name")
      g[from=vertex_set[1,], to=vertex_set[2,]] = TRUE
    }
  }

  return(g)
}

vertex.overlap = function(graph, labels, attr, gids = FALSE){
  #new.edges = vertex.overlap(g, rownames(data), "members")
  edge.matrix = NULL
  z = get.vertex.attribute(graph, attr)
  for(i in 1:length(labels)){
    res = lapply(z, function(ch) match(labels[i], ch)) # find which nodes have the members
    vertices = which( !is.na(t(res))) # flip the weird list, pull out node indices which hit in
res
    if(length(vertices) >= 2){ # so only pairs get linked
      new.edges = combn(vertices, 2) # put all pairs of those things together to join them all
to each other
      edge.matrix = cbind(edge.matrix, new.edges) # big ol thing to hold all the new edges
    }
```

```
  }
  # hack fix
  if(gids){
    if(length(V(graph)$gid) > 0){ # make sure it doesn't break
      row1 = V(graph)$gid[edge.matrix[1,]]
      row2 = V(graph)$gid[edge.matrix[2,]]
      edge.matrix = rbind(row1, row2)
    } else{warning("No gids present, ignoring gids argument and returning vertex ids")}
  }

  return(edge.matrix)
}

ask_h = function(data, method, max.k, distance.matrix = FALSE){
  old.par = par(mfrow = c(1,2))
  if(method == "single" | method == "complete"){
    data = unique(data) # avoids weird joins at 0, doesn't
                        # work when centroids or averages used
  }
  if(distance.matrix){C = hclust(as.dist(data), method = method)
  } else{C = hclust(dist(data), method = method)
  }
  hist(C$height[C$height > 0], breaks = min(ceiling(length(C$height)/2), 20), col = "red")
```

```
  if(nrow(C$merge) > 1){
    plot(C)
  }
  h = NA
  while(is.na(h)|h <= 0){
    h = readline("What height should we cut the clusters at, according to this graph?: ")
    h = ifelse(grepl("[^0-9.]",h),-1,as.numeric(h))
  }
  if(h > max(C$height)){h = max(C$height)}

  return(h)
  par(old.par)
}

## used to remap a vector such that the maximum value is 1 and the minimum value is 0,
linear scale
scale01 = function(vector){
  vector = (vector - min(vector))/(max(vector)-min(vector))
  return(vector)
}

knn.estimate.weights = function(data, k, distance.matrix = FALSE, dist.sort = FALSE,
dim = NA){
```

```
  ## check to make sure data isn't too high-dimensional
  if(!dist.sort){

    if(distance.matrix){
      while(is.na(dim)|dim <= 0){
        dim = readline("What dimensional space is the data from? (i.e. R^n): ")
        dim = ifelse(grepl("[^0-9.]",k),-1,as.numeric(dim))
      }
    } else{dim = ncol(data)
    }
    if(dim > 341){ # 341 is experimentally the largest value that can be entered into cq
      warning("Dimensionality too large, cannot produce knn estimates on larger than 341
dimensional spaces")
      return(-1) #error code -1 = dimensionality too large
    }

    if(distance.matrix){
      d = as.dist(data)
    } else{d = dist(data, diag = TRUE, upper = TRUE)
    }
    n = nrow(as.matrix(d)) # number of data points
    Rk = NULL
    m <- data.frame(t(combn(1:n,2)), as.numeric(d))
```

```
  for(i in 1:n){
    Rk[i] = (sort(m[m[,1]==i|m[,2]==i, 3])[k])
  }
 } else{
  Rk = data[,k]
 }
 estimates = NULL
 cQ = cq(q = dim) # calculated once to save time
 Rk =

  estimates = k/(n*cQ*Rk) # formula in reference below
  # http://www.ssc.wisc.edu/~bhansen/718/NonParametrics10.pdf
 return(estimates)
}

cq = function(q){return(pi^(q/2)/gamma((q+2)/2))} # used in the above function
 # is the volume of a q-dimensional unit ball

knn.weights = function(data, k, distance.matrix = FALSE){
 if(distance.matrix){
  d = as.dist(data)
 } else{d = dist(data, diag = TRUE, upper = TRUE)
 }
```

```
  len = length(data[,1])
  weights = NULL

  m <- data.frame(t(combn(1:len,2)), as.numeric(d))
  for(i in 1:len){
    weights[i] = sum(sort(m[m[,1]==i|m[,2]==i, 3])[1:k])
  }
  return(weights)
}

dist.sort = function(data, distance.matrix = FALSE){
  if(distance.matrix){
    d = as.dist(data)
  } else{d = dist(data, diag = TRUE, upper = TRUE)
  }
  len = length(data[,1])
  weights = NULL

  m <- data.frame(t(combn(1:len,2)), as.numeric(d))
  for(i in 1:len){
    weights = rbind(weights, sort(m[m[,1]==i|m[,2]==i, 3]))
  }
  return(weights)
```

```
}

dist.corr = function(x, y, theta){ ## this function doesn't work yet
  dsq = 0
  dsq = ((x[1]+x[2]*cos(theta))-(y[1]+y[2]*cos(theta))) + (x[2] - y[2])
  d = sqrt(dsq)
  return(d)
}

find.slice = function(data, distance.matrix = FALSE){
  hist(knn.weights(data, distance.matrix), col = "red", main = "Distance to nearest
neighbor")
}

ask.k = function(data, k.low = 4, k.high = 15, distance.matrix = FALSE){ ## presents a
series of histograms for the user to pick the best dist
  if(distance.matrix){
    d = as.dist(data)
    len = length(attributes(d)$Labels)
  } else{
    d = dist(data, diag = TRUE, upper = TRUE)
    len = nrow(data)
  }
```

```
  weights = NULL

  m <- data.frame(t(combn(1:len,2)), as.numeric(d))
  for(i in 1:len){
    weights = rbind(weights, sort(m[m[,1]==i|m[,2]==i, 3]))
  }

  k = 0
  par(mfrow=c(2,3)) # change the plotting window so we can see many graphs at once
  if(k.low < 2){ k.low = 2} # rowSums breaks if we hand it only a single column
  for(i in k.low:k.high){
    hist(rowSums(weights[,1:i]),
         main = paste("Sum of distances to the ", i, "-th nearest neighbors", sep = ""),
         col = "red")
  }
  while(is.na(k)|k <= 0){
    k = readline("How many neighbors should be use according to these graphs?: ")
    k = ifelse(grepl("[^0-9.]",k),-1,as.numeric(k))
  }
  par(mfrow=c(1,1)) # put the plotting window back because this would suck for other
plots
  return(k) # fix the plot frame so it doesn't print 2x3 anymore
```

```
}

id.to.responses = function(ids, map, imp.data, response.data){
  # sadly need to give the function all of the objects to chain together
  imp.index = unique(unlist((V(map)$members)[ids]))
  response.index = imp.index
  return(response.data[response.index,])
}

member.to.responses = function(members, imp.data, response.data){
  # basically the function above, started halfway through, depending on
  # whether we have the node IDs or the member indices
  imp.index = unique(members)
  response.index = as.integer(rownames(imp.data)[imp.index])
  return(response.data[response.index,])
}

write.Gephi = function(my.map, filepattern){
  if(length(unique(unlist(V(my.map)$name))) != length(unlist(V(my.map)$name))){
    V(my.map)$old.names = V(my.map)$name
    V(my.map)$name = c(1:length(unlist(V(my.map)$name)))
    warning("Wrote old names to attribute $old.names, new names given")
  }
```

```
  my.df.v = get.data.frame(my.map, what = "vertices")
  my.df.v = my.df.v[,!grepl("members$", colnames(my.df.v))]
  colnames(my.df.v)[1] = c("Id")
  my.df.e = get.data.frame(my.map, what = "edges")
  if(ncol(my.df.e) > 2){
    colnames(my.df.e)[1:3] = c("Source", "Target", "Weight")
  } else{
    colnames(my.df.e)[1:2] = c("Source", "Target")
  }

  node.path = paste(filepattern, ".node.list.csv", sep = "")
  edge.path = paste(filepattern, ".edge.list.csv", sep = "")
  write.table(as.matrix(my.df.v), file = node.path, row.names = FALSE, sep = ",")
  write.table(as.matrix(my.df.e), file = edge.path, row.names = FALSE, sep = ",")
}

inject.attribute = function(map, truths, name, rel.per = NULL){
  for(i in 1:length(V(map))){
    subtruths = truths[V(map)$members[[i]]]
    raw = sum(subtruths, na.rm = T)
    present = raw > 0
    percent = raw / V(map)$membersize[[i]]
    if(length(rel.per) > 0){
```

```
      relative = percent/rel.per
    } else{
      relative = NULL
    }

    map = set.vertex.attribute(map, paste(name, '_present', sep = ""), i, present)
    map = set.vertex.attribute(map, paste(name, '_raw', sep = ""), i, raw)
    map = set.vertex.attribute(map, paste(name, '_percent', sep = ""), i, percent)
    map = set.vertex.attribute(map, paste(name, '_relative', sep = ""), i, relative)
  }
  return(map)
}

# id.to.responses = function(ids, map, imp.data, response.data){
#   # sadly need to give the function all of the objects to chain together
#   imp.index = unique(unlist((V(map)$members)[ids]))
#   response.index = imp.index
#   return(response.data[response.index,])
# }
```

Factor Analysis.Rmd

---

title: "InICE Factor Descriptions"

author: "Jackie Doyle"

date: "Wednesday, October 28, 2015"

output: pdf_document

Warning, this code overwrites the fa() function because of bad namespacing.

\section{Changelog}

2/27/2017 - added (in comments) the factor names to the question patterns so it's easier to link individual questions to their eventual factors.

3/16/16 - changes made to the cleaning file reduced N from 2966 to 2916. Added MLE imputation before factor space creation.

12/2/15 - Merged changes to factor names, altered final chunk to produce a factor space and a factor subspace congruent with our decision on which 13 factors we wanted to use. Adjusted a couple ofthe factor loading patterns after removing the core identity questions (8a, 8c, 7Pa, 7Ma) and rerunning the factor analysis there; there is some disagreement here?

12/14/15 - Changed negatively worded items to reverse coded: question (3t, 3u,3v, 3w,3x, 5b, 5c, 5e, 5f, 10c, 10h, 10k, 10a, 10q, 10d, 10e, 10o) Changed previously negatively worded items to regularly coded (10j, 10m). This was because 10j, and 10m are measuring extroversion and not introversion like was thought before. Changed the version number and initials in the file name. Coding at the end to flip extroversion codinghas been turned to comments because the negative coding has been reversed in the beginning of the code.

\section{Code}

If we naively take the descriptions of the factors that we made with the pilot information, we can translate those to the questions presented in the current form of the survey in the form of the following patterns

```{r}
setwd("C:/Users/Jackie/Downloads/InICE Data Files")
load(file = "INICE_v2.RData")
source("C:/Users/Jackie/Dropbox/R Files/custom.R")
library(psych)
library(moments)
library(nFactors)
library(corrplot)
library(MissMech)
```

Added 3/16/16: do an imputation of the quantitative data, using all the quantitative (but not qualitative/demographic data). We would use Q14, except the imputation does not converge then, so we just use 1-10.

```{r}
quant.Qs =grepl("^Q(1|2|3|4|5|6|7|8|9|10)[EPM]*[a-z]", colnames(INICE$data))
#however, Q8Eng_e and Q8Eng_f are colinear, so we need to remove one before imputing.
#quant.Qs = quant.Qs & !grepl("Q8Eng_f", colnames(INICE$data))
quant.data = INICE$data[,quant.Qs]

qImp = Impute(as.matrix(quant.data), imputation.method = "dist.free")
# this imputes data outside the allowable range
# and which is not integer
# so we round and crop it
qdat = round(qImp$yimp)
qdat[qdat > 6] = 6
qdat[qdat < 0] = 0

copy = INICE$data # to safely make changes
copy[,quant.Qs] = qdat # replace those columns with imputed data
```

```
#copy$Q8Eng_f = copy$Q8Eng_e
```

```{r}
factor.patterns = c("^Q1[a-z]$", # f1,1            Belongingness
                    # f1,2 these questions were cut
                    # f1,3 these questions were cut
                    "^Q2(a|b|c|d)$", # f2,1        Performance Approach
                    "^Q2(e|f|g)$", # f2,2          Mastery Approach
                    "^Q2(h|i|j)$", # f2,3          Work Avoid
                    "^Q3(j|k|l|m|n)$", # f3,1       Expectancy
                    "^Q3(t|u|v|w|x)$", # f3,2       Connectedness
                    "^Q3(a|f|g)$", # f3,3          Instrumentality
                    "^Q3(o|p|q|r|s)$", # f3,4       Value
                    "^Q3(b|c|d|e)$", # f3,5         Perceptions of Future
                    # f3,6, bad factor with only two questions, one of which is missing
                    # f4s, all cut, potential "demographic" variables
                    "^Q5(a|d|g|h)$", # f5,1         Grit: PoE
                    "^Q5(b|c|e|f)$", # f5,2         Grit: CoI
                    "^Q8Eng_(k|l|m|n|o)$", # f6,1    EID: P/C
                    "^Q8Eng_(d|e|f|g)$", # f6,2     EID: R
```

```
                "^Q8Eng_(b|h|i|j)$", # f6,3     EID: I
                "^Q9(f|g|h|i|j)$", # f7,1       Eng.AB
                "^Q9(a|b|c|d|e)$", # f7,2       Sci.AB
                "^Q10(u|x|v|p|s)$", # f8,1      Neuroticism
                "^Q10(m|o|e|j|d)$", # f8,2      Extraversion
                "^Q10(g|f|b|l|i)$", # f8,3      Agreeableness
                "^Q10(k|c|a|h)$", # f8,4  #     Conscientiousness
                "^Q10(t|r|q|w)$", # f8,5 #      Openness to Experience
                "^Q7Phys_(i|j|k|l|n)$", # f9,1  PID: PC
                "^Q7Phys_(b|c|d|e|m)$", # f9,2  PID: R
                "^Q7Phys_(f|g|h)$", # f9,3      PID: I
                "^Q7Math_(b|c|d|e|m)$", # f10,1 MID: PC
                "^Q7Math_(i|j|k|l|n)$", # f10,2 MID: R
                "^Q7Math_(f|g|h)$") # f10,3     MID: I
# these items were negatively coded in their respective factors
negative.items = c("Q3h",
"Q10n","Q3t","Q3u","Q3v","Q3w","Q3x","Q5b","Q5c","Q5e","Q5f","Q10c","Q10h","Q
10k","Q10a","Q10d","Q10e","Q10o","Q10q")

names = colnames(INICE$data)
```

With these tools set up, we can build our data frame of factor-scores.

```{r}
# copy = INICE$data # because we're gonna alter it
copy[,negative.items] = 6-copy[,negative.items] # now they're "positive"

for(pattern in factor.patterns){
  copy[,pattern] = rowMeans(copy[,grepl(pattern, names)]) # leaving na.rm = FALSE
}

factor.names  = c("Belongingness","Performance Approach", "Mastery Approach","Work Avoid", "Expectancy",  "Connectedness", "Instrumentality", "Value", "Perceptions of Future", "Grit: Persistence of Effort", "Grit: Consistency of Int", "Eng ID: Perf\\Comp", "Eng ID: Recognition", "Eng ID: Interest", "Eng AB", "Sci AB", "Neuroticism", "Extroversion", "Agreeableness", "Conscienciousness",  "Openness", "Phys ID: Perf\\Comp", "Phys ID: Recognition", "Phys ID: Interest", "Math ID: Recognition", "Math ID: Perf\\Comp", "Math ID: Interest")

factor.space = copy[,factor.patterns]
colnames(factor.space) = factor.names # make it pretty

# save(factor.space, file = "factor.space_V2.RData")

factor.cor = cor(copy[,factor.patterns], use = "pairwise.complete.obs")
```

```
# so the later correlation plot looks pretty and is readable
attributes(factor.cor)$dimnames[[1]] = factor.names
attributes(factor.cor)$dimnames[[2]] = factor.names
corrplot(factor.cor, order = "hclust", hclust.method = "complete")
```

What about the factor loadings for these new factors?

```{r, cache = TRUE}
cutoff = 0.4
question.patterns = c("^Q2[a-z]$", "^Q3[a-z]$", "^Q5[a-z]$", "^Q8Eng_(b|[de]|[g-z])$",
"^Q9[a-z]$", "^Q10[a-z]$", "^Q7Phys_[b-z]$", "^Q7Math_[b-z]$")
numfact = rbind(3, 5, 2, 3, 2, 5, 3, 3)
rownames(numfact) = question.patterns
fa = NULL

for(pattern in question.patterns){
  mydata = na.omit(copy[,grepl(pattern, colnames(copy))])

  ev <- eigen(cor(mydata)) # get eigenvalues
  ap <- parallel(subject=nrow(mydata),var=ncol(mydata), rep=100,cent=.05)
  nS <- nScree(x=ev$values, aparallel=ap$eigen$qevpea)
  plotnScree(nS)
```

```
  fa[[pattern]]= factanal(mydata, factors = numfact[pattern,], rotation = "promax")
  print(fa[[pattern]]$loadings, cutoff = cutoff, sort = FALSE)
}

## do the belongingness factor with just 1
mydata = na.omit(copy[,grep(factor.patterns[1], colnames(copy))])
ap <- parallel(subject=nrow(mydata),var=ncol(mydata), rep=100,cent=.05)
nS <- nScree(x=ev$values, aparallel=ap$eigen$qevpea)
plotnScree(nS)

fa[[factor.patterns[[1]]]]= factanal(mydata, factors = 1, rotation = "promax")
print(fa[[factor.patterns[1]]]$loadings, cutoff = cutoff, sort = FALSE)
```

And then take these loadings and find the average loading value for all of the items which loaded higher than the cutoff for each item.

```{r, cache = TRUE}
i = 2 # because Q1 has only one factor, so we skipped it in the factor analysis
average.loadings = NULL
for(qpattern in question.patterns){
  for(n in 1:numfact[qpattern,]){
```

```
    # need to see which factor it is
    rownumber = which(fa[[qpattern]]$loadings[,n] == max(fa[[qpattern]]$loadings[,n]))
    rowname = rownames(fa[[qpattern]]$loadings)[rownumber]
    qletter = strsplit(rowname, split = "")[[1]][length(strsplit(rowname, split = "")[[1]])]
    which.fact = factor.patterns[i:(i+numfact[qpattern,]-1)][grepl(qletter,
factor.patterns[i:(i+numfact[qpattern,]-1)])]
    # there MUST be an easier way to do this... but now we have which factor it is

    highloads = abs(fa[[qpattern]]$loadings[,n]) > cutoff
    average.loadings[which.fact] = mean(abs(fa[[qpattern]]$loadings[highloads,n]))
  }
  i = i + numfact[qpattern,]
}
name.pairs = cbind(factor.patterns, factor.names)
rownames(name.pairs) = factor.patterns
print(average.loadings)
name.loadings = NULL
for(pattern in factor.patterns){
  name.loadings[name.pairs[pattern, 2]] = average.loadings[pattern]
}

print(name.loadings)
```

```
hist(average.loadings)
```

We can look at the item reliability using Cronbach's alpha for each factor taken individually.

```{r, cache = TRUE}
names = colnames(copy)
alphas = NULL
for(i in 1:length(factor.patterns)){
  cols = grepl(factor.patterns[i], names)
  subframe = copy[,cols]
  alphas[[i]] = alpha(subframe, title = factor.patterns[i], check.keys =FALSE)
  print(alphas[[i]])
}
```

We can now repeat the same sort of procedure we did with the questions as a whole, but now with the factors we just created. Here's a bunch of histograms plus skew/kurtosis.

\newpage

```{r, cache = TRUE}
```

```
breaks = c(-0.5, 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5)
for(i in factor.patterns){
    hist(copy[,i], breaks = breaks, xlab = i, main = paste("Histogram of ", name.pairs[i,2],
sep = ""))
    print(paste("Skewness:",skewness(copy[,i], na.rm = T)))
    print(paste("Kurtosis:",kurtosis(copy[,i], na.rm = T)))
}
```

```{r, cache=TRUE, echo = FALSE}
old.par = par(mfrow = c(2,2))

for(i in factor.patterns){
    for(school in c("Purdue", "FIU", "Clemson", "UNR")){
    hist(copy[copy$school == school,i], breaks = breaks, xlab = i, main =
paste("Histogram of ", i, " at ", school, sep = "") )
    }
    for(school in c("Purdue", "FIU", "Clemson", "UNR")){
    print(school)
    print(paste("Skewness:",skewness(copy[copy$school == school, i], na.rm = T)))
    print(paste("Kurtosis:",kurtosis(copy[copy$school == school, i], na.rm = T)))
    print("")
    }
```

```
    for(j in 1:10){
      print("") # needed for proper spacing
    }



}

par(old.par)
```

We can also see if any of the schools have different distributions on any of these factors.

```{r}
p.vals = NULL
for(i in factor.patterns){
  w.list = list(copy[copy$school == "Clemson", i],
                copy[copy$school == "FIU", i],
                copy[copy$school == "Purdue", i],
                copy[copy$school == "UNR", i])
  print(i)
  kwt = kruskal.test(w.list)
  print(kwt)
  p.vals = c(p.vals, kwt$p.value)
```

```
}

print(p.vals)

print(p.vals < 0.05)
```

```
# We flip the Extroversion factor since it was originally printed as a measure of

introversion, and then store the stuff to save/do stuff with.

#

# ```{r}

# factor.space = copy[,factor.patterns]

# colnames(factor.space) = factor.names

# factor.space[,"Extroversion"] = 6 - factor.space[,"Extroversion"]

```{r}

subspace = factor.space[c(8, 4, 6, 9, 17, 18, 1, 2, 7, 11, 12, 13, 23)]

```
```

Group Differences.Rmd

```
---
title: "Group Descriptions"
author: "Jackie Doyle"
date: "February 8, 2016"
output:
  pdf_document:
    fig_width: 8
    fig_height: 6
    fig_caption: true
geometry: margin=0.5in
---

Here's a bunch of descriptions of the various groups we've found, as they're positioned relative to the normative group. We begin with a description of the normative group.

```{r}
# this has more than is required, but oh well
setwd('C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Participant Selection')
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Mapping/env_v2_group_comparisons_done.RData")
```
```

load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Raw data/INICE_v2.RData")

library(fmsb)

library(igraph)

```

```{r}

# build the raw data frames

for(i in 1:length(SN$ids)){

  SN$raw[[i]] = INICE$data[rownames(SN$factors[[i]]),]

}

for(i in 1:length(NG$ids)){

  NG$raw[[i]] = INICE$data[rownames(NG$factors[[i]]),]

}

for(i in 1:length(BG$ids)){

  BG$raw[[i]] = INICE$data[rownames(BG$factors[[i]]),]

}

for(i in 1:length(OS$ids)){

  OS$raw[[i]] = INICE$data[rownames(OS$factors[[i]]),]

}

# in case we didn't build snail DF

SN$frame = data.frame(SN$centroid)

```
colnames(SN$frame) = c("SN_1", "SN_2", "SN_3")

NG$frame = data.frame(NG$centroid)

colnames(NG$frame) = c("NG_1", "NG_2")

BG$frame = data.frame(BG$centroid)

colnames(BG$frame) = c("BG_1", "BG_2", "BG_3", "BG_4")

OS$frame = data.frame(OS$centroid)

colnames(OS$frame) = c("OS_1", "OS_2", "OS_3")

maxes = c(rep(6, times = NC))

mins = c(rep(0, times = NC))

star.frame = data.frame(t(data.frame(NG$frame, SN$frame, BG$frame, OS$frame)))

starplus = data.frame(rbind(maxes, mins, star.frame))

palette(rainbow(nrow(star.frame)))

radarchart(starplus, seg = 6,

        pcol = c(1:nrow(star.frame)),

        maxmin= TRUE, centerzero = TRUE, plwd = 3)

legend(x = 1.2, y = -.2, legend = c(rownames(star.frame)), fill = c(1:nrow(star.frame)))

starorder = starplus[,order(NG$centroid[[1]])]

radarchart(starorder, seg = 6, pcol = c(1:nrow(star.frame)), maxmin = TRUE, centerzero

= TRUE, plwd =3)

legend(x = 1.20, y = -.2, legend = rownames(star.frame), fill = c(1:nrow(star.frame)))
```

```

\newpage

\section{The Normative Group}

The normative group is charcterized by the following:

* High scores on Instrumentality, Perceptions of Future, Connectedness, and Belongingness.
* Average scores on Extroversion, Grit:CoI, Performance Approach, PhysID:Rec, Value, EngID:Rec, and EngID:Perf/Comp.
* Low scores on Neuroticism, and Work Avoid.

```{r}
barplot(NG$centroid[[1]][order(NG$centroid[[1]])])
barplot(NG$centroid[[2]][order(NG$centroid[[2]])])
# This is the order the factors will be
# presented in all throughout this document
# (since the legend under the bars doesn't
# always show the entire list of names)
the.order = order(NG$centroid[[1]])
the.names = names(NG$centroid[[1]])[the.order]
the.names
```
```

```
shortnames = c("WA", "Neurot", "Extr", "Grit:CoI", "PA", "PhysID:R", "Val",
"EngID:R", "EngID:PC", "Belong", "Conn", "PoFut", "Instr")

nrow(NG$raw[[1]]) # Total number of students
nrow(NG$raw[[2]])
table(NG$raw[[1]]$school) # Breakdown of students by school
table(NG$raw[[1]]$school)/ nrow(NG$raw[[1]]) # percentage of students by school
table(NG$raw[[2]]$school) # Breakdown of students by school
table(NG$raw[[2]]$school)/ nrow(NG$raw[[2]]) # percentage of students by school

table(NG$raw[[1]]$school[NG$raw[[1]]$has.email == TRUE]) # with emails
table(NG$raw[[2]]$school[NG$raw[[2]]$has.email == TRUE]) # with emails
```

\newpage
\section{Supernormal Group 3}
Branching group one has MORE

* ($p < 0.001$) Value
* ($p < 0.01$) PhysID: Rec
* ($p < 0.05$) Perceptions of Future, Belongingness, EngID:PC

Branching group one has LESS

* ($p < 0.05$) Neurotic

```{r}
barplot(as.matrix(starorder[7,] - starorder[3,]), names.arg = shortnames, cex.names = .3)

# NG is blue
radarchart(starorder[c(1:2, 3, 7),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,
centerzero = TRUE, plwd =3)
```

Population statistics for supernormal group 3

```{r}
nrow(SN$raw[[3]]) # Total number of students
table(SN$raw[[3]]$school) # Breakdown of students by school
table(SN$raw[[3]]$school)/ nrow(SN$raw[[3]]) # percentage of students by school

table(SN$raw[[3]]$school[SN$raw[[3]]$has.email == TRUE]) # with emails
```

\newpage
\section{Branching Group 1}

Branching group one has MORE

* ($p < 0.05$) Engineering ID: Recognition

Branching group one has LESS

* ($p < 0.001$) Value
* ($p < 0.05$) Performance Approach

```{r}
barplot(as.matrix(starorder[8,] - starorder[3,]), names.arg = shortnames, cex.names = .3)

# NG is blue
radarchart(starorder[c(1:2, 3, 8),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,
centerzero = TRUE, plwd =3)
```

Population statistics for branching group 1

```{r}
nrow(BG$raw[[1]]) # Total number of students
table(BG$raw[[1]]$school) # Breakdown of students by school
table(BG$raw[[1]]$school)/ nrow(BG$raw[[1]]) # percentage of students by school
```

```
table(BG$raw[[1]]$school[BG$raw[[1]]$has.email == TRUE]) # with emails
```

\newpage

\section{Branching Group 2}

Branching group two has MORE

* ($p < 0.05$) Neuroticism

Branching group two has LESS

* ($p < 0.001$) Value, Connectedness, Eng ID: Rec
* ($p < 0.01$) Phys ID: Rec
* ($p < 0.05$) Perceptions of Future, Extroversion, Eng ID: PC

```{r}
barplot(as.matrix(starorder[9,] - starorder[3,]), names.arg = shortnames, cex.names = .3)

# NG is blue
radarchart(starorder[c(1:2, 3, 9),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,
centerzero = TRUE, plwd =3)
```

```

Population statistics for branching group 2

```{r}
nrow(BG$raw[[2]]) # Total number of students
table(BG$raw[[2]]$school) # Breakdown of students by school
table(BG$raw[[2]]$school)/ nrow(BG$raw[[2]]) # percentage of students by school

table(BG$raw[[2]]$school[BG$raw[[2]]$has.email == TRUE]) # with emails
```

\newpage
\section{Branching Group 3}
Branching group three has LESS

* (p < 0.001) Performance Approach, Grit: CoI
* (p < 0.01) Connectedness
* (p < 0.05) Extroversion

```{r}
barplot(as.matrix(starorder[10,] - starorder[3,]), names.arg = shortnames, cex.names = .3)
```
```

radarchart(starorder[c(1:2, 3, 10),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,

centerzero = TRUE, plwd =3)

```

Population statistics for branching group 3

```{r}

nrow(BG$raw[[3]]) # Total number of students

table(BG$raw[[3]]$school) # Breakdown of students by school

table(BG$raw[[3]]$school)/ nrow(BG$raw[[3]]) # percentage of students by school

table(BG$raw[[3]]$school[BG$raw[[3]]$has.email == TRUE]) # with emails

```

\newpage

\section{Branching Group 4}

Branching group four has MORE

* ($p < 0.01$) Grit: CoI

Branching group four has LESS

* ($p < 0.001$) Eng ID: Rec
```

* (p < 0.01) Extroversion, Performance Approach

* (p < 0.05) Eng ID: PC, Work Avoid, Neuroticism

```{r}
barplot(as.matrix(starorder[11,] - starorder[3,]), names.arg = shortnames, cex.names = .3)

radarchart(starorder[c(1:2, 3, 11),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,
centerzero = TRUE, plwd =3)
```

Population statistics for branching group 3

```{r}
nrow(BG$raw[[4]]) # Total number of students
table(BG$raw[[4]]$school) # Breakdown of students by school
table(BG$raw[[4]]$school)/ nrow(BG$raw[[4]]) # percentage of students by school

table(BG$raw[[4]]$school[BG$raw[[4]]$has.email == TRUE]) # with emails
```

\newpage

\section{Outlier Set 1}

Outlier set 1 has MORE

* (p < 0.001) Extroversion

Outlier set 1 has LESS

* (p < 0.001) Recognition

* (p < 0.01) Value

```{r}
barplot(as.matrix(starorder[12,] - starorder[3,]), names.arg = shortnames, cex.names = .3)

# NG is blue
radarchart(starorder[c(1:2, 3, 12),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,
centerzero = TRUE, plwd =3)
```

Population statistics for outlier set 1

```{r}
nrow(OS$raw[[1]]) # Total number of students
table(OS$raw[[1]]$school) # Breakdown of students by school
table(OS$raw[[1]]$school)/ nrow(OS$raw[[1]]) # percentage of students by school
```

table(OS$raw[[1]]$school[OS$raw[[1]]$has.email == TRUE]) # with emails

```

\newpage

\section{Outlier Set 2}

Outlier set 2 has MORE

* (p < 0.001) Eng ID: PC

* (p < 0.05) Belongingness, Instrumentality

Outlier set 2 has LESS

* (p < 0.001) Eng ID: Rec

* (p < 0.05) Phys ID: Rec

```{r}

barplot(as.matrix(starorder[13,] - starorder[3,]), names.arg = shortnames, cex.names = .3)

# NG is blue

radarchart(starorder[c(1:2, 3, 13),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,

centerzero = TRUE, plwd =3)

```
```

Population statistics for outlier set 2

```{r}
nrow(OS$raw[[2]]) # Total number of students
table(OS$raw[[2]]$school) # Breakdown of students by school
table(OS$raw[[2]]$school)/ nrow(OS$raw[[2]]) # percentage of students by school

table(OS$raw[[2]]$school[OS$raw[[2]]$has.email == TRUE]) # with emails
```

\newpage

\section{Outlier Set 3}

Outlier set 3 has MORE

* ($p < 0.01$) Connectedness

Outlier set 3 has LESS

* ($p < 0.001$) Eng ID: Rec, Work Avoid
* ($p < 0.05$) Phys ID: Rec, Eng ID: PC

```{r}
barplot(as.matrix(starorder[14,] - starorder[3,]), names.arg = shortnames, cex.names = .3)
```

```
# NG is blue
radarchart(starorder[c(1:2, 3, 14),], pcol = rainbow(8)[c(6, 1)], maxmin = TRUE,
centerzero = TRUE, plwd =3)
```

Population statistics for outlier set 3

```{r}
nrow(OS$raw[[3]]) # Total number of students
table(OS$raw[[3]]$school) # Breakdown of students by school
table(OS$raw[[3]]$school)/ nrow(OS$raw[[3]]) # percentage of students by school

table(OS$raw[[3]]$school[OS$raw[[3]]$has.email == TRUE]) # with emails
```

```{r}
# for(i in 1:13){
#   print(colnames(norm.data[i]))
#   print(t.test(men[,i], norm.data[,i])$p.value)
#   print(t.test(women[,i], norm.data[,i])$p.value)
# }
```

```

Make new group names for stuff, store.

```{r updated group data}
ids = NULL

factors = NULL

centroid = data.frame()

raw = NULL


ids$NG = c(NG$ids[[1]], NG$ids[[2]], SN$ids[[1]], SN$ids[[2]], SN$ids[[3]])

#unique to get rid of the one duplicated row

factors$NG = rbind(NG$factors[[1]], NG$factors[[2]], SN$factors[[1]], SN$factors[[2]],

SN$factors[[3]])

factors$NG = factors$NG[unique(rownames(factors$NG)),]

centroid = data.frame(colMeans(factors$NG))


NG.names = rownames(id.to.responses(ids$NG, chosen.map, INICE$data, INICE$data))


raw$NG = INICE$data[NG.names,]


for(i in 1:4){

  ids$NnG[[i]] = BG$ids[[i]]

  factors$NnG[[i]] = BG$factors[[i]]
```
```

```
  centroid = cbind(centroid, BG$centroid[i])
  BG.names = rownames(BG$factors[[i]])
  raw$NnG[[i]] = INICE$data[BG.names,]
}

for(i in 1:3){
  ids$NnG[[i+4]] = OS$ids[[i]]
  factors$NnG[[i+4]] = OS$factors[[i]]
  centroid = cbind(centroid, OS$centroid[i])
  OS.names = rownames(OS$factors[[i]])
  raw$NnG[[i+4]] = INICE$data[OS.names,]
}

all.rows = c(rownames(factors$NG), unlist(lapply(factors$NnG, rownames)))
potential.rows = rownames(INICE$data)
DG.rows = setdiff(potential.rows, all.rows)
# ids$DG = potential.ids[!potential.ids %in% all.ids]  ## too many missing from map to
be sensical
factors$DG = factor.subspace[DG.rows,]
raw$DG = INICE$data[DG.rows,]

colnames(centroid) = c("NG", "NnG_1", "NnG_2", "NnG_3", "NnG_4", "NnG_5",
"NnG_6", "NnG_7")
```

```
group.data = NULL

group.data$ids = ids

group.data$factors = factors

group.data$centroid = centroid

group.data$raw = raw


save(group.data, file = "C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study

Data/Mapping/groupdata.RData")
```

IVA.Rmd

```
---
title: "Identity vs Attitude"
author: "Jackie Doyle"
date: "April 1, 2016"
output: pdf_document
---

```{r load data}
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Raw data/INICE_v2.RData")
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Major-Question cleaning/major.copy_v2.RData")
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Major-Question cleaning/major.matrix_v2.RData")

load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Mapping/factor.space_v2.RData")
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Major-Question cleaning/major.zero_v2.RData")
source("C:/Users/Jackie/Dropbox/R Files/custom.R")
library(car)
library(lm.beta)
```
```

```
library(corrplot)
```

```{r make interests}
interests = INICE$data[,grepl("Q14", colnames(INICE$data))]
attitudes = factor.space
sub.names = c("Value", "Work Avoid", "Connectedness", "Perceptions of Future",
"Neuroticism", "Extroversion", "Belongingness", "Performance Approach",
"Instrumentality", "Grit: Consistency of Int", "Eng ID: Perf\\Comp", "Eng ID:
Recognition", "Phys ID: Recognition")
factor.subspace = factor.space[,sub.names]
sub.attitudes = factor.subspace
```

Our column names for the factor space are not pretty things because they often consist of multiple words. This destroys things like "auto-make a formula" and other stuff that tries to use the name as a variable in a formula. So we replace all the names with nicer names for what we're doing. To be explicit about the changes, we'll write them all out using recode.

```{r make attitude/interest frames}
colnames(attitudes) = c("Belongingness", "Perform.App", "Mastery.App",
"Work.Avoid", "Expectancy", "Connectedness", "Instrumentality", "Value",
```

```
"Perf.of.Future", "Grit.PoE", "Grit.CoI", "EngID.PC", "EngID.Rec", "EngID.Int",
"Eng.AB", "Sci.AB", "Neuroticism", "Extroversion", "Agreeableness",
"Conscientiousness", "Openness", "PhysID.PC", "PhysID.Rec", "PhysID.Int",
"MathID.Rec", "MathID.PC", "MathID.Int")
colnames(interests) = c(
  "AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE_CSE", "CON", "EE", "EP",
"EEE", "IND", "IT", "MSE", "ME", "MIE", "NUKE", "OSTEM", "ONON"
)
class(interests) = "data.frame"
class(attitudes) = "data.frame"
IVA.frame = cbind(interests, attitudes)
IVS.frame = cbind(interests, sub.attitudes)
```

Here are some diagnostic plots about the distribution of interests. We want to see how normally distributed they are, and how bimodal the distribution is. For the most part, they all are, though the second (ABE) and third (BME) are distinct.

```{r visualize interest plots}
ikde.list = NULL
for(i in 1:18){
  hist(na.omit(interests[,i]), breaks = seq(-0.5, 6.5, by=1), main = colnames(interests)[i],
freq = FALSE, xlim = c(0, 6), ylim = c(0, .3))
```

```
  ikde = density(interests[,i], na.rm = T, bw = .5, from = 0, to = 6)
  ikde.list[[i]] = ikde$y
  points(cbind(ikde$x, ikde.list[[i]]), type = "l", col = "red", lwd = 2)
  #plot(ikde, main = colnames(interests)[i], zero.line = TRUE, xlab = )
}

plot(cbind(ikde$x, ikde.list[[17]]), type = "l", col = 1, lwd = 2, xlim = c(0, 6), ylim = c(0, .3))
for(i in 1:18){
  points(cbind(ikde$x, ikde.list[[i]]), type = "l", col = i, lwd = 2, lty = i %% 5 + 1)
}
```

The majors that make up the interest questions are not quite the same as the ones that make up our major matrix, as determined by Jackie when she cleaned up the list. To bring them more in line with each other, we do some overlap merging, and trim a couple answers off to be dealt with by hand later (especially important since that includes the General Engineering, First Year Engineering, and Exploratory Engineering groups).

```{r clean demographics}
#
# ABE_BSEdemo = ifelse(major.zero$ABE == 1 | major.zero$BSE == 1, 1, 0)
# BE_BMEdemo = ifelse(major.zero$BE == 1 | major.zero$BME == 1, 1, 0)
```

```
# CE_CSEdemo = ifelse(major.zero$CE == 1 | major.zero$CSE == 1, 1, 0)
# EE_ECEdemo = ifelse(major.zero$EE == 1 | major.zero$ECE == 1, 1, 0)
#
# crossmatrix = cbind(interests[,c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL",
"CE_CSE", "EE", "IND", "MSE", "ME", "NUKE")], major.zero$AAE, ABE_BSEdemo,
BE_BMEdemo, major.zero$CME, major.zero$CVL, CE_CSEdemo, EE_ECEdemo,
major.zero$IND, major.zero$MSE, major.zero$ME, major.zero$NUKE)
# # oh god the column names
#
# newnames = c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE_CSE",
"EE_ECE", "IND", "MSE", "ME", "NUKE")
# demonames = lapply(newnames, paste, "demo", sep = "")
# colnames(crossmatrix) = c(newnames, demonames)
# sub.int = interests[,c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE_CSE",
"EE", "IND", "MSE", "ME", "NUKE")]
# sub.demo = crossmatrix[,unlist(demonames)]
# ```
#
# We also need total measures for Engineering and Physics identity.
#
# ```{r make identity measures}
EngID = rowMeans(attitudes[,c("EngID.PC", "EngID.Rec", "EngID.Int")])
PhysID = rowMeans(attitudes[,c("PhysID.PC", "PhysID.Rec", "PhysID.Int")])
```

```
MathID = rowMeans(attitudes[,c("MathID.PC", "MathID.Rec", "MathID.Int")])
#
no.frame = attitudes[,!grepl("ID", colnames(attitudes))]
no.form = paste(names(no.frame), "+ MathID", collapse = " + ")
ID.frame = cbind(EngID, PhysID, MathID, no.frame)
#
# ```
#
# We can really only do regressions with sufficiently large populations. So we take all of the major groupings with >100 students (lowest is 121, next lower is 71) and make subpopulations to study.
#
# ```{r demographic regression with large populations}
# large.pop = cbind(sub.demo$BE_BMEdemo, sub.demo$CMEdemo,
sub.demo$CVLdemo, sub.demo$CE_CSEdemo,
#                   sub.demo$EE_ECEdemo, sub.demo$MEdemo, major.zero$FYE,
major.zero$GEN)
# colnames(large.pop) = c("BE_BME", "CME", "CVL", "CE_CSE", "EE_ECE", "ME",
"FYE", "GEN")
# have.demo = complete.cases(large.pop)
# ID.usable = ID.frame[have.demo,]
#
# models = NULL
```

```
#
# for(col in 1:ncol(large.pop)){
#   maj.name = colnames(large.pop)[col]
#   in.major = large.pop[have.demo, col] == 1
#   students = ID.usable[in.major,]
#   new.formula = paste("EngID ~", no.form)
#   models[[maj.name]] = summary(lm(new.formula, students))
# }
#
# sigs = NULL
# BF.correct = 1
#
# for(mod in 1:length(models)){
#   sigs = rbind(sigs, t(how.sig(coef(models[[mod]])[,4]*BF.correct)))
# }
# sigs = data.frame(t(sigs))
# colnames(sigs) = colnames(large.pop)
# ```
#
# ```{r binary interest demographics}
# int.demo = interests
# for(name in names(interests)){
#   qt3 = summary(interests[,name])[[5]]
```

```
#   int.demo[,name][interests[,name] >= qt3] = 1
#   int.demo[,name][interests[,name] < qt3] = 0
# }
# ```
#
# Here we do a base regression of Identity (engineering in this case) as predicted by our other attitudinal factors (not including the other primary identity variable, but yes including Math identity). We run the first regression with all of the attitudes, and then stepwise remove those which are highly non-significant in order until the entire set of variables is significant.
#
# Geoff suggests 0.01 significance as a final cutoff, and ad-hoc Bonferroni correction. Even though we run the linear model a half-dozen times, we should not count all of those as independent statistical tests (for correction calculations) because they are highly related to each other and are all doing the same test time and again on the same data.
#
# ```{r base ELM}
# # ID.frame is the EID/PID/MID + other interests
# att.names = c(colnames(no.frame), "MathID")
# att.form = paste(att.names, collapse = " + ")
# summary(lm(paste('EngID ~ ', att.form), ID.frame))
#
# start.ELM = (lm(paste('EngID ~ ', att.form), ID.frame))
```

```
#
# ## Neuroticism and Conscientiousness are insignificant at > 0.8, remove and retry
# remove = c("Neuroticism", "Conscientiousness")
# att.names = att.names[!grepl(paste(remove, collapse = "|"), att.names)]
# att.form = paste(att.names, collapse = " + ")
# summary(lm(paste('EngID ~ ', att.form), ID.frame))
#
# ## Grit COI is very non-significant, remove that
# remove = c("Grit.CoI")
# att.names = att.names[!grepl(paste(remove, collapse = "|"), att.names)]
# att.form = paste(att.names, collapse = " + ")
# summary(lm(paste('EngID ~ ', att.form), ID.frame))
#
# ## Instrumentality
# remove = c("Instrumentality")
# att.names = att.names[!grepl(paste(remove, collapse = "|"), att.names)]
# att.form = paste(att.names, collapse = " + ")
# summary(lm(paste('EngID ~ ', att.form), ID.frame))
#
# ## Performance Approach
# remove = c("Perform.App")
# att.names = att.names[!grepl(paste(remove, collapse = "|"), att.names)]
# att.form = paste(att.names, collapse = " + ")
```

```
# summary(lm(paste('EngID ~ ', att.form), ID.frame))
#
# ## Agreeableness
# remove = c("Agreeableness")
# att.names = att.names[!grepl(paste(remove, collapse = "|"), att.names)]
# att.form = paste(att.names, collapse = " + ")
# summary(lm(paste('EngID ~ ', att.form), ID.frame))
#
# ## no longer any massively insignificant things; Connectedness and Value are largest >
0.34
# remove = c("Connectedness", "Value")
# att.names = att.names[!grepl(paste(remove, collapse = "|"), att.names)]
# att.form = paste(att.names, collapse = " + ")
# summary(lm(paste('EngID ~ ', att.form), ID.frame))
#
# ## Mastery
# remove = c("Mastery.App")
# att.names = att.names[!grepl(paste(remove, collapse = "|"), att.names)]
# att.form = paste(att.names, collapse = " + ")
# summary(lm.beta(lm(paste('EngID ~ ', att.form), ID.frame)))
#
# ## Everything now significant, though Sci.AB and Extroversion are only at **
significance ( 0.01 > p > 0.001) while everything else is e-5 or much lower
```

this is a change from when we accidentally left off MathID; mediating effect on the effect of Extro and SciAB?

#

final.ELM = (lm(paste('EngID ~ ', att.form), ID.frame))

final.ELM.form = att.form

vif(final.ELM)

```

Now we repeat that process of iterative linear modeling, except with Physics Identity as the predicted variable.

```{r automated iterative PLM building}

IDAT.frame = cbind(ID.frame, interests)

new.PLM = NULL

att.form =  "Belongingness + Perform.App + Mastery.App + Work.Avoid + Expectancy + Connectedness + Instrumentality + Value + Perf.of.Future + Grit.PoE + Grit.CoI + Eng.AB + Sci.AB + Neuroticism + Extroversion + Agreeableness + Conscientiousness + Openness + MathID"

max.pvalue = 0.05 # significance to try to reduce error

ending = FALSE

```
rsqlist = NULL
aiclist = NULL
while(!ending){
  physPLM = (lm(paste("PhysID ~ ", att.form), IDAT.frame))
  clist = coef(summary(physPLM))
  plist = clist[2:nrow(clist),4]
  padjlist = p.adjust(plist, "holm", n=20)
  if(max(padjlist) > max.pvalue){
    rnum = which(plist == max(plist))
    remove = rownames(clist)[2:18][which(plist == max(plist))]
    print(paste(remove, plist[rnum]))
    rsqlist = c(rsqlist, summary(physPLM)$adj.r.squared)
    aiclist = c(aiclist, AIC(physPLM))
    formlist = strsplit(att.form, " + ", fixed = TRUE)[[1]]
    newformlist = formlist[!grepl(remove, formlist)]
    att.form = paste(newformlist, collapse = " + ")
  }
  else{
    ending = TRUE
  }
}
# plot(rsqlist, type = 'b')
plot(aiclist, type = "b")
```

```

There ae 8 attitudes which are significant after Holm-correcting @n=20, and two more which are significant if we pretend we only have the 12 that remain after culling. Instrumentality, Neuroticism and Agreeableness are @12 significant, but not @20, and Conscientiousness survives culling but is not Holm-significant in either case.

The @20 significance levels are identical if we just throw all of the attitudes in from the beginning and then correct for multiple-testing; the same ones are removed as eventually fell out of the culling + correcting, without the weirdness that happens from the culling (i.e. Agreeableness and Neuroticism not being signficant until after other factors are culled).

Normality tests to see if we can really trust our Beta coefficients.

```{r quartile normality test}
att.names = c(colnames(no.frame), "MathID")
for(i in 1:length(att.names)){
  qqnorm(ID.frame[,att.names[i]], main = att.names[i])
  qqline(ID.frame[,att.names[i]])
}
```
```

Our variables aren't super normal. I think looking at Estimates rather than Betas gives us a better idea of how the variable behaves, especially when all of these factors work on the same scale.

```{r build interest and attitude frame IDAT.frame}
#
# #### Engineering ####
# int.ELM = NULL
# for(interest in colnames(interests)){
#   att.form = paste(sapply(strsplit(final.ELM.form, " + ", fixed = TRUE), paste, "*", interest, sep = ""), collapse = " + ")
#   int.ELM[[interest]] = summary(lm(paste('EngID ~ ', att.form), IDAT.frame))
# }
#
# #### Physics ####
# reduced.PLM = summary(lm.beta(lm(PhysID ~ Belongingness + Expectancy + Connectedness + Perf.of.Future  + Eng.AB + Sci.AB + Openness + MathID, data = IDAT.frame)))
#
# physplm.form = "Belongingness + Expectancy + Connectedness + Perf.of.Future + Eng.AB + Sci.AB + Openness + MathID"
# int.PLM = NULL
# for(interest in colnames(interests)){
```

```
#   att.form = paste(sapply(strsplit(physplm.form, " + ", fixed = TRUE), paste, "*",
interest, sep = ""), collapse = " + ")
#   int.PLM[[interest]] = (lm(paste('PhysID ~ ', att.form), IDAT.frame))
# }
```

```

Turns out I *can* automate this if I let it look terrible. The below code iteratively removes the interaction term with the highest p-value and then re-runs the regression. It stops when all interactions are gone, or when everything is significant. Doesn't remove "the interest itself is non-significant" even if it's the last one, so we need to check that by hand.

```{r automating it all}
base_num = 9 + 1 + 1 # eight attitudes, 1 interest, 1 intercept
max.pvalue = 0.05 # significance to try to reduce error
physplm.form = "Belongingness + Expectancy + Connectedness + Instrumentality +
Perf.of.Future + Eng.AB + Sci.AB + Openness + MathID"

## old analysis code for when we included all the different interaction terms, rather than just Interest

# for(interest in colnames(interests)){


```
```

```
#   #print("")
#   #print(interest)
#   att.form = paste(physplm.form, interest, paste(sapply(strsplit(physplm.form, " + ",
fixed = TRUE), paste, ":", interest, sep = ""), collapse = " + "), sep = " + ")
#   ending = FALSE
#   rsqlist = NULL
#   aiclist = NULL
#   while(!ending){
#     new.PLM[[interest]] = (lm(paste("PhysID ~ ", att.form), IDAT.frame))
#     clist = coef(summary(new.PLM[[interest]]))
#     plist = clist[(base_num+1):nrow(clist),4]
#     padj = p.adjust(plist, n = 18)  # was 162
#     if(max(padj) > max.pvalue){
#       rnum = which(plist == max(plist))
#       remove = rownames(clist)[base_num + which(plist == max(plist))]
#       #print(paste(remove, plist[rnum]))
#       rsqlist = c(rsqlist, summary(new.PLM[[interest]])$adj.r.squared)
#       aiclist = c(aiclist, AIC(new.PLM[[interest]]))
#       formlist = strsplit(att.form, " + ", fixed = TRUE)[[1]]
#       newformlist = formlist[!grepl(remove, formlist)]
#       att.form = paste(newformlist, collapse = " + ")
#       if(length(newformlist) == 10){
#         new.PLM[[interest]] = (lm(paste("PhysID ~ ", att.form), IDAT.frame))
```

```
#         ending = TRUE # end the loop before breaking
#       }
#     }
#     else{
#       ending = TRUE
#     }
#   }
#   # plot(rsqlist, type = 'b')
#   #plot(aiclist, type = "b")
# }

for(interest in colnames(interests)){
  att.form = paste("PhysID ~ ", physplm.form, " + ", interest)
  new.PLM[[interest]] = lm(att.form, data = IDAT.frame)
}

vif(physPLM) # all less than 2, maximum 1.90 (Belongingness)
new.PLM$vif = lapply(new.PLM, vif) # all less than 2, maximum 1.925 (Belongingness)

for(test in colnames(interests)){
  n.test = 11 # was 8*18
  new.PLM$adj.p[[test]] = p.adjust(coef(summary(new.PLM[[test]]))[,4], n = n.test)
  #print(how.sig(t(t(new.PLM$adj.p[[test]]))))
```

```
}
# for(i in 1:length(new.PLM)){print(paste(names(new.PLM[i]), " ",
summary(new.PLM[[i]])$adj.r.squared))}
## if we want the adj.rsquared values
```

```{r}
newPs = NULL
for(interest in names(new.PLM$adj.p)){
  newPs = cbind(newPs, new.PLM$adj.p[[interest]][1:11])
}
colnames(newPs) = names(new.PLM$adj.p)
rownames(newPs)[11] = "Major"
newSig = NULL
for(col in 1:ncol(newPs)){
  newSig = cbind(newSig, as.matrix(how.sig(t(t(newPs[,col])))))
}
newSig = data.frame(newSig)
colnames(newSig) = colnames(newPs)

## base ps
```

```
baseadjp = p.adjust(coef(summary(physPLM))[,4], n = 20)

baseadjp[11] = 1

baseSig = how.sig(baseadjp)

rownames(baseSig)[11] = "Major"

usedSig = baseSig[intersect(rownames(newSig), rownames(baseSig)),]

combSig = cbind(usedSig, newSig)

rownames(combSig) = rownames(newPs)

t(combSig)
```

Discriminant factor analysis for whether we can safely use interest rather than demographic major.

```{r Discriminant factor analysis}
library(MASS)

interest.names = c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE_CSE",
"EE_ECE", "IND", "MSE", "ME", "NUKE")
interest.string = paste(interest.names, collapse = " + ")

# ## this stuff will fail
# for(major in colnames(interests)){
#   grouping = factor(major.zero[,major])
```

```
#   new.df = data.frame(interests, grouping)
#
# }
# qdanal = qda(grouping ~ AAE + ABE_BSE + BE_BME + CME + CVL + CE_CSE +
EE_ECE + IND + MSE + ME + NUKE, new.df)
```

Turns out DFA is hard to match all the assumptions for (multivariate normality, homogeneity of variance/covariance, multicollinearity, and independence). Independence should be good, but the normality is probably all kinds of violated, for example.

```{r "Let's try it again with a GLM to do logistic regression"}
major.zero$ABE_BSE = ifelse(major.zero$ABE == 1 | major.zero$BSE == 1, 1, 0)
major.zero$BE_BME = ifelse(major.zero$BE == 1 | major.zero$BME == 1, 1, 0)
major.zero$CE_CSE = ifelse(major.zero$CE == 1 | major.zero$CSE == 1, 1, 0)
#major.zero$EE_ECE = ifelse(major.zero$EE == 1 | major.zero$ECE == 1, 1, 0)

major.new = major.zero[,c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL",
"CE_CSE",  "EE", "ENV", "MSE", "ME", "NUKE")]
```

```
colnames(major.new) = c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL",
"CE_CSE",  "EE", "EEE", "MSE", "ME", "NUKE") # change ENV to EEE to match
below
# we're leaving out a lot of the majors here. In particular, we leave out EXP, FYE, GEN,
IND, OTH, and Undeclared

### need to remove CON and EP from interests for the next loop because only a single
person declared it as their major
new_int = c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE_CSE", "EE",
"EEE", "MSE", "ME", "NUKE") # to match above
sub.int = interests[,colnames(interests) %in% new_int]


## this seems to work better
major.glm = NULL
glm.est = NULL
glm.p = NULL
for(major in colnames(major.new)){
  grouping = factor(major.new[,major])
  new.df = data.frame(sub.int, grouping)
  major.glm[[major]] = glm(grouping~., family = binomial(link="logit"), data = new.df)
  glm.est = cbind(glm.est, coef(summary(major.glm[[major]]))[,1])
  glm.p = cbind(glm.p, how.sig(coef(summary(major.glm[[major]]))[,4]))
```

```
}
colnames(glm.est) = new_int
colnames(glm.p) = new_int

main.estimates = get.diag(glm.est[2:12,])
#lapply(major.glm,summary)

levelplot(glm.est[2:12,], at = seq(-ceiling(max(glm.est)), ceiling(max(glm.est)), by =
.222), xlab="Interest", ylab="Declared Major", scale=list(x=list(rot=45)))
#levelplot(exp(glm.est[2:12,]), at = seq(2-ceiling(exp(max(glm.est))),
ceiling(exp(max(glm.est))), by = 3))
```

Or, in an easier way with biserial correlations

```{r}
newname = sapply(colnames(major.new), paste, "_declared", sep = "")
colnames(major.new) = newname
BS = polyserial(sub.int, major.new)
corrplot(BS)

summary(BS[c(2:4, 8), 10]) # correlation of "life science" engineerings with ME
```

If we want to check our explanation of Belongingness, let's regress Belongingness with the subconstructs of Physics Identity

```{r}
belonging.subreg.data = factor.space[,c("Belongingness", "Phys ID: Perf\\Comp", "Phys ID: Recognition", "Phys ID: Interest")]
colnames(belonging.subreg.data) = c("Belongingness", "PhysID.PC", "PhysID.Rec", "PhysID.Int")
summary(lm.beta(lm(Belongingness ~ PhysID.PC + PhysID.Rec + PhysID.Int, data = belonging.subreg.data)))
```

Bringing back the ikde lists from earlier to try to plot distributions of answers to Q14

```{r}
names(ikde.list) = colnames(interest)
plot(cbind(ikde$x, ikde.list[[17]]), type = "l", col = 1, lwd = 2, xlim = c(0, 6), ylim = c(0, .3))
for(i in 1:18){
  points(cbind(ikde$x, ikde.list[[i]]), type = "l", col = i, lwd = 2, lty = i %% 5 + 1)
}
ikde.matrix = matrix(ncol = 18, nrow = 512) # ikde.list has 512 data points
for(i in 1:length(ikde.list)){
  ikde.matrix[,i] = ikde.list[[i]]
```

```
}

matplot(ikde$x, ikde.matrix, type = c("l"),col = rainbow(18), lwd = 3, xlab = "Interest in pursuing major", ylab = "Density of Responses") #plot
legend("topleft", legend = colnames(interests), col=rainbow(18), lwd = 3) # optional legend

```

Changes in estimates from base to different

```{r}
base.estimate = coef(summary(lm.beta(physPLM)))[1:10,1]
new.estimates = matrix(ncol = 18, nrow = 10)
for(i in 1:ncol(new.estimates)){
  new.estimates[1:10,i] = coef(summary(lm.beta(new.PLM[[i]])))[1:10,1]
}
colnames(new.estimates) = colnames(interests)
rownames(new.estimates) = rownames(coef(summary(physPLM)))

base.beta = coef(summary(lm.beta(physPLM)))[1:10,2]
new.beta = matrix(ncol = 18, nrow = 10)
for(i in 1:ncol(new.estimates)){
  new.beta[1:10,i] = coef(summary(lm.beta(new.PLM[[i]])))[1:10,2]
```

```
}


diff.est = new.estimates-base.estimate

diff.est = diff.est[2:10,]


diff.beta = new.beta-base.beta

diff.beta = diff.beta[2:10,]


se = coef(summary(physPLM))[2:10,2]

new.se = se

#for(i in 1:ncol(new.estimates)){

#  print((new.se - coef(summary(new.PLM[[i]]))[2:10,2])/se)

#}

## all the new standard errors are slightly less but approximately equal to the original

standard error

## so shared variance doubles and standard deviation increases by a factor of sqrt(2)

frac.est = diff.est

frac.beta = diff.beta

for(i in 1:ncol(diff.est)){

  frac.est[,i] = diff.est[,i] / (sqrt(2)*se)

  frac.beta[,i] = diff.beta[,i] / (sqrt(2)*se)

}
```

```
rownames(diff.est) = c("Belongingness", "Expectancy", "Connectedness",
"Instrumentality", "Perceptions of Future", "Engineering Agency Beliefs", "Science
Agency Beliefs", "Openness", "Math Identity")

#levelplot(t(diff.est), at = seq(-.15, .15, by = 0.0333), xlab="Test", ylab="Estimate",
scale=list(x=list(rot=45)))
#levelplot(t(frac.est), at = seq(-3.6, 3.6, by = 0.333), xlab="Test", ylab="Estimate",
scale=list(x=list(rot=45)))
levelplot(t(frac.est)[1:17,], at = c(-10, -1.96, 1.96, 10), xlab="Test", ylab="Estimate",
scale=list(x=list(rot=45)))
levelplot(t(diff.est)[1:17,], xlab = "Added Major", ylab = "Difference in Estimate from
Base Model", scale=list(x=list(rot=45)))

## this plot shows us, interestingly enough, that the only estimates which showed a
statistically significant difference
## were the ones that had a significant intereaction with the Interest question
```

```{r}
EP.subreg = cbind(interests$EP, factor.space[,c("Phys ID: Perf\\Comp", "Phys ID:
Recognition", "Phys ID: Interest")])
colnames(EP.subreg) = c("EP", "PhysID.PC", "PhysID.Rec", "PhysID.Int")
summary(lm.beta(lm(EP ~ PhysID.PC + PhysID.Rec + PhysID.Int, data = EP.subreg)))
```

```

```{r}
summary(lm.beta(lm(PhysID ~ Belongingness + Expectancy + Instrumentality +
Perf.of.Future + Sci.AB + Eng.AB + MathID + Openness + Connectedness + AAE +
ABE_BSE + BE_BME + CVL + CON + CME + IT + IND + CE_CSE + EE + EP + EEE
+ ME + MIE + MSE + NUKE, data = IDAT.frame)))

## interesting. AAE > ME
att.form = "Belongingness + Expectancy + Instrumentality + Perf.of.Future + Sci.AB +
Eng.AB + MathID + Openness + Connectedness + AAE + ABE_BSE + BE_BME +
CVL + CON + CME + IT + IND + CE_CSE + EE + EP + EEE + ME + MIE + MSE +
NUKE"
ending = FALSE
rsqlist = NULL
aiclist = NULL
while(!ending){
  bigLM = (lm(paste("PhysID ~ ", att.form), IDAT.frame))
  clist = coef(summary(bigLM))
  plist = clist[11:nrow(clist),4]
  padjlist = p.adjust(plist, "holm", n=25)
  if(max(plist) > max.pvalue){
```
```

```
    rnum = which(plist == max(plist))
    remove = rownames(clist)[11:25][which(plist == max(plist))]
    print(paste(remove, plist[rnum]))
    rsqlist = c(rsqlist, summary(bigLM)$adj.r.squared)
    aiclist = c(aiclist, AIC(bigLM))
    formlist = strsplit(att.form, " + ", fixed = TRUE)[[1]]
    newformlist = formlist[!grepl(remove, formlist)]
    att.form = paste(newformlist, collapse = " + ")
  }
  else{
    ending = TRUE
  }
}

## what if we just do the interests?
summary(lm.beta(lm(PhysID ~ AAE + ABE_BSE + BE_BME + CVL + CON + CME +
IT + IND + CE_CSE + EE + EP + EEE + ME + MIE + MSE + NUKE, data =
IDAT.frame)))
att.form =  "AAE + ABE_BSE + BE_BME + CVL + CON + CME + IT + IND +
CE_CSE + EE + EP + EEE + ME + MIE + MSE + NUKE"
ending = FALSE
rsqlist = NULL
aiclist = NULL
```

```
while(!ending){
  intLM = (lm(paste("PhysID ~ ", att.form), IDAT.frame))
  clist = coef(summary(intLM))
  plist = clist[2:nrow(clist),4]
  padjlist = p.adjust(plist, "holm", n=18)
  if(max(plist) > max.pvalue){
    rnum = which(plist == max(plist))
    remove = rownames(clist)[2:25][which(plist == max(plist))]
    print(paste(remove, plist[rnum]))
    rsqlist = c(rsqlist, summary(intLM)$adj.r.squared)
    aiclist = c(aiclist, AIC(intLM))
    formlist = strsplit(att.form, " + ", fixed = TRUE)[[1]]
    newformlist = formlist[!grepl(remove, formlist)]
    att.form = paste(newformlist, collapse = " + ")
  }
  else{
    ending = TRUE
  }
}
```

```{r betas for the different majors}
int.beta = NULL
```

```
for(i in 1:ncol(new.estimates)){
  int.beta[i] = coef(summary(lm.beta(new.PLM[[i]])))[11,2]
}
int.beta = data.frame(int.beta, row.names = colnames(interests))
```

below this line needs to be corrected/updated

##

EEE, IT, and ONON have an insignificant base-interest, and no interaction terms. MSE and OSTEM has an insignificant base-interest but significant interaction terms with that interest.

```{r below this line needs to be corrected/updated ##### ############### #### and now we make a big significance matrix}
sigs = NULL
bonferroni.correction = 380
sigs = rbind(sigs, c(t(how.sig(coef(summary(physPLM))[2:11,4]*bonferroni.correction)), ""))
## base levels of significance per model
for(mod in 1:length(new.PLM)){
  sigs = rbind(sigs,
t(how.sig(coef(summary(new.PLM[[mod]]))[2:12,4]*bonferroni.correction)))
```

```
}
sigs = data.frame(t(sigs))
colnames(sigs) = c("Overall", names(new.PLM))
rownames(sigs) = c(rownames(sigs)[1:10], "Interest")
```

Can we automatically add the interaction terms or do we need to do it by hand?

We need new rows for Interest x (Connectedness, Perf.of.Future, Sci.AB, Belongingness, Agreeableness, Instrumentality, Eng.AB, Expectancy, Openness). That's 9 more rows added to our matrix which is 19 columns wide. Everything except MathID has at least one interaction term among the interests.

```{r adding interaction to significance matrix}
blank.raw = matrix(data = "", nrow = 10, ncol = 19)
rownames(blank.raw) = c("Belongingness:Interest", "Expectancy:Interest",
"Connectedness:Interest", "Instrumentality:Interest", "Perf.of.Future:Interest",
"Eng.AB:Interest", "Sci.AB:Interest", "Agreeableness:Interest", "Openness:Interest",
"Math.ID:Interest")
colnames(blank.raw) = colnames(sigs)
# relevel sigs, and then bind
sig.frame = as.matrix(rbind(sigs, blank.raw))
## add sigs by hand...blah
```

```
#sig.frame[c("Connectedness:Interest", "Perf.of.Future:Interest"),c("AAE")] = "*"

sig.frame[c("Sci.AB:Interest"),c("BE_BME")] = "***"

# sig.frame[c("Sci.AB:Interest", "Perf.of.Future:Interest"),c("CME")] = "*"
sig.frame[c("Perf.of.Future:Interest"),c("CME")] = "**"

#sig.frame[c("Perf.of.Future:Interest"),c("CE_CSE")] = "*"

#sig.frame[c("Connectedness:Interest", "Instrumentality:Interest", "Eng.AB:Interest",
"Agreeableness:Interest"),c("EE_ECE")] = "*"

#sig.frame[c("Sci.AB:Interest", "Instrumentality:Interest"), c("EP")] = "*"
sig.frame[c("Agreeableness:Interest"), c("EP")] = "**"
sig.frame[c("Perf.of.Future:Interest"),c("EP")] = "***"

#sig.frame[c("Expectancy:Interest", "Agreeableness:Interest"),c("IT")] = "*"

#sig.frame[c("Eng.AB:Interest"),c("MSE")] = "*"
#sig.frame[c("Agreeableness:Interest"),c("MSE")] = "**"

#sig.frame[c("Perf.of.Future:Interest", "Expectancy:Interest"),c("ME")] = "*"
#sig.frame[c("Belongingness:Interest"),c("ME")] = "*"
```

```
sig.frame[c("Connectedness:Interest"),c("OSTEM")] = "**"
sig.frame[c("Openness:Interest"),c("OSTEM")] = "***"

sig.frame = data.frame(sig.frame)
```

How many / what fraction of students declared into something like "Gen Eng" or otherwise didn't declare for a major?

```{r}
general.categories = c("EXP", "FYE", "GEN", "Undeclared")
major.totals = colSums(major.zero, na.rm = T)
num_gen = sum(major.totals[general.categories])
total = nrow(INICE$data)
fraction = num_gen/total
```

Which schools use which name for their gen eng programs?

```{r}
school.major = cbind(major.zero, INICE$data$school)
for(cat in general.categories){
```

```
  print(cat)
  print(summary(lm(paste(cat, "~ INICE$data$school"), school.major)))
}
```

The neutral category is Clemson. A nearly-zero intercept says Clemson doesn't use that, while a larger positive one means they do. Significant non-zero positive estimates mean Clemson doesn't and that school does, while significant non-zero negative estimates mean Clemson does and they don't. Pretty sure I said that all correctly.

Exploratory: FIU and Purdue both use this.
First-Year-Engineering: Purdue is the only one who uses this.
General Engineering: Clemson is the only one who uses this (tiny bit Purdue?)
Undeclared: primarily UNR, little bit of Purdue.

Maybe we can see this information another way.

```{r}
weird.cat = NULL
for(cat in general.categories){
  weird.cat = cbind(weird.cat,paste(school.major[,cat],
school.major$'INICE$data$school', sep = ""))
}
```

```
colnames(weird.cat) = general.categories
for(cat in general.categories){
  print(cat)
  print(table(weird.cat[,cat]))
}
weird.cat = data.frame(weird.cat)
```

Here we can see, for example, the combination "1" + "school" only appears in certain schools.

Exp: FIU and Purdue

FYE: Purdue

GEN: Clemson, Purdue (1 FIU, 10 UNR)

UnDec: 2 clemson, 26 Purdue, 55 UNR

Better representation

```{r}
for(cat in general.categories){
  cat.mat = t(table(as.matrix(weird.cat[grepl("1", weird.cat[,cat]),cat])))
  rownames(cat.mat) = cat
  print(cat.mat)
```

}

```

I realize now I could have done this whole thing simpler with something like

```{r}
table(INICE$data$school[major.matrix$FYE == 1]) # for FYE
table(INICE$data$school[major.matrix$GEN == 1])
table(INICE$data$school[major.matrix$EXP == 1])
table(INICE$data$school[major.matrix$Undeclared== 1])
```

Check whether coefficients match

```{r}
newframe = NULL
for(i in 1:length(new.PLM)){
  newframe = rbind(newframe,coef(summary(new.PLM[[i]]))[2:12,1])
}
oldreg = c(coef(summary(physPLM))[2:11,1], 0)
newframe = rbind(oldreg, newframe)
scaleframe = scale(newframe, center = oldreg, scale = FALSE)
rownames(scaleframe) = c("BASE", names(new.PLM))
```
```

```
rownames(newframe) = rownames(scaleframe)
colnames(scaleframe)[11] = "Interest"
colnames(newframe) = colnames(scaleframe)
for(i in 1:ncol(scaleframe)){
  barplot(scaleframe[,i], names.arg = rownames(scaleframe), cex.names = 1, main =
colnames(scaleframe)[i], las = 2)
}
for(i in 1:ncol(newframe)){
  barplot(newframe[,i], names.arg = rownames(newframe), cex.names = 1, main =
colnames(newframe)[i], las = 2)
}
```

Core Methods.Rmd

```
---
title: "Core Methods"
author: "Jackie Doyle"
date: "May 3, 2017"
output:
  pdf_document: default
  html_document: default
---


```{r setup, include=FALSE}
#knitr::opts_chunk$set(echo = TRUE)
```


```{r data setup and loading}
#setwd('C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Group Differences')
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Mapping/env_v2_group_comparisons_done.RData")
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Raw data/INICE_v2.RData")
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Mapping/groupdata.RData")
```
```

```
source("C:/Users/Jackie/Dropbox/R Files/custom.R")
library(pwr)
```

Now we want proportions of various groups in each of these. The usual groups of interest for our population include Women/female-identified, Asian, Hispanic.

Interest scales for ME, BME

```{r make reusable funciton to check membership}
is.thing = function(data, question, val = 1){
  yes = data[,question] == val & !is.na(data[,question])
  return(sum(yes))
}
# in hindsight, probably better served by sum(any.filled(data[,question]))
```

```{r}
female = "Q17a"
asian = "Q16d"
hispanic = "Q16b"
nng.names = c("NnG_1", "NnG_2", "NnG_3", "NnG_4", "NnG_5", "NnG_6", "NnG_7")
```

```
group.values = matrix(dimnames = list(c("NG", nng.names, "DG", "Overall"),
c("Female", "Asian", "Hispanic", "ME", "BE_BME", "Total")), nrow = 10, ncol = 6)


group.values["NG",] = c(is.thing(group.data$raw$NG, female),
                        is.thing(group.data$raw$NG, asian),
                        is.thing(group.data$raw$NG, hispanic),
                        mean(group.data$raw$NG$Q14n, na.rm = T), # ME
                        mean(group.data$raw$NG$Q14c, na.rm = T), # BE_BME
                        nrow(group.data$raw$NG))
for(i in 1:7){
  group.values[(i+1),] = c(is.thing(group.data$raw$NnG[[i]], female),
                        is.thing(group.data$raw$NnG[[i]], asian),
                        is.thing(group.data$raw$NnG[[i]], hispanic),
                        mean(group.data$raw$NnG[[i]]$Q14n, na.rm = T), # ME
                        mean(group.data$raw$NnG[[i]]$Q14c, na.rm = T), # BE_BME
                        nrow(group.data$raw$NnG[[i]]))
}


group.values["DG",] = c(is.thing(group.data$raw$DG, female),
                        is.thing(group.data$raw$DG, asian),
                        is.thing(group.data$raw$DG, hispanic),
                        mean(group.data$raw$DG$Q14n, na.rm = T), # ME
```

```
                mean(group.data$raw$DG$Q14c, na.rm = T), # BE_BME
                nrow(group.data$raw$DG))

# overall, only 1694 of 2916 students have been counted so far, about 58 %

group.values["Overall",] = c(is.thing(INICE$data, female),
                is.thing(INICE$data, asian),
                is.thing(INICE$data, hispanic),
                mean(INICE$data$Q14n, na.rm = T), # ME
                mean(INICE$data$Q14c, na.rm = T), # BE_BME
                nrow(INICE$data))

group.values = data.frame(group.values)

group.prop = group.values
group.prop[,1:3] = group.values[,1:3]/group.prop$Total
print(group.prop, digits = 3)
```

```{r proportion testing and mean difference testing}
prop.names = c("Female", "Asian", "Hispanic")
mean.names = c("ME", "BE_BME")
```

```
test = NULL
test$matrix =  matrix(dimnames = list(c(nng.names, "DG", "Overall"), c(prop.names,
mean.names, "Total")), nrow = 9, ncol = 6)
test$p = test$matrix
test$conf.l = test$matrix
test$conf.u = test$matrix

for(row in rownames(test$matrix)){
  for(col in prop.names){
    successes = c(group.values[row, col], group.values["NG", col])
    trials = c(group.values[row, "Total"], group.values["NG", "Total"])
    pt = prop.test(successes, trials)
    test$matrix[row, col] = how.sig(pt$p.value)[[1]]
    test$p[row, col] = pt$p.value
    test$conf.l[row,col] = pt$conf.int[1]
    test$conf.u[row,col] = pt$conf.int[2]
  }
  for(col in mean.names){
    q = switch(col, "ME" = "Q14n", "BE_BME" = "Q14c")
    i = which(row == c(nng.names, "DG", "Overall"))
    if(i < 8){
      x = group.data$raw$NnG[[i]][,q]
    } else if(row == "DG"){
```

```
    x = group.data$raw[[row]][,q]
  } else if(row == "Overall"){
    x = INICE$data[,q]
  }

  y = group.data$raw$NG[,q]
  twpt = twoway.perm.test(x, y, type="wilcox")
  test$matrix[row, col] = how.sig(pvalue(twpt)[1])[[1]]
  test$p[row, col] = pvalue(twpt)[1]
  test$conf.l[row,col] = attributes(pvalue(twpt))$conf.int[1]
  test$conf.u[row,col] = attributes(pvalue(twpt))$conf.int[1]
 }
}
test.NG = test

## and vs overall
test = NULL
test$matrix =  matrix(dimnames = list(c("NG", nng.names, "DG"), c("Female", "Asian",
"Hispanic", "ME", "BE_BME", "Total")), nrow = 9, ncol = 6)
test$p = test$matrix
test$conf.l = test$matrix
test$conf.u = test$matrix
```

```
for(row in rownames(test$matrix)){
  for(col in prop.names){
    successes = c(group.values[row, col], group.values["Overall", col])
    trials = c(group.values[row, "Total"], group.values["Overall", "Total"])
    pt = prop.test(successes, trials)
    test$matrix[row, col] = how.sig(pt$p.value)[[1]]
    test$p[row, col] = pt$p.value
    test$conf.l[row,col] = pt$conf.int[1]
    test$conf.u[row,col] = pt$conf.int[2]
  }
  for(col in mean.names){
    q = switch(col, "ME" = "Q14n", "BE_BME" = "Q14c")
    i = which(row == c(nng.names, "DG", "NG"))
    if(i < 8){
      x = group.data$raw$NnG[[i]][,q]
    } else{
      x = group.data$raw[[row]][,q]
    }
    y = group.data$raw$NG[,q]
    twpt = twoway.perm.test(x, y, type="wilcox")
    test$matrix[row, col] = how.sig(pvalue(twpt)[1])[[1]]
    test$p[row, col] = pvalue(twpt)[1]
    test$conf.l[row,col] = attributes(pvalue(twpt))$conf.int[1]
```

```
    test$conf.u[row,col] = attributes(pvalue(twpt))$conf.int[1]
  }

}
test.overall = test

print(test.NG)
print(test.overall)
```

Do we have the power to see things that we want?

```{r this has become our wilcox.power function}
# n = 1000
# pval <- replicate(n, wilcox.test(rnorm(547,4.309,.012),
rnorm(41,3.675,1.7597))$p.value)
# summary(pval)
# sum(pval < .05) / n
```

```{r power calculations for wilcox test by monte carlo simulation, cache = TRUE}
```

```
first.power = matrix(nrow = 9, ncol = 2, dimnames = list(c(nng.names, "DG", "Overall"),
mean.names))

for(row in c(nng.names, "DG", "Overall")){
  for(col in mean.names){
    q = switch(col, "ME" = "Q14n", "BE_BME" = "Q14c")
    i = which(row == c(nng.names, "DG", "Overall"))
    if(i < 8){
      x = group.data$raw$NnG[[i]][,q]
    } else if(row == "DG"){
      x = group.data$raw[[row]][,q]
    } else if(row == "Overall"){
      x = INICE$data[,q]
    }
    y = group.data$raw$NG[,q]
    wp = wilcox.power(x, y, n = 1000)
    first.power[row, col] = wp*100
  }
}

second.power  = matrix(nrow = 9, ncol = 2, dimnames = list(c("NG", nng.names, "DG"),
mean.names))
```

```
for(row in c("NG", nng.names, "DG")){
  for(col in mean.names){
    q = switch(col, "ME" = "Q14n", "BE_BME" = "Q14c")
    i = which(row == c(nng.names, "DG", "NG"))
    if(i < 8){
      x = group.data$raw$NnG[[i]][,q]
    } else{
      x = group.data$raw[[row]][,q]
    }
    y = group.data$raw$NG[,q]
    wp = wilcox.power(x, y, n = 1000)
    second.power[row, col] = wp*100
  }
}

print(first.power, digits = 2)
print(second.power, digits = 2)
```

Proportion power: we don't have enough people to get power of 0.8

```{r proportion power}
```

```
# power.prop.test(p1= group.prop["NG", "Female"], p2=group.prop["NnG_1",
"Female"], power = .8)
# power.prop.test(p1= group.prop["NG", "Asian"], p2=group.prop["NnG_1", "Asian"],
power = .8)
# power.prop.test(p1= group.prop["NG", "Hispanic"], p2=group.prop["NnG_1",
"Hispanic"], power = .8)

powers = matrix(nrow = 9, ncol = 3, dimnames = list(c(nng.names, "DG", "Overall"),
prop.names))
for(col in prop.names){
  for(row in rownames(powers)){
    p1 = group.prop["NG", col]
    p2 = group.prop[row, col]
    h = ES.h(p1, p2)
    powers[row, col] = pwr.2p2n.test(h, n1 = group.prop["NG", "Total"], n2 =
group.prop[row, "Total"])$power
  }
}

powers.ng = powers
print(powers.ng*100, digits = 2)
```

```
powers = matrix(nrow = 9, ncol = 3, dimnames = list(c("NG", nng.names, "DG"),
prop.names))
for(col in prop.names){
  for(row in rownames(powers)){
    p1 = group.prop["Overall", col]
    p2 = group.prop[row, col]
    h = ES.h(p1, p2)
    powers[row, col] = pwr.2p2n.test(h, n1 = group.prop["Overall", "Total"], n2 =
group.prop[row, "Total"])$power
  }
}
powers.overall = powers
print(powers.overall*100, digits = 2)
```

There are a few tests which have power > 80% which did not turn up as significant to the prop.test earlier; these are Overall vs NnG_2 for Female, and Overall vs NnG_4 for Hispanic. NG_1 vs Overall hispanic and NG_1 vs NnG_4 hispanic was also non-significant.

So there's something here, but it's not really a lot, and lots of the table would be completely blank if we tried to fill it out, due to all the parts where there's insufficient power to even detect something (~5%, etc)

########

\section Comparing NG to Overall Sample

Since the overall sample vs NG was the only one that consistently had statistical power (since NG is large enough) we're going to make comparisons on different axes about those groups

Demographics to consider: Female, Asian, Hispanic. (Able-bodied?)

Factors to consider: ?

Interests to consider: ME, CME, CVL, EE, BME, CS

```{r compare interests}

NG.interests = group.data$raw$NG[,grepl("Q14", colnames(INICE$data))]

#OV.interests = INICE$data[,grepl("Q14", colnames(INICE$data))]

DG.interests = group.data$raw$DG[,grepl("Q14", colnames(INICE$data))]

colnames(NG.interests) = c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE", "CON", "EE", "EP", "EEE", "IND", "IT", "MSE", "ME", "MIE", "NUKE", "OSTEM", "ONON")

#colnames(OV.interests) = colnames(NG.interests)

colnames(DG.interests) = colnames(NG.interests)
```

```
select.interests =  c("AAE", "ME", "EE", "CVL", "CME", "BE_BME", "CE", "IT")


# subset to the ones we care about
NG.interests = NG.interests[,select.interests]
#OV.interests = OV.interests[,select.interests]
DG.interests = DG.interests[,select.interests]


# comp.matrix = matrix(nrow = ncol(NG.interests), ncol = 5)
# rownames(comp.matrix) = colnames(NG.interests)
# for(col in colnames(NG.interests)){
#   test = t.test(NG.interests[,col], OV.interests[,col], conf.int = TRUE)
#   comp.matrix[col,] = c(test$estimate[1] - test$estimate[2], test$p.value,
how.sig(test$p.value)[1], test$conf.int[1], test$conf.int[2])
# }
# comp.df = data.frame(comp.matrix)
# colnames(comp.df) = c("NG - OV", "P.value", "Sig.", "CI.lower", "CI.upper")
# for(i in c(1, 2, 4, 5)){
#   comp.df[,i] = as.numeric(as.character(comp.df[,i]))
# }
# comp.OV = comp.df


## again but with DG
```

```
comp.matrix = matrix(nrow = ncol(NG.interests), ncol = 6)
rownames(comp.matrix) = colnames(NG.interests)
for(col in colnames(NG.interests)){
  test = t.test(NG.interests[,col], DG.interests[,col], conf.int = TRUE)
  comp.matrix[col,] = c(test$estimate[2] - test$estimate[1], p.adjust(test$p.value, n =
length(select.interests)), how.sig(p.adjust(test$p.value[1], n=length(select.interests))),
test$conf.int[1], test$conf.int[2], cohens.d(NG.interests[,col], DG.interests[,col]))

}
comp.df = data.frame(comp.matrix)
colnames(comp.df) = c("DG - NG", "P.value", "Sig.", "CI.lower", "CI.upper", "Cohens
D")
for(i in c(1, 2, 4, 5, 6)){
  comp.df[,i] = as.numeric(as.character(comp.df[,i]))
}
comp.DG = comp.df
```

```{r doing the same as above, just for a bunch of NnG instead of DG}
nng.matrix = NULL
for(g in c(1:7)){
```

```
  g.data = group.data$raw$NnG[[g]][,grepl("Q14", colnames(INICE$data))]
  colnames(g.data) = c("AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE",
"CON", "EE", "EP", "EEE", "IND", "IT", "MSE", "ME", "MIE", "NUKE", "OSTEM",
"ONON")
  nng.matrix[[g]] = matrix(nrow = ncol(NG.interests), ncol = 5)
  rownames(nng.matrix[[g]]) = colnames(NG.interests)
  colnames(nng.matrix[[g]]) = c("DG - NG", "P.value", "Sig.", "CI.lower", "CI.upper")
  for(col in colnames(NG.interests)){
    test = t.test(NG.interests[,col], g.data[,col], conf.int = TRUE)
    # nng.matrix[[g]][col,] = c(test$estimate[2] - test$estimate[1], p.adjust(test$p.value, n
= length(select.interests)), how.sig(p.adjust(test$p.value[1], n=length(select.interests))),
test$conf.int[1], test$conf.int[2])
    nng.matrix[[g]][col,] = c(test$estimate[2] - test$estimate[1], (test$p.value),
how.sig((test$p.value[1])), test$conf.int[1], test$conf.int[2])
  }
  nng.matrix[[g]] = data.frame(nng.matrix[[g]])
  colnames(nng.matrix[[g]]) = c("DG - NG", "P.value", "Sig.", "CI.lower", "CI.upper")
  for(i in c(1, 2, 4, 5)){
    nng.matrix[[g]][,i] = as.numeric(as.character(nng.matrix[[g]][,i]))
  }
}
comp.NnG = nng.matrix
```

And we can compare proportions on some demographic variables

```{r compare demographics}
demo.subset = grepl("(Q17a)|(Q16d)|(Q16b)|(Q15e)|(Q16c)|(Q16g)|(Q18a)|(Q17c)",
colnames(INICE$data))
NG.demo = group.data$raw$NG[,demo.subset]
OV.demo = INICE$data[,demo.subset]
DG.demo = group.data$raw$DG[,demo.subset]

colnames(NG.demo) = c("Able-Bodied", "Hispanic", "White", "Asian", "Black",
"Female", "Male", "Straight")
colnames(OV.demo) = colnames(NG.demo)
colnames(DG.demo) = colnames(NG.demo)

prop.matrix.OV = matrix(nrow = ncol(NG.demo), ncol = 7, dimnames =
list(colnames(NG.demo), c("NG.est", "OV.est", "Est.diff", "P.value", "sig", "CI.upper",
"CI.lower")))
prop.matrix.DG = matrix(nrow = ncol(NG.demo), ncol = 7, dimnames =
list(colnames(NG.demo), c("NG.est", "DG.est", "Est.diff", "P.value", "sig", "CI.upper",
"CI.lower")))
NG.trials = nrow(NG.demo)
OV.trials = nrow(OV.demo)
```

```
DG.trials = nrow(DG.demo)


for(col in colnames(NG.demo)){
  NG.succ = is.thing(NG.demo, col)
  OV.succ = is.thing(OV.demo, col)
  DG.succ = is.thing(DG.demo, col)

  OV.test = prop.test(c(NG.succ, OV.succ), c(NG.trials, OV.trials))
  DG.test = prop.test(c(NG.succ, DG.succ), c(NG.trials, DG.trials))

  prop.matrix.OV[col,] = c(OV.test$estimate[1], OV.test$estimate[2],
OV.test$estimate[1]-OV.test$estimate[2], OV.test$p.value, how.sig(OV.test$p.value)[1],
OV.test$conf.int[1], OV.test$conf.int[2])
  prop.matrix.DG[col,] = c(DG.test$estimate[1], DG.test$estimate[2],
DG.test$estimate[1]-DG.test$estimate[2], DG.test$p.value, how.sig(DG.test$p.value)[1],
DG.test$conf.int[1], DG.test$conf.int[2])
}
prop.OV = data.frame(prop.matrix.OV)
prop.DG = data.frame(prop.matrix.DG)
for(i in c(1:4,6:7)){
  prop.OV[,i] = as.numeric(as.character(prop.OV[,i]))
  prop.DG[,i] = as.numeric(as.character(prop.DG[,i]))
}
```

```
print(prop.OV, digits = 2)
print(prop.DG, digits = 2) # i think DG is actually the appropriate comparison to make.
```

So NG has 5% less Hispanic, 10% more white, 6.5% less female, and 7% more male representation than DG. Well that sucks, but also it's a story we sort of believed already. Note, Female + Male proportions add up to over 100% in the NG; one student answered that they were both Male and Female (we did not have a bigender option, though we did have a genderqueer option)

```{r check for missingness differences?}
all.demo = grepl("Q(1[5-9])|(2[0-2])", colnames(INICE$data))
NG.answered = any.filled(group.data$raw$NG[,all.demo])
DG.answered = any.filled(group.data$raw$DG[,all.demo])
ans.diff = prop.test(c(sum(NG.answered), sum(DG.answered)), c(NG.trials, DG.trials))
```

No difference in the proportion of students answering "the demographic questions" as a whole. By this we mean, if a student gave any response to any of questions 15-22, then they provided a response; 93.7% of both NG and DG provided at least some response.

Is there a difference on individual questions?

```{r check for missingness differences on the individual demographic questions}
question.groups = c("Q15", "Q16", "Q17", "Q18", "Q19", "Q20", "Q21", "Q22")
diff.matrix.DG = matrix(nrow = length(question.groups), ncol = 7, dimnames =
list(question.groups, c("NG.est", "DG.est", "Est.diff", "P.value", "sig", "CI.upper",
"CI.lower")))
for(qs in question.groups){
  demo.qs = grepl(qs, colnames(INICE$data))
  NG.succ = sum(any.filled(group.data$raw$NG[,demo.qs]))
  DG.succ = sum(any.filled(group.data$raw$DG[,demo.qs]))

  DG.test = prop.test(c(NG.succ, DG.succ), c(NG.trials, DG.trials))

  diff.matrix.DG[qs,] = c(DG.test$estimate[1], DG.test$estimate[2], DG.test$estimate[1]-
DG.test$estimate[2], DG.test$p.value, how.sig(DG.test$p.value)[1], DG.test$conf.int[1],
DG.test$conf.int[2])
}
diff.DG = data.frame(diff.matrix.DG)
for(i in c(1:4,6:7)){
  diff.DG[,i] = as.numeric(as.character(diff.DG[,i]))
}
rownames(diff.DG) = c("Ability", "Race", "Gender", "Sexuality", "Parent.Gender",
"Parent.Ed", "Relative.STEM", "US.status")
```

```
print(diff.DG, digits = 2)
```

```

There's no significant difference in the response rate of each of these blocks of questions between NG and DG. The closest one was "have your relatives worked in STEM or other", with $p < 0.06$, but that's also not really something we compared before. Most questions had >90% response rate, except for ability which had "only" an 84 and 86% response rate for the two groups.

```{r combining response rate differences with estimate differences}
true.prop.matrix = matrix(nrow = ncol(NG.demo), ncol = 7, dimnames =
list(colnames(NG.demo), c("NG.est", "DG.est", "Est.diff", "P.value", "sig", "CI.upper",
"CI.lower")))

for(col in colnames(NG.demo)){
  NG.succ = is.thing(NG.demo, col)
  DG.succ = is.thing(DG.demo, col)

  # ugly brute force way to get this done, not scalable
  demo.qs = grepl("Q16", colnames(INICE$data))
  if(col == "Able-bodied"){demo.qs = grepl("Q15", colnames(INICE$data))}


```
```

```
  if(col == "Female"){demo.qs = grepl("Q17", colnames(INICE$data))}
  if(col == "Straight"){demo.qs = grepl("Q16", colnames(INICE$data))}
  NG.trials = sum(any.filled(group.data$raw$NG[,demo.qs]))
  DG.trials = sum(any.filled(group.data$raw$DG[,demo.qs]))

  DG.test = prop.test(c(NG.succ, DG.succ), c(NG.trials, DG.trials))

  true.prop.matrix[col,] = c(DG.test$estimate[1], DG.test$estimate[2],
DG.test$estimate[1]-DG.test$estimate[2], DG.test$p.value, how.sig(DG.test$p.value)[1],
DG.test$conf.int[1], DG.test$conf.int[2])
}

true.prop.DG = data.frame(true.prop.matrix)
for(i in c(1:4,6:7)){
  true.prop.DG[,i] = as.numeric(as.character(true.prop.DG[,i]))
}

print(true.prop.DG, digits = 2)
print(true.prop.DG$Est.diff - prop.DG$Est.diff, drop = FALSE, digits = 3)
```

true.prop is a better measure of what the actual proportion of various populations is, because previous estimates counted everyone who didn't respond "positively" as a

"negative" response, i.e. Female proportion was "those who responded female" to "those who didn't respond female", rather than "those who responded with something other than female". The estimates were different by only small amount (less than 1%, maximum 0.68% for "straight", though that was only a difference of 1% to begin with).

```{r}
print(true.prop.DG$Est.diff - prop.DG$Est.diff, drop = FALSE, digits = 3) # repeated from above
print((true.prop.DG$Est.diff - prop.DG$Est.diff)/prop.DG$Est.diff * 100, digits = 3)

print((true.prop.DG$NG.est - prop.DG$NG.est)/prop.DG$NG.est * 100, digits = 3)
summary((true.prop.DG$NG.est - prop.DG$NG.est)/prop.DG$NG.est * 100)

print((true.prop.DG$DG.est - prop.DG$DG.est)/prop.DG$NG.est * 100, digits = 3)
summary((true.prop.DG$DG.est - prop.DG$DG.est)/prop.DG$NG.est * 100)

old = c(prop.DG$NG.est, prop.DG$DG.est)
true = c(true.prop.DG$NG.est, true.prop.DG$DG.est)
plot(old, true)
summary(lm(old~true))
```

As a percentage of the naive calculation estimate differences, we see that "Straight" changed by >60%, but the results for which we found significant differences changed by 10.6%, 2.8%, and 9.7%, respectively for Hispanic, White, and Female. The 10% results seem to be large enough that it's worth going through the hassle of calculating true percentages. The larger the difference in estimates, the less of a percent change there is however; on an absolute scale, the differences remained small (as stated in the previous section, all differences in differences were less than 1%(abs)).

The difference in the estimates themselves, however, was approximately 8% across the board for NG, but varied between 8% and 15% for the DG. Overall, the old values are extremely highly correlated with the new values (r=0.999977), with the old being ~91% of the true.

From here out, we'll use the true values when talking about the literal proportions present in the groups, as they seem like a better representative of the truth of the situation.

```{r make demographic data.frame reasonable}
OV.zero = NA.to.zero(OV.demo) # takes a bit, segregated to only run once while we debug
```

```{r make logit models for answered-the-question-block}
```

```
p.mat = matrix(ncol = length(question.groups), nrow = ncol(OV.zero)+1, dimnames =
list(c("Intercept", colnames(OV.zero)), question.groups))
e.mat = p.mat
model.list = NULL
for(qs in question.groups){
  demo.qs = grepl(qs, colnames(INICE$data))
  answered = any.filled(INICE$data[,demo.qs])
  OVplus = cbind(OV.zero, answered)
  model = glm(answered ~., family = binomial(link="logit"), data=OVplus)
  model.list[[qs]] = model
  p.mat[,qs] = summary(model)$coefficients[,4]
  e.mat[,qs] = summary(model)$coefficients[,1]
}
```

```{r print logit models, eval = FALSE}
for(qs in question.groups){
  print(qs)
  print(summary(model.list[[qs]]))
}
```

```{r fancy estimate table for significant results}
```

```
print.sig.df(e.mat, p.mat, digits = 4, p.sig = 0.05, trans = T)
```

```{r factor race factor gender}
OV.races = OV.zero[,2:5]
OV.genders = OV.zero[,6:7]

factor.demo = function(row){
  if(!any.filled(row)){
    return(NA)
  } else if(sum(row) == 1){
    return(colnames(row)[which(row == 1)])
  } else if(sum(row) > 1){
    return("Multi")
  } else{return("Other")}
}
OV.able = NULL
OV.straight = NULL
OV.race = NULL
OV.gender = NULL
for(row in rownames(OV.races)){
  OV.race[row] = factor.demo(OV.races[row,])
  OV.gender[row] = factor.demo(OV.genders[row,])
```

```
}

OV.sub = data.frame(OV.zero$"Able-Bodied", OV.zero$Straight, OV.race, OV.gender)
colnames(OV.sub) = c("Able", "Straight", "Race", "Gender")
OV.sub$Gender = relevel(OV.sub$Gender, "Male")
OV.sub$Race = relevel(OV.sub$Race, "White")

pf.mat = matrix(ncol = length(question.groups), nrow = 11, dimnames =
list(c("Intercept", c("Able", "Straight", "Asian", "Black", "Hispanic", "R.Multi",
"R.Other", "Female", "G.Multi", "G.Other")), question.groups))
ef.mat = pf.mat
modelf.list = NULL
for(qs in question.groups){
  demo.qs = grepl(qs, colnames(INICE$data))
  answered = any.filled(INICE$data[,demo.qs])
  OVplus = data.frame(OV.sub, answered)

  model = glm(answered ~., family = binomial(link="logit"), data=OVplus)
  modelf.list[[qs]] = model
  pf.mat[,qs] = summary(model)$coefficients[,4]
  ef.mat[,qs] = summary(model)$coefficients[,1]
}
```

```
print.sig.df(ef.mat, pf.mat, trans = T)
```

Can we predict group membership by demographic factor?

```{r}
NG.names = rownames(NG.demo)
DG.names = rownames(DG.demo)
OV.sub$Group = NA
OV.sub[NG.names, c("Group")] = 1
OV.sub[DG.names, c("Group")] = 0

model.NG = glm(Group ~., family = binomial(link="logit"), data = OV.sub)
all.coef.NG = summary(model.NG)$coefficient[,1]
sig.coef.NG = all.coef.NG[summary(model.NG)$coefficient[,4] < 0.05]

CI.NG = confint(model.NG, parm = names(sig.coef.NG))
exp(CI.NG)

est.frame = cbind(CI.NG[,1], sig.coef.NG, CI.NG[,2])
colnames(est.frame) = c("2.5 %", "Est", "97.5 %")
odds.frame = exp(est.frame)
```

```
# OV.sub[NG.names, c("Group")] = 0
# OV.sub[DG.names, c("Group")] = 1
#
# model.DG = glm(Group ~., family = binomial(link="logit"), data = OV.sub)
# all.coef.DG = summary(model.DG)$coefficient[,1]
# sig.coef.DG = all.coef.DG[summary(model.DG)$coefficient[,4] < 0.05]
#
# CI.DG = confint(model.DG, parm = names(sig.coef.DG))
#
# exp(CI.DG)


# exp(sig.coef.NG)*exp(sig.coef.DG)
## turns out from the above that sig.coef.DG are just the inverse log ods of sig.coef.NG. Which should have made sense from the start.
```

Significant results: Intercept (-1.2204), Asian (-0.5332), Black (-0.7188), Hispanic (-0.6520), Female (-0.3925). Both multiracial and other-racial were right on the edge of significance, and were also negative.

From odds frame, we see that Asian, Black, and Hispanic people are roughly 50% as likely as White people to be in NG, and Female-identified people are roughly 2/3rds as likely as Male-identified people to be in NG.

Able and Straight aren't doing anything for group membership, so we'll pull those out for now, and intersect race and gender first.

```{r semi-intersectional logit}
factor.demo = function(row){ # improved function to deal with intersections
  if(!any.filled(row)){
    return(NA)
  } else{
    return(paste(colnames(row)[which(row == 1)], collapse = " "))
  }
}

OV.RG = INICE$data[,grepl("Q(16|17)[a-h]$", colnames(INICE$data))]
OV.R = INICE$data[,grepl("Q16[a-h]$", colnames(INICE$data))]
OV.G = INICE$data[,grepl("Q17[a-h]$", colnames(INICE$data))]
OV.RG = NA.to.zero(OV.RG)
OV.R = NA.to.zero(OV.R)
OV.G = NA.to.zero(OV.G)
colnames(OV.RG) = c( "Am.Indian", "Hispanic", "White", "Asian", "Middle.Eastern",
"Another.race", "Black", "Hawaiian"
                    , "Female", "Trans", "Male", "Cis", "Genderqueer", "Another.gender"
,"Agender"
```

```
                )
colnames(OV.R) = c( "Am.Indian", "Hispanic", "White", "Asian", "Middle.Eastern",
"Another.race", "Black", "Hawaiian"
                )
colnames(OV.G) = c( "Female", "Trans", "Male", "Cis", "Genderqueer",
"Another.gender" ,"Agender"
                )
OV.int.demo = NULL
OV.int.R = NULL
OV.int.G = NULL
#
# OV.able = NA.to.zero(INICE$data[,grepl("Q15[a-h]$", colnames(INICE$data))])
# OV.sex = NA.to.zero(INICE$data[,grepl("Q18[a-z]$", colnames(INICE$data))])
#
# OV.sex.f = NULL
# OV.able.f = NULL
#
# for(row in rownames(OV.RG)){
#   OV.sex.f[row] = factor.demo(OV.sex[row,])
#   OV.able.f[row] = factor.demo(OV.able[row,])
# }
#
Straight = recode(OV.sex.f, "'Q18a' = 1; NA = NA; else = 0")
```

```
Not.straight = recode(OV.sex.f, "'Q18a' = 0; NA = NA; else = 1")
Able = recode(OV.able.f, "'Q15e' = 1; NA = NA; else = 0")
Not.able = recode(OV.able.f, "'Q15e' = 0; NA = NA; else = 1")
#
# OV.RG = data.frame(Able, Not.able, OV.RG, Straight, Not.straight)

for(row in rownames(OV.RG)){
  OV.int.demo[row] = factor.demo(OV.RG[row,])
  OV.int.R[row] = factor.demo(OV.R[row,])
  OV.int.G[row] = factor.demo(OV.G[row,])
}

OV.int.g = data.frame(OV.int.demo, OV.sub$Group)
OV.int.G = data.frame(OV.int.G, OV.sub$Group)
OV.int.R = data.frame(OV.int.R, OV.sub$Group)
OV.int.GR = data.frame(OV.int.G, OV.int.R, OV.sub$Group)

colnames(OV.int.g) = c("Factor", "Group")
colnames(OV.int.G) = c("Factor", "Group")
colnames(OV.int.R) = c("Factor", "Group")
colnames(OV.int.GR) = c("Gender", "Race", "Group")
OV.int.g$Factor = relevel(OV.int.g$Factor, "White Male")
OV.int.G$Factor = relevel(OV.int.G$Factor, "Male")
```

```
OV.int.R$Factor = relevel(OV.int.R$Factor, "White")
OV.int.GR$Gender = relevel(OV.int.GR$Gender, "Male")
OV.int.GR$Race = relevel(OV.int.GR$Race, "White")

nrow(OV.int.GR[!OV.int.GR$Gender%in%c("Male", "Female") &
!OV.int.GR$Race%in%c("White", "Asian", "Black", "Hispanic", "White Hispanic") &
!is.na(OV.int.GR$Group) & ( !is.na(OV.int.GR$Gender) | !is.na(OV.int.GR$Race) ),]) #
but it turns out I don't need this number

model = glm(Group~Factor, family=binomial(link="logit"), data=OV.int.g)
model.G = glm(Group~Factor, family=binomial(link="logit"), data=OV.int.G)
model.R = glm(Group~Factor, family=binomial(link="logit"), data=OV.int.R)
model.GR = glm(Group~Gender + Race, family = binomial(link="logit"),
data=OV.int.GR)
sig.rows = summary(model)$coefficients[,4] < 0.05
sig.coef = summary(model)$coefficients[sig.rows,1]

sig.R = summary(model.R)$coefficients[summary(model.R)$coefficients[,4] < 0.05, 1]
sig.G = summary(model.G)$coefficients[summary(model.G)$coefficients[,4] < 0.05, 1]

CI = confint(model, parm = names(sig.coef))
est.frame = cbind(CI[,1], sig.coef, CI[,2])
colnames(est.frame) = c("2.5 %", "Est", "97.5 %")
```

```
odds.frame = exp(est.frame)
```

Not-disabled and straight are the only ones with sufficient numbers in these intersections to show up as significant (i.e. with enough power). Without ability, Black Female shows up (negative estimate), but Hispanic Female vanishes. Without sexuality, Hispanic Female shows up (negative estimate), but Black Female is not present. Without sexuality or ability, both Black Female and Hispanic Female are present.

Ability and sexuality don't show any significant differences, though there do seem to be differences in the proportions and likelihoods. I wonder if this is simply a matter of statistical power since we have so few people that fit into these categories. (for example, 2608 straight, 38 Bi, and 21 Gay (with 249 NAs)).

So we limit our intersection to Race x Gender, and get the 7 significant estimates found.

```{r subsetting the intersection based on gender or race}
OV.f = OV.int.g
OV.female = OV.f
OV.female$Factor = relevel(OV.f$Factor, "White Female")
OV.female = OV.female[OV.female$Factor %in% c("White Female", "Black Female", "Hispanic Female", "Hispanic White Female", "Asian Female"),]
OV.female$Factor = factor(OV.female$Factor)
```

```
OV.male = OV.f[grepl("Male", OV.f$Factor),]
OV.male = OV.male[OV.male$Factor %in% c("White Male", "Black Male", "Hispanic Male", "Asian Male", "Hispanic White Male"),]
OV.male$Factor = relevel(OV.male$Factor, "White Male")
OV.male$Factor = factor(OV.male$Factor)

model = glm(Group ~., family=binomial(link="logit"), data = OV.female)
```

vs doing it with propr testing. Is this actually a better version of what we're trying to show?

```{r doing race x gender with proportions}
OV.female$Factor = factor(OV.female$Factor)
test = prop.test(table(OV.female))
test = prop.test(table(OV.female)[c(1,3),c(2, 1)]) # almost significant.

# power level of these differences?
counts = table(OV.female)
for(i in 2:4){
  n1 = sum(counts[1,])
  n2 = sum(counts[i,])
  p1 = counts[1,1] / n1
```

```
  p2 = counts[i,1] / n2
  h = ES.h(p1, p2)
  test = pwr.2p2n.test(h, n1, n2)
  print(test$power)
}
```

So for women we have 11%, 77% (almost!) and 13% power. Can't do much.

```{r}
print.powers = function(factor.frame){
  counts = table(factor.frame)
  for(i in 2:nrow(counts)){
    n1 = sum(counts[1,])
    n2 = sum(counts[i,])
    p1 = counts[1,1] / n1
    p2 = counts[i,1] / n2
    h = ES.h(p1, p2)
    test = pwr.2p2n.test(h, n1, n2)
    print(paste(rownames(counts)[i], ": ", test$power, sep = ""))
  }
}
```

```
OV.male$Factor = factor(OV.male$Factor)
print.powers(OV.male) #sort of obvious, but no Black Male power
```

```{r}
OV.asian = OV.f[grepl("Asian", OV.f$Factor),]
OV.asian = OV.asian[OV.asian$Factor %in% c("Asian Male", "Asian Female"),]
OV.asian$Factor = relevel(OV.asian$Factor, "Asian Male")
OV.asian$Factor = factor(OV.asian$Factor)

OV.black = OV.f[grepl("Black", OV.f$Factor),]
OV.black = OV.black[OV.black$Factor %in% c("Black Male", "Black Female"),]
OV.black$Factor = relevel(OV.black$Factor, "Black Male")
OV.black$Factor = factor(OV.black$Factor)

OV.hispanic = OV.f[grepl("Hispanic", OV.f$Factor),]
OV.hispanic = OV.hispanic[OV.hispanic$Factor %in% c("Hispanic Male", "Hispanic
Female", "Hispanic White Male"),]
OV.hispanic$Factor = relevel(OV.hispanic$Factor, "Hispanic Male")
OV.hispanic$Factor = factor(OV.hispanic$Factor)

print.powers(OV.asian)
print.powers(OV.black)
```

print.powers(OV.hispanic)

```

```{r}

asian = INICE$data$Q16d

hispanic = INICE$data$Q16b

white = INICE$data$Q16c

black = INICE$data$Q16g

male = INICE$data$Q17c

female = INICE$data$Q17a

wh = white & !is.na(white) & hispanic & !is.na(hispanic)

wo = white & !wh & !is.na(white)

ho = hispanic & !wh & !is.na(hispanic)

fh = (female & hispanic & !is.na(female) & !is.na(hispanic))

fa = (female & asian & !is.na(female) & !is.na(asian))

fb = (female & black & !is.na(female) & !is.na(black))

fw = (female & white & !is.na(female) & !is.na(white))

mh = (male & hispanic & !is.na(male) & !is.na(hispanic))

ma = (male & asian & !is.na(male) & !is.na(asian))

mb = (male & black & !is.na(male) & !is.na(black))

```
mw = (male & white & !is.na(male) & !is.na(white))

mhw = mh & mw

mh = mh & !mhw

mw = mw & !mhw

fhw = fh & fw

fho = fh & !fw

fwo = fw & ! fh

mat = matrix(nrow = 14, ncol = 2)

rownames(mat) = c("fh", "fa", "fb", "fw", "mh", "ma", "mb", "mw", "h", "a", "b", "w",

"f", "m")

colnames(mat) = c("succ", "trial")

mat["fh",] = c(table(OV.f$Group[fh])[[2]], sum(table(OV.f$Group[fh])))

mat["fa",] = c(table(OV.f$Group[fa])[[2]], sum(table(OV.f$Group[fa])))

mat["fb",] = c(table(OV.f$Group[fb])[[2]], sum(table(OV.f$Group[fb])))

mat["fw",] = c(table(OV.f$Group[fw])[[2]], sum(table(OV.f$Group[fw])))

mat["mh",] = c(table(OV.f$Group[mh])[[2]], sum(table(OV.f$Group[mh])))

mat["ma",] = c(table(OV.f$Group[ma])[[2]], sum(table(OV.f$Group[ma])))

mat["mb",] = c(table(OV.f$Group[mb])[[2]], sum(table(OV.f$Group[mb])))

mat["mw",] = c(table(OV.f$Group[mw])[[2]], sum(table(OV.f$Group[mw])))
```

```
mat["h",] = c(table(OV.f$Group[!is.na(hispanic)])[[2]],
sum(table(OV.f$Group[!is.na(hispanic)])))
mat["a",] = c(table(OV.f$Group[!is.na(asian)])[[2]],
sum(table(OV.f$Group[!is.na(asian)])))
mat["b",] = c(table(OV.f$Group[!is.na(black)])[[2]],
sum(table(OV.f$Group[!is.na(black)])))
mat["w",] = c(table(OV.f$Group[!is.na(white)])[[2]],
sum(table(OV.f$Group[!is.na(white)])))
mat["f",] = c(table(OV.f$Group[!is.na(female)])[[2]],
sum(table(OV.f$Group[!is.na(female)])))
mat["m",] = c(table(OV.f$Group[!is.na(male)])[[2]],
sum(table(OV.f$Group[!is.na(male)])))

t.mat = matrix(nrow = 14, ncol= 14, dimnames = list(rownames(mat), rownames(mat)))
for(i in 1:nrow(t.mat)){
  for(j in 1:nrow(t.mat)){
    t.mat[i, j] = prop.test(cbind(mat[i,], mat[j,]))$p.value
  }
}
```

```{r}
```

```
hisp.table = rbind(table(OV.f$Group[wh]), table(OV.f$Group[ho]),
table(OV.f$Group[mhw]), table(OV.f$Group[mh]), table(OV.f$Group[fhw]),
table(OV.f$Group[fho]))
rownames(hisp.table) = c("white.hisp", "hisp", "male white.hisp", "male hisp", "female
white.hisp", "female hisp")
hisp.table = hisp.table[,2:1]
```

```{r do smaller factor logit regression}
group = OV.int.g$Group
demo.sub.frame = data.frame(asian, hispanic, black, white, female, male)
demo.sub.frame = NA.to.zero(demo.sub.frame)
demo.factor = NULL
for(i in 1:nrow(demo.sub.frame)){
  demo.factor[i] = factor.demo(demo.sub.frame[i,])
}
demo.factor = relevel(factor(demo.factor), "white male")
demo.group = data.frame(demo.factor, group)

limited.model = glm(group~., family=binomial(link="logit"), data=demo.group)
sig.limit = summary(limited.model)$coefficients[,4] < 0.05
sig.limit.coef = summary(limited.model)$coefficients[sig.limit, 1]
```

```
OV.rf.female = OV.int.g
OV.rf.female$Factor = relevel(OV.int.g$Factor, "White Female")
model = glm(Group~Factor, family=binomial(link="logit"), data=OV.rf.female)
sig.rows = summary(model)$coefficients[,4] < 0.05
sig.coef = summary(model)$coefficients[sig.rows,1]
print(sig.coef)

for(refact in c("White Female", "Asian Male", "Black Male", "Hispanic Male", "Hispanic White Male")){
  OV.rf.female$Factor = relevel(OV.int.g$Factor, refact)
  model = glm(Group~Factor, family=binomial(link="logit"), data=OV.rf.female)
  sig.rows = summary(model)$coefficients[,4] < 0.05
  sig.coef = summary(model)$coefficients[sig.rows,1]
  print(sig.coef)
}
```

```{r better "power calculation", finding }
pwr.2p2n.test(n1 = 2040, n2 = 519, power = 0.8) # = 0.1377

ES.h.p = function(h, p){
  asp = asin(sqrt(p))
  p2.down = sin(0.5* h - asp)^2
```

```
  p2.up   = sin(0.5* h + asp)^2
  return(c(p2.down, p2.up))
}

prop.list = seq(0.00, 0.95, by = 0.0005)
prop.down = seq(0.05, 0.95, by = 0.05)
prop.up = seq(0.05, 0.95, by = 0.05)
for(i in 1:length(prop.list)){
  prop.down[i] = ES.h.p(0.137, prop.list[i])[1]
  prop.up[i] = ES.h.p(0.137, prop.list[i])[2]
}

plot(prop.list, prop.up, type = "l")
points(prop.list, prop.down, type = "l")

mean.diff = (prop.up-prop.down)/2

plot(c(prop.list, prop.list), c(prop.up-prop.list, prop.down-prop.list), type = "l")
plot(prop.list, (prop.up-prop.list)/prop.list)
plot(prop.list, (mean.diff), type ="l", xlim = c(0.3, 0.7))
```

```{r}
library(plotly)
d <- data.frame(prop.list, mean.diff)
colnames(d) = c("base.prop", "mean.difference")
# without log scales
p <- plot_ly(d, x = ~base.prop, y = ~mean.difference) %>% add_markers()

p <- layout(p, xaxis = list(type = "log"))
```

\section Talking about Group Differences that we found

```{r ######## new section ########}
```

```{r double check the group differences}

NC = ncol(group.data$factors$NG) # number of columns of factors
NG = length(group.data$factors$NnG) # number of NnG
for(id in 1:length(group.data$factors$NnG)){
  sd = rep(NA, times = NC)
  sig = rep(NA, times = NC)
  p = rep(NA, times = NC)
```

```
  for(col in 1:NC){
    sd[col] = sd(group.data$factors$NnG[[id]][,col])
    twpt = twoway.perm.test(group.data$factors$NG[,col],
group.data$factors$NnG[[id]][,col])
    p[col] = p.adjust(pvalue(twpt)[1], n = NC*NG) # n = 7*13 = 91
  }
  sig = how.sig(p)

  group.data$factors$p$NnG[[id]] = p
  group.data$factors$sig$NnG[[id]]= sig
  group.data$factors$sd$NnG[[id]] = sd
}
```

So the old classifications we had are worthwhole, it seems. Even if we check them against the overall NG factors (where NG = NG1, NG2, SN1, and SN2). Because SN3 shows no significant differences after correcting for the number of tests, we'll also roll that into it.

Let's make the significance table into something pretty.

```{r}
sig.mat = matrix(nrow = NC, ncol = NG)
```

```
for(g in 1:NG){
  sig.mat[,g] = as.matrix(group.data$factors$sig$NnG[[g]])
}
# since the matrix has all three levels of significance (*, **, and ***), we can make it a df nicely
sig.df = data.frame(sig.mat)
colnames(sig.df) = nng.names
rownames(sig.df) = c("Value", "Work.Avoid", "Connectedness", "Perc.of.Fut", "Neuroticism", "Extroversion", "Belongingness", "Perform.App", "Instrumentality", "Grit.CoI", "EngID.PC", "EngID.Rec", "PhysID.Rec")
```

Every group except NnG_1 and NnG_6 has 3 factors on which they significantly differ from the normative group. NnG_1 has two, and NnG_6 has one.

Perceptions of Future, Neuroticism, Belongingness, and Instrumentality have no groups which show differences from NG. The other nine factors have at least one group which shows a difference.

```{r effect size of significant differences}

for(g in 1:NG){
  d = NULL
```

```
  for(fact in 1:NC){
    d[fact] = cohens.d(group.data$factors$NG[,fact], group.data$factors$NnG[[g]][,fact])
  }
  names(d) = c("Value", "Work.Avoid", "Connectedness", "Perc.of.Fut", "Neuroticism",
"Extroversion", "Belongingness", "Perform.App", "Instrumentality", "Grit.CoI",
"EngID.PC", "EngID.Rec", "PhysID.Rec")
  group.data$factors$d$NnG[[g]] = d
}

d.mat = matrix(ncol = NG, nrow = NC)
for(g in 1:NG){
  d.mat[,g] = group.data$factors$d$NnG[[g]]
}
d.df = data.frame(d.mat)
rownames(d.df) = names(d)
colnames(d.df) = nng.names

# to extract just the significant ones in a messier data frame)
d.mat[sig.df == ""] = ""
d.df2 = data.frame(d.mat)
rownames(d.df2) = names(d)

for(g in 1:NG){
```

```
  diff = NULL
  for(fact in 1:NC){
    diff[fact] = mean(group.data$factors$NG[,fact], na.rm = T) - mean(
group.data$factors$NnG[[g]][,fact], na.rm = T)
  }
  names(diff) = c("Value", "Work.Avoid", "Connectedness", "Per.of.Fut", "Neuroticism",
"Extroversion", "Belongingness", "Perform.App", "Instrumentality", "Grit.CoI",
"EngID.PC", "EngID.Rec", "PhysID.Rec")
  group.data$factors$diff$NnG[[g]] = diff
}
diff.mat = matrix(ncol = NG, nrow = NC)
for(g in 1:NG){
  diff.mat[,g] = group.data$factors$diff$NnG[[g]]
}
diff.df = data.frame(diff.mat)
rownames(diff.df) = names(d)
colnames(diff.df) = nng.names

# to extract just the significant ones in a messier data frame)
diff.mat[sig.df == ""] = ""
diff.df2 = data.frame(diff.mat)
rownames(diff.df2) = names(d)
```

Ok so we have some effect sizes. Need to talk with people to see if it's ok to bring Cohen's d into the picture when we're selecting subgroups (and thus artificially limiting the variance -> boosting the effect size)

```{r}
library(fmsb)
maxes = c(rep(6, times = NC))
mins = c(rep(0, times = NC))

starplus  = data.frame(rbind(maxes, mins, t(group.data$centroid)))
colnames(starplus) = c("Value", "Work.Avoid", "Connectedness", "Per.of.Fut", "Neuroticism", "Extroversion", "Belongingness", "Perform.App", "Instrumentality", "Grit.CoI", "EngID.PC", "EngID.Rec", "PhysID.Rec")
rownames(starplus) = c("maxes" ,  "mins" ,   "NG"  ,"NnG_1"  , "NnG_2" ,  "NnG_3" , "NnG_4"  , "NnG_5", "NnG_6" ,  "NnG_7")

asc.order = order(group.data$centroid[,1])
starorder = starplus[,asc.order]
starorder[3:10,] = starorder[c(4:10,3),] # move NG to the end
rownames(starplus) = c("maxes" ,  "mins"  ,"NnG_1"  , "NnG_2" ,  "NnG_3" ,  "NnG_4" , "NnG_5", "NnG_6" ,  "NnG_7", "NG")
```

```
radarchart(starorder, seg = 6, pcol = c(2:8, 1), maxmin = TRUE, centerzero = TRUE,
plwd =3, plty = 1)
legend("topleft", legend = rownames(starorder)[c(4:10,3)], fill = c(2:8, 1))
```

This gives us a big old smear of lines, so we break them out one by one, and also graph them with standard errors to show overlap/non-overlap.

The standard error on the normative group is rather small (0.02-0.05) so mostly this will be driven by the other factor.

```{r}
NGsd = sapply(group.data$factors$NG, sd)
NGupper = group.data$centroid[,"NG"] + NGsd/sqrt(nrow(group.data$factors$NG))
NGlower = group.data$centroid[,"NG"] - NGsd/sqrt(nrow(group.data$factors$NG))
for(g in 1:NG){
  sd = sapply(group.data$factors$NnG[[g]], sd, na.rm = T)
  upper = group.data$centroid[,(g+1)] + sd/sqrt(nrow(group.data$factors$NnG[[g]]))
  lower = group.data$centroid[,(g+1)] - sd/sqrt(nrow(group.data$factors$NnG[[g]]))

  new.star = data.frame(rbind(maxes, mins, group.data$centroid[,"NG"], NGupper,
NGlower, group.data$centroid[,(g+1)], upper, lower))
  new.star = new.star[,asc.order]
```

```
  radarchart(new.star, seg = 6, pcol = rep(c(1, g+1), each = 3), plty = rep(c(1,3,3), times =
2), maxmin = TRUE, centerzero = TRUE, plwd =2)
  legend("topleft", legend = c("NG", colnames(group.data$centroid)[g+1]), fill = c(1,
g+1))
}
```

```{r skew of factor subspace distributions}
skew(unlist(factor.subspace))
sd(unlist(factor.subspace))
mean(unlist(factor.subspace))
summary(unlist(factor.subspace))
```

```{r}
overlap.plot = function(col1, col2){
  plot(group.data$factors$DG[,c(col1, col2)], xlim = c(0,6), ylim = c(0,6), pch = 22)
  points(group.data$factors$NG[,c(col1,col2)], xlim = c(0,6), ylim = c(0,6), pch = 20, col
= "red", cex = 1.5)
}
```

```{r make a vector and/or group relation graph}
library(igraph)

fact.mat = matrix(nrow = nrow(sig.df), ncol = nrow(sig.df), data= 0, dimnames =
list(rownames(sig.df), rownames(sig.df)))
group.mat = matrix(nrow = ncol(sig.df), ncol = ncol(sig.df), data= 0, dimnames =
list(colnames(sig.df), colnames(sig.df)))

tf.mat = sig.df != ""

for(name in rownames(fact.mat)){
  for(i in 1:ncol(tf.mat)){
    if(tf.mat[name,i]){
      for(f.name in rownames(fact.mat)[rownames(fact.mat) != name]){
        if(tf.mat[f.name, i]){
          fact.mat[name, f.name] = fact.mat[name, f.name] + 1
        }
      }
    }
  }
}
```

```
for(name in rownames(group.mat)){
  for(i in 1:nrow(tf.mat)){
    if(tf.mat[i,name]){
      for(f.name in rownames(group.mat)[rownames(group.mat) != name]){
        if(tf.mat[i, f.name]){
          group.mat[name, f.name] = group.mat[name, f.name] + 1
        }
      }
    }
  }
}

g.fact <- graph_from_adjacency_matrix(fact.mat, weighted=TRUE, mode =
"undirected")
g.group = graph_from_adjacency_matrix(group.mat, weighted=TRUE, mode =
"undirected")

write.Gephi.regular = function(my.map, filepattern){
  if(length(unique(unlist(V(my.map)$name))) != length(unlist(V(my.map)$name))){
    V(my.map)$old.names = V(my.map)$name
    V(my.map)$name = c(1:length(unlist(V(my.map)$name)))
    warning("Wrote old names to attribute $old.names, new names given")
  }
```

```
  my.df.v = get.data.frame(my.map, what = "vertices")
  my.df.v = my.df.v[,!grepl("members$", colnames(my.df.v))] # sadly the member list is
poorly behaved
  colnames(my.df.v)[1] = c("Id") # seriously Gephi? You're weird
  my.df.e = get.data.frame(my.map, what = "edges")
  if(ncol(my.df.e) > 2){
    colnames(my.df.e)[1:3] = c("Source", "Target", "Weight") # seriously Gephi? You're
weird
  } else{
    colnames(my.df.e)[1:2] = c("Source", "Target")
  }

  node.path = paste(filepattern, ".node.list.csv", sep = "")
  edge.path = paste(filepattern, ".edge.list.csv", sep = "")
  write.table(as.matrix(my.df.v), file = node.path, row.names = FALSE, sep = ",")
  write.table(as.matrix(my.df.e), file = edge.path, row.names = FALSE, sep = ",")
}
```

Qualitative_salience.Rmd

```
---
title: "Qualitative Salience"
author: "Jackie Doyle"
date: "May 15, 2017"
output: pdf_document
---

Get information about the qualitative interview participants. Most of the stuff is just done in RQDA.

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r load libraries and files}
library(RQDA)
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Mapping/groupdata.RData")
load("C:/Users/Jackie/Downloads/InICE Data Files/openemails_v2.RData")
load("C:/Users/Jackie/Dropbox/InIce Proposal 2014/Analysis/Study Data/Mapping/env_v2_group_comparisons_done.RData")
source('C:/Users/Jackie/Dropbox/R Files/custom.R', echo=FALSE)
```
```

```

```{r getting the emails}
(this section has been redacted for privacy and IRB compliance)
```

So now we have the row numbers for each student

```{r make student data frame}
student.rows = c(allison, betty, cara, elisa, pilar)
student.raw = INICE$data[student.rows,]
rownames(student.raw) = c("Allison", "Betty", "Cara", "Elisa", "Pilar")
student.factors = factor.space[student.rows,]
rownames(student.factors) = c("Allison", "Betty", "Cara", "Elisa", "Pilar")
```

```{r make composite identity factors}
student.factors$PhysID = rowMeans(student.factors[,grep("Phys ID",
colnames(student.factors))])
student.factors$MathID = rowMeans(student.factors[,grep("Math ID",
colnames(student.factors))])
student.factors$EngID = rowMeans(student.factors[,grep("Eng ID",
colnames(student.factors))])

```
all.factors = factor.space
all.factors$PhysID = rowMeans(all.factors[,grep("Phys ID", colnames(all.factors))])
all.factors$MathID = rowMeans(all.factors[,grep("Math ID", colnames(all.factors))])
all.factors$EngID = rowMeans(all.factors[,grep("Eng ID", colnames(all.factors))])

for(i in (ncol(all.factors)-2):ncol(all.factors)){
  print(colnames(all.factors)[i])
  print(mean(all.factors[,i]))
  print(sd(all.factors[,i]))
  print(se(all.factors[,i]))
}

# > student.factors[,28:30] - 4.24
#            PhysID     MathID      EngID
# Allison  1.6266667 1.56000000  1.2600000
# Betty   -0.9288889 1.69333333 -0.1900000
# Cara    -1.6400000 1.62666667  0.5933333
# Elisa    0.1600000 0.82666667  1.4266667
# Pilar   -0.6177778 0.04888889  0.8433333
# > student.factors[,28:30] - 4.66
#            PhysID     MathID      EngID
# Allison  1.206667  1.1400000  0.8400000
```

```
# Betty   -1.348889  1.2733333 -0.6100000
# Cara    -2.060000  1.2066667  0.1733333
# Elisa   -0.260000  0.4066667  1.0066667
# Pilar   -1.037778 -0.3711111  0.4233333
# > student.factors[,28:30] - 4.80
#           PhysID     MathID       EngID
# Allison  1.066667  1.0000000  0.70000000
# Betty   -1.488889  1.1333333 -0.75000000
# Cara    -2.200000  1.0666667  0.03333333
# Elisa   -0.400000  0.2666667  0.86666667
# Pilar   -1.177778 -0.5111111  0.28333333
#
# > cor(student.factors[,28:30])
#           PhysID       MathID      EngID
# PhysID 1.000000000  0.002992156  0.6481753
# MathID 0.002992156  1.000000000 -0.3861278
# EngID  0.648175325 -0.386127777  1.0000000
# > cor(t(student.factors[,28:30]))
#           Allison      Betty       Cara      Elisa      Pilar
# Allison  1.00000000 0.07526572 -0.3713797 -0.9277446 -0.9548484
# Betty    0.07526572 1.00000000  0.8978953  0.3023297  0.2243835
# Cara    -0.37137972 0.89789529  1.0000000  0.6910692  0.6304571
# Elisa   -0.92774461 0.30232968  0.6910692  1.0000000  0.9967354
```

Pilar   -0.95484837 0.22438353 0.6304571 0.9967354 1.0000000

```

From correlations between people, we see that Eliza and Pilar have similar scores (r~1), and they're very different from Allison (r~ -0.99), while  Betty and Cara are most like each other (r~ 0.89), and a little like Elisa and Pilar (but weak, r ~ 0.3)

So we'e got a pretty good spread of values across the space. And from the deltas, we're not looking at "everyone who is low PhysID", or EngID. However, Math Identity is generally higher (except Eliza and Pilar, who are sort of average).

```{r demographic questions}
demographics = student.raw[,grep("Q(1[5-9])|(2[0-2])", colnames(student.raw))]
interests = student.raw[,grep("Q14", colnames(student.raw))]
colnames(interests) = c(
  "AAE", "ABE_BSE", "BE_BME", "CME", "CVL", "CE_CSE", "CON", "EE", "EP",
"EEE", "IND", "IT", "MSE", "ME", "MIE", "NUKE", "OSTEM", "ONON"
)

interests - rowMeans(interests)
```
```

```{r identity subscores}
id.subscores = student.factors[,grep("ID:", colnames(student.factors))]
colnames(id.subscores) = c("EngID:PC", "EngID:Rec", "EngID:Int", "PhysID:PC",
"PhysID:Rec", "PhysID:Int", "MathID:Rec", "MathID:PC", "MathID:Int")
id.subscores[,7:8] = id.subscores[,8:7] # to match the pattern of the other two
print(id.subscores, digits = 2)
```

```{r distance from the normative group}
normative.group = group.data$centroid$NG
student.subspace = factor.subspace[student.rows,]
dist.df = rbind(normative.group, student.subspace)
rownames(dist.df) = c("normative group", rownames(student.factors))
dist(dist.df)
```

```{r FIU underrepresented in the normative group}
school = INICE$data$school
group = OV.sub$Group
summary(glm(group ~ school, family = binomial(link="logit")))
```

So students from FIU are statistically about half as likely to be in the normative group, which makes sense because of the high proportion of non-white students at FIU, and the same regression as before.

VITA

JACQUELINE DOYLE

Born, Boston, Massachusetts

2011	B.S., Physics Purdue University West Lafayette, Indiana
	Teaching Assistant Purdue University West Lafayette, Indiana
2014	M.S., Physics Purdue University West Lafayette, Indiana
	Teaching Assistant Purdue University West Lafayette, Indiana
2017	Doctoral Candidate Florida International University Miami, Florida
	Teaching Assistant Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Doyle, J., & Potvin, G. (2015). In search of distinct graduate admission strategies in physics: An exploratory study using topological data analysis. *2015 Physics Education Research Conference Proceedings*, 107–110. https://doi.org/10.1119/perc.2015.pr.022

Doyle, J. (2016). *Emergent Attitudinal Profiles of Introductory Engineering Students*. Presentation, 19th Annual Meeting of the Physics Education Research Conference.

Doyle, J. (2015). *Quantitative Intersectionality and Attitudinal Profiles with Cluster Analysis*. Presentation, CU Boulder, Boulder, CO.

Doyle, J. (2016). *"Who Can Be An Engineer?" Investigating Attitudes and Self-Identification*. Presentation, 2016 Summer Meeting of the American Association of Physics Teachers.

Doyle, J. (2016). *In search of distinct graduate admission strategies in physics: An exploratory study using topological data analysis.* Presentation, 2015 Summer Meeting of the American Association of Physics Teachers.

Doyle, J., & Potvin, G. (2015). In search of distinct graduate admission strategies in physics: An exploratory study using topological data analysis. *2015 Physics Education Research Conference Proceedings*, 107–110. https://doi.org/10.1119/perc.2015.pr.022

Kirn, A., Godwin, A., Benson, L., Potvin, G., Doyle, J. M., Verdín, D., & Boone, H. (2016). Intersectionality of Non-normative Identities in the Cultures of Engineering (InIce). *Proceedings of the 123rd ASEE Annual Conference and Exposition*, abstract accepted, final paper due Feb. 1, 2016. https://doi.org/10.18260/p.25448